Perceptual Alignment in Artificial Vision: Bio-Inspired Design and Psychophysical Evaluation

Pablo Hernández Cámara

Directors:
Valero Laparra Pérez-Muelas
Jesús Malo López

VNIVERSITAT DE VALÈNCIA

Doctorado en Ingeniería Electrónica
Escuela Técnica Superior de Ingeniería
Universitat de València

September, 2025

Note to the reader

According to the University of Valencia Doctorate Regulation[1,2] this PhD dissertation is presented as a compendium of at least three publications in international journals containing the results of the conducted work. It also describes work that has recently been submitted to scientific journals.

Furthermore, in accordance with the aforementioned regulation and with the aim of fostering the language of the University of Valencia in research and education activity, this PhD dissertation starts with three abstracts in Spanish, Valencian and English, ends with a summary dissertation in Spanish and includes the complete versions of the journal publications.

[1]Reglament sobre depòsit, avaluació i defensa de la tesi doctoral aprovat pel Consell de Govern de 28 de Juny de 2016. ACGUV 172/2016.

[2]Pla d'increment de la docència en valencià (ACGUV 129/2012) aprovat i modificat pel Consell de Govern de 22 de desembre de 2016. ACGUV 308/2016.

Acknowledgements

En primer lugar, si te estas leyendo la tesis o piensas leértela, gracias. Supongo que es señal de que te importo, te interesa mi trabajo o quizá ambas cosas.

Escribir y depositar la tesis ha sido un proceso algo estresante, más por lo que ha sucedido alrededor que por la escritura en sí misma. Aun así, me ha servido para recopilar todo lo que he hecho hasta ahora, en qué punto estoy y quién me ha acompañado en el camino:

Gracias a todas las personas que participan en el mundo científico, publicando y revisando trabajos (algunas revisiones son buenas, otras no tanto...) de manera altruista, simplemente por contribuir al avance de la ciencia. Gracias también a quienes publican en abierto, compartiendo datos, código y recursos: gracias a gente así, la ciencia avanza más rápido.

Gracias a todos los investigadores que he conocido en estos años, de todos he aprendido algo. Gracias a los investigadores y profesores tanto del IPL como del departamento, aunque no erais mis tutores siempre que he necesitado algo me habéis ayudado. Gracias a la gente de Barcelona por el que fue mi primer congreso, gracias a la gente de Bristol, Raúl y Alex, por acogerme tan bien las veces que he ido. Gracias a la gente de Alemania por invitarnos a ir, espero que podamos trabajar juntos.

Gracias a todos los que han formado parte de mi proceso de aprendizaje y me han hecho llegar hasta aquí. A los profesores del colegio, en especial a Pere, por inculcarme la pasión por la física, las matemáticas y la ciencia.

Gracias a mis amigos, tanto a los de toda la vida como a los que fui encontrando en la universidad. Gracias por estar siempre ahí, por apoyarme, por escucharme hablar sin descanso de mi trabajo y, sobre todo, por ayudarme a desconectar. Por las fiestas, las cervezas, por salir conmigo en bici. Creo que algunas de las mejores ideas me han venido justo después de dejar de pensar un rato en la investigación.

Gracias a mis compañeros de laboratorio. Venir al IPL se hace mucho

más agradable cuando estáis aquí. Gracias Jorge, Nuria, Paula, Jose Manuel y Alexandra, por ayudarme siempre que lo he necesitado, por los ratos de café, por las tonterías que hacíamos (y seguimos haciendo) para desconectar y por los viajes juntos.

Gracias a mis padres por habérmelo dado todo. Siempre me habéis apoyado y guiado en todo. Gracias a vosotros soy la persona que soy. Son demasiadas las cosas por las que daros las gracias, así que lo resumo en una sola palabra: GRACIAS. Gracias también al resto de mi familia: a mi hermano, a mis abuelos y a todos los que me habéis apoyado siempre.

Gracias a Emma, por todo. Por ser tú, por estar siempre conmigo, por ayudarme y apoyarme. Sin ti no sería tan feliz. Por aguantarme en los peores momentos, incluso cuando ni yo mismo me aguanto. Por darme tu punto de vista de las cosas (que normalmente, no siempre, eh, suele ser el correcto). Por todo lo que estamos construyendo juntos y lo que aún nos queda por delante.

Gracias a mis tutores, Valero y Jesús. Habéis sido mucho más que simples directores de tesis. Me habéis enseñado a investigar, a escribir, a publicar y a mirar la ciencia con otros ojos. Gracias por ser como sois: por no aparentar, por intentar siempre que las cosas se hagan bien, por no publicar cualquier cosa y por ir siempre de frente.

Gracias, Jesús, por todos estos años en el IPL. Es imposible contar las horas de charlas, discusiones de café y pizarra, de programar juntos, que has invertido en mi. Estoy convencido de que ningún otro tutor habría hecho eso. Mucho de lo que he aprendido a lo largo del doctorado ha sido gracias a ti.

Gracias, Valero, por darme la oportunidad de hacer el doctorado. No estaría aquí sin ti. Gracias por aceptar que hiciera el TFM contigo y por todo el apoyo que siempre me has dado. Por guiarme desde mis primeros experimentos hasta hoy, por decirme qué ideas tenían sentido y cuáles no, y por todos los proyectos que escribiste para ayudarme. Pese a lo ocupado que estás, siempre has sacado un rato cuando lo he necesitado. Simplemente, gracias.

Contents

Acronyms

Throughout the thesis, there are words often written in their acronym form. Each acronym is always presented in parentheses the first time the word appears. However, here we present a curated list of all the acronyms that apppear in the thesis, sorted by their appearance position:

- **CNN**: **C**onvolutional **N**eural **N**etwork
- **ReLU**: **Re**ctified **L**inear **U**nit
- **ViT**: **Vi**sion **T**ransformer
- **CLIP**: **C**ontrastive **L**anguage **I**mage **P**re-training
- **VGG**: **V**isual **G**eometry **G**roup
- **DN**: **D**ivisive **N**ormalization
- **IoU**: **I**ntersection **o**ver **U**nion
- **CSF**: **C**ontrast **S**ensitivity **F**unction
- **RMSE**: **R**oot **M**ean **S**quared **E**rror
- **IQA**: **I**mage **Q**uality **A**ssessment
- **MOS**: **M**ean **O**pinion **S**core
- **SROCC**: **S**pearman **R**ank-**O**rder **C**orrelation **C**oeficient
- **SSIM**: **S**tructual **S**imilarity **I**ndex **M**easure
- **LPIPS**: **L**earned **P**erceptual **I**mage **P**atch **S**imilarity
- **DISTS**: **D**eep **I**mage **S**tructure and **T**exture **S**imilarity

Resumen en castellano

Introducción y contexto

El aprendizaje profundo ha transformado de manera radical el campo de la visión artificial durante las dos últimas décadas. Desde los primeros modelos basados en perceptrones y redes neuronales multicapa hasta la consolidación de las redes convolucionales (CNNs) y, más recientemente, los modelos basados en transformadores, la capacidad de los sistemas artificiales para procesar y reconocer patrones visuales ha crecido de forma extraordinaria. La disponibilidad de grandes conjuntos de datos, el aumento en la potencia computacional y los avances en algoritmos de entrenamiento han impulsado el rendimiento de los modelos hasta superar, en determinadas tareas, el nivel humano.

Sin embargo, este progreso ha estado acompañado de una pregunta persistente: ¿en qué medida estos modelos procesan la información visual de una manera comparable a la percepción humana? La visión humana es robusta, flexible y capaz de adaptarse a variaciones de los datos fácilmente, mientras que los modelos artificiales tienden a ser frágiles frente a cambios, ruido adversarial o correlaciones no deseadas en los datos de entrenamiento. Ejemplos paradigmáticos de esta fragilidad se encuentran en los denominados ataques adversariales: pequeñas perturbaciones invisibles para el ojo humano son suficientes para que un sistema de última generación clasifique erróneamente una imagen. Asimismo, se ha demostrado que muchas redes aprenden a apoyarse en atajos estadísticos, como la textura de un objeto o el contexto de la escena, en lugar de en características más abstractas y estables, como la forma global. Estas discrepancias ponen en evidencia que rendimiento en métricas estándar y comportamiento perceptual humano no son dimensiones equivalentes.

Por estas razones, la relación entre los sistemas artificiales y la visión

biológica es ambivalente. Por un lado, las redes neuronales profundas se inspiran directamente en principios neurocientíficos. Las CNNs recuperan la idea de campos receptivos descubiertos por Hubel y Wiesel en la corteza visual primaria; mecanismos como la normalización divisiva tienen su origen en estudios de la corteza visual temprana; y las arquitecturas basadas en atención se relacionan con fenómenos de focalización selectiva ampliamente descritos en la literatura neurocientífica y psicológica. Por otro lado, en muchos casos los nuevos modelos se desarrollan y optimizan siguiendo criterios empíricos de rendimiento más que de plausibilidad biológica. Este giro hacia el pragmatismo ha permitido grandes avances en conjuntos de datos de evaluación, pero también ha contribuido a que los mecanismos internos de los modelos se alejen de las estrategias perceptuales humanas, generando un desfase cada vez más evidente.

En este contexto surge el concepto de alineamiento humano, entendido como el grado en que un modelo artificial procesa y representa la información visual de manera comparable al sistema visual humano. Dicho alineamiento puede analizarse en varios niveles complementarios. En primer lugar, a nivel conductual, observando si los modelos responden de forma similar a los humanos ante estímulos visuales controlados. En segundo lugar, a nivel representacional, estudiando si la organización interna de las representaciones aprendidas refleja las distancias perceptuales humanas. Y, en tercer lugar, a nivel computacional, preguntándose si los cálculos implementados en las redes reproducen operaciones identificadas en la neurociencia, como mecanismos de ganancia adaptativa o sensibilidad al contraste. Esta triple perspectiva permite ir más allá de la precisión en tareas concretas para abordar la cuestión de fondo: ¿ven los modelos lo mismo que vemos los humanos?

La presente tesis se inscribe en este marco, con una visión doble. Por un lado, (1) evaluar si la incorporación de mecanismos inspirados en la biología puede mejorar la robustez y el rendimiento de los modelos de visión, ampliando su capacidad para generalizar más allá de las condiciones de entrenamiento. Por otro lado, (2) analizar hasta qué punto las representaciones y comportamientos de dichos modelos se alinean con los de la percepción humana, utilizando para ello protocolos inspirados en la psicofísica clásica. Al situarse en la intersección entre neurociencia, psicología experimental y aprendizaje profundo, este trabajo aspira a contribuir a una comprensión más rica de la relación entre sistemas artificiales y humanos, y a sentar las bases para el diseño de arquitecturas de visión artificial que no solo sean

precisas, sino también perceptualmente afines a nuestra experiencia visual.

Objetivos de la tesis

El trabajo de investigación se articula en torno a un conjunto de objetivos interrelacionados que buscan tender un puente entre el mundo de la visión biológica y los desarrollos recientes en aprendizaje profundo. Estos objetivos responden a una preocupación central: comprender si los mecanismos que dotan a la percepción humana de robustez, adaptabilidad y coherencia pueden trasladarse a arquitecturas artificiales, y en qué medida estas arquitecturas logran reproducir comportamientos perceptuales análogos a los nuestros.

El primer objetivo consiste en integrar cálculos biológicamente inspirados dentro de las arquitecturas de deep learning. En este sentido, la tesis se centra en la normalización divisiva (Divisive Normalization, DN), un mecanismo ampliamente documentado en el córtex visual temprano, especialmente en el área V1. La DN actúa como un proceso de control de ganancia que regula la respuesta neuronal en función de la actividad del entorno, garantizando eficiencia y estabilidad en la percepción humana. La investigación explora su implementación en modelos de segmentación como U-Net, analizando si la introducción de este mecanismo aporta ventajas prácticas en términos de robustez frente a condiciones visuales adversas como niebla, baja iluminación o reducción de contraste. Con ello se busca comprobar si la inspiración biológica puede traducirse en un mejor desempeño computacional sin sacrificar eficiencia.

El segundo objetivo aborda el desarrollo y aplicación de metodologías que permitan medir de manera rigurosa el alineamiento entre sistemas artificiales y percepción humana. Para ello, la tesis propone herramientas inspiradas en la psicofísica y en la ciencia de la visión, disciplinas que han estudiado durante décadas los fundamentos de la percepción mediante experimentos controlados. Se plantea la construcción de baterías de pruebas que evalúen propiedades de bajo nivel, como la sensibilidad al contraste o la discriminación cromática, junto con marcos que permitan analizar el grado de similitud en representaciones internas y respuestas conductuales. Este objetivo busca superar la dependencia exclusiva de métricas tradicionales de precisión o exactitud, aportando indicadores más cercanos a la experiencia perceptual humana.

Finalmente, el tercer objetivo consiste en analizar los factores que determinan el grado de alineamiento en modelos de aprendizaje profundo. El interés no se limita únicamente a las modificaciones arquitectónicas, sino que se extiende a elementos como las estadísticas de los conjuntos de datos de entrenamiento, los objetivos de optimización empleados, la duración de los procesos de aprendizaje, las técnicas de regularización y las estrategias de lectura e interpretación de las representaciones internas. Se pretende desentrañar qué condiciones favorecen o, por el contrario, dificultan la emergencia de propiedades comparables a la percepción humana. Este análisis ofrece una perspectiva amplia que permite comprender la alineación perceptual como un fenómeno multifactorial, resultado de la interacción entre arquitectura, datos y objetivos de aprendizaje.

En conjunto, estos objetivos marcan un itinerario de investigación que combina diseño arquitectónico, desarrollo metodológico y análisis crítico de factores determinantes. La meta última es avanzar hacia modelos de visión artificial que no solo destaquen por su precisión en distintos conjuntos de datos, sino que también se aproximen a la manera en que los seres humanos perciben, interpretan y responden al mundo visual.

Metodología

La estrategia metodológica de esta tesis se fundamenta en la combinación de tres aproximaciones complementarias que, en conjunto, permiten abordar el problema del alineamiento perceptual desde una perspectiva amplia y multidimensional. En lugar de limitarse al análisis de un único tipo de arquitectura o métrica de rendimiento, el trabajo propone un marco metodológico que integra el diseño de modelos bioinspirados, la creación de marcos de evaluación perceptual basados en la psicofísica y la comparación sistemática entre distintas arquitecturas de aprendizaje profundo. Esta combinación busca no solo medir la eficacia técnica de los modelos, sino también indagar en su grado de semejanza con la percepción humana.

En primer lugar, se desarrollaron modelos bioinspirados que incorporan mecanismos computacionales inspirados en la neurociencia visual. La estrategia más relevante en este ámbito fue la integración de la normalización divisiva en variantes de la arquitectura U-Net, ampliamente utilizada en tareas de segmentación de imágenes. Estas capas de normalización, diseñadas como módulos diferenciables dentro del marco del aprendizaje profundo,

fueron concebidas para emular propiedades funcionales del córtex visual temprano, particularmente aquellas relacionadas con el control de ganancia y la adaptación al contraste local. Para evaluar su impacto, se compararon modelos con y sin estas capas en contextos adversos que suelen representar un desafío tanto para la visión artificial como para la visión humana: escenas con niebla, condiciones de baja iluminación y variaciones significativas en el contraste. El propósito de este análisis fue doble. Por un lado, se pretendía determinar si la bioinspiración podía traducirse en una mayor robustez y generalización de los modelos; por otro, se evaluaba si estas modificaciones arquitectónicas eran capaces de inducir propiedades perceptuales más cercanas a las humanas, algo que trasciende la mera mejora en métricas de segmentación.

En segundo lugar, se desarrollaron marcos de evaluación perceptual específicamente diseñados para estudiar el grado de alineamiento entre los modelos de aprendizaje profundo y la percepción humana. Este aspecto metodológico constituye una de las aportaciones más originales de la tesis, ya que supone un alejamiento de los indicadores tradicionales de rendimiento y se adentra en la psicofísica como referente evaluativo. Se elaboró, en primer lugar, un decálogo de fenómenos psicofísicos de bajo nivel que incluye pruebas como la sensibilidad al contraste, el enmascaramiento contextual, la adaptación a estímulos repetidos y otros fenómenos fundamentales en el estudio de la visión. Este conjunto de experimentos permitió analizar en qué medida los modelos replican o divergen de los patrones observados en observadores humanos. A este marco se añadió la adaptación de las conocidas elipses de MacAdam al dominio de las redes neuronales, lo que posibilitó medir umbrales de discriminación cromática en los modelos y compararlos con los datos psicofísicos humanos. Complementariamente, se implementaron pruebas de funciones de sensibilidad al contraste (CSF) en modelos de lenguaje multimodal, que permiten estimar la capacidad de los modelos para responder a variaciones de frecuencia espacial y de contraste, sin necesidad de acceder directamente a sus representaciones internas. Finalmente, en el caso de CLIP, se propuso un enfoque por niveles de abstracción, diferenciando tareas de bajo, medio y alto nivel perceptual con el fin de analizar cómo evoluciona el alineamiento a lo largo de las capas y a medida que progresa el entrenamiento. En conjunto, estos marcos metodológicos constituyen una batería de herramientas de evaluación perceptual que complementan y amplían las métricas convencionales.

El tercer pilar metodológico consistió en llevar a cabo comparaciones sistemáticas entre distintas arquitecturas de redes neuronales. Se analizaron tanto modelos puramente visuales, como las redes convolucionales y los Vision Transformers, como modelos multimodales más recientes, tales como CLIP y los grandes modelos de lenguaje multimodal (MLLMs). En todos los casos, la evaluación se realizó en dos dimensiones complementarias: por un lado, se consideró el rendimiento en conjuntos de datos estándar de la literatura, lo que permitió situar los modelos en el estado del arte; por otro, se analizó su comportamiento en los marcos perceptuales descritos anteriormente, con el objetivo de determinar en qué medida las diferencias arquitectónicas, las estrategias de entrenamiento o los datos empleados influyen en el alineamiento perceptual con los humanos. Este enfoque comparativo permitió identificar qué factores —ya sean estructurales, relacionados con el conjunto de datos de entrenamiento o derivados de la regularización— favorecen o dificultan la convergencia entre la percepción artificial y la humana.

En resumen, la metodología adoptada en esta tesis responde a la necesidad de articular un enfoque integrador que combine bioinspiración, psicofísica y comparación entre arquitecturas. De esta manera, no solo se persigue avanzar en el diseño de modelos más robustos y precisos, sino también comprender en qué condiciones dichas arquitecturas se aproximan o se alejan de los procesos perceptuales humanos. Este marco metodológico constituye, por tanto, la base empírica y conceptual sobre la cual se construyen los resultados y conclusiones presentados en la investigación.

Resultados principales

La presente tesis doctoral ha producido una serie de resultados que, en conjunto, permiten comprender mejor la compleja relación entre la bioinspiración arquitectónica, la robustez computacional y el alineamiento perceptual entre redes neuronales profundas y el sistema visual humano. Los hallazgos se pueden sintetizar en varios ejes temáticos, aunque cada uno de ellos está interconectado y contribuye a la formulación de conclusiones de mayor alcance conceptual.

Bioinspiración y robustez

Uno de los primeros resultados relevantes se obtuvo en el ámbito del diseño de modelos bioinspirados. La incorporación de mecanismos de nor-

malización divisiva (DN) en arquitecturas de tipo U-Net mostró un impacto directo en la robustez frente a perturbaciones visuales. Este hallazgo resulta especialmente significativo porque confirma empíricamente lo que la literatura en neurociencia lleva tiempo sugiriendo: la normalización divisiva, ampliamente observada en la corteza visual temprana, constituye un mecanismo fundamental de control de ganancia y adaptación al contexto que permite a los organismos biológicos procesar información visual de manera estable en entornos cambiantes.

En el caso de los experimentos realizados en esta tesis, las U-Nets con capas DN demostraron un mejor rendimiento en imágenes degradadas por condiciones adversas tales como niebla, baja iluminación o alteraciones en el contraste. Frente a estas distorsiones, que suelen ser críticas para aplicaciones como la conducción autónoma o la vigilancia en entornos poco controlados, los modelos mejorados con DN mantuvieron una capacidad de segmentación significativamente más alta que sus homólogos estándar. Además, esta mejora no supuso un incremento notable en la complejidad computacional: el aumento en el número de parámetros fue marginal, lo que demuestra que la integración de mecanismos bioinspirados puede aportar beneficios sustanciales sin comprometer la escalabilidad de las arquitecturas modernas. En suma, estos resultados ponen de manifiesto que la transferencia de ideas provenientes de la biología al diseño de modelos artificiales no solo es conceptualmente coherente, sino también prácticamente eficaz para dotar a las redes de una mayor generalización frente a perturbaciones.

Falta de alineamiento perceptual

No obstante, el análisis detallado reveló un contraste llamativo entre robustez y alineamiento perceptual. Los mismos modelos bioinspirados que incorporaban capas de normalización divisiva, y que habían demostrado ventajas en entornos degradados, no lograron replicar fenómenos psicofísicos básicos cuando se evaluaron mediante el marco del Decálogo. Este conjunto de pruebas, diseñado para explorar propiedades de bajo nivel como la sensibilidad al contraste, la adaptación y el enmascaramiento contextual, mostró que las U-Nets con DN se desviaban significativamente de las curvas y patrones característicos de la percepción humana.

Por ejemplo, mientras que un observador humano presenta un umbral bien definido de sensibilidad a frecuencias espaciales y un comportamiento predecible frente a enmascaramiento lateral, los modelos analizados no repro-

ducían dichas regularidades. Su respuesta permanecía anclada a dinámicas puramente computacionales, sin reflejar la riqueza adaptativa de la percepción humana. Este resultado es clave porque muestra que mejorar la robustez de un modelo no equivale a hacerlo más humano. De hecho, un sistema puede ser capaz de resistir perturbaciones de manera eficiente sin que ello implique que sus estrategias internas se asemejen a las del sistema visual biológico. La conclusión que se desprende es que la bioinspiración, aunque útil, no garantiza por sí sola el alineamiento perceptual.

Color y estadística de los datos

El análisis de la discriminación cromática a través de las elipses de MacAdam ofreció otra perspectiva fundamental. Estos experimentos pusieron de relieve que la capacidad de un modelo para distinguir colores depende más de la riqueza cromática del conjunto de datos con el que ha sido entrenado que de la propia arquitectura. Dicho de otro modo, la estadística del entorno visual —representada por la diversidad cromática en los datos— constituye un factor determinante para la aproximación al comportamiento humano en tareas cromáticas.

Los resultados mostraron que redes entrenadas con datasets cromáticamente limitados producían elipses de discriminación distorsionadas y alejadas de las observadas en sujetos humanos. En cambio, aquellas redes entrenadas con distribuciones cromáticas más ricas y variadas generaban regiones de discriminación que se aproximaban mucho más a las elipses humanas clásicas. Este hallazgo refuerza la idea de que el alineamiento perceptual no solo se ve condicionado por la bioinspiración arquitectónica, sino también, y en gran medida, por las propiedades estadísticas del entorno de entrenamiento. El papel del dataset, por tanto, no se limita a proporcionar ejemplos suficientes para la generalización, sino que puede moldear directamente la estructura perceptual emergente en los modelos.

Multimodalidad y pérdida de alineamiento de bajo nivel

Los experimentos con modelos multimodales introducen una capa adicional de complejidad. En el caso de CLIP, se observó que en las fases iniciales del entrenamiento las representaciones se apoyaban fundamentalmente en texturas, lo que producía una mayor coincidencia con la percepción humana de bajo nivel. Sin embargo, a medida que el entrenamiento avanzaba y la supervisión lingüística adquiría un papel central, las representaciones

se desplazaban hacia formas globales y abstracciones semánticas. Este desplazamiento resultó ventajoso para la categorización y la generalización de alto nivel, pero supuso una pérdida de sensibilidad hacia los detalles locales, precisamente aquellos que caracterizan la percepción humana temprana.

Algo similar se constató en los grandes modelos de lenguaje multimodal (MLLMs). A pesar de sus sorprendentes capacidades de razonamiento y su éxito en tareas complejas, las pruebas de función de sensibilidad al contraste (CSF) revelaron limitaciones notables. Estos modelos no replicaban de manera fidedigna la sensibilidad humana a la frecuencia espacial ni al contraste, aunque en algunos casos llegaban a aproximarse cualitativamente a ciertos patrones. Ninguno, sin embargo, alcanzaba el nivel de fidelidad propio de la visión humana. Este hallazgo pone en evidencia un problema recurrente: los modelos multimodales son extraordinarios en razonamiento semántico y transferencia de conocimiento, pero carecen de los fundamentos perceptuales básicos que caracterizan la cognición visual humana.

Trade-off entre precisión y alineamiento

Finalmente, uno de los resultados más reveladores de la tesis fue la constatación de un compromiso intrínseco —o trade-off— entre precisión y alineamiento perceptual. Los experimentos demostraron que la relación entre el rendimiento de los modelos y su similitud con la percepción humana no es monótona, sino que adopta una forma de U invertida. En otras palabras, los modelos más grandes, entrenados durante más tiempo o con técnicas de regularización intensiva, lograban puntuaciones superiores en benchmarks convencionales, pero al mismo tiempo mostraban un menor grado de alineamiento con el comportamiento humano en tareas perceptuales.

Paradójicamente, arquitecturas más simples, como AlexNet, o representaciones intermedias dentro de CLIP, resultaban más cercanas a la percepción humana en dominios de bajo nivel. Esto sugiere que optimizar exclusivamente para la precisión puede alejar a los modelos de la estructura perceptual humana, mientras que configuraciones más modestas preservan, en mayor medida, regularidades compartidas con la visión biológica. El hallazgo invita a reconsiderar los objetivos del diseño de arquitecturas y plantea la necesidad de métricas complementarias a la precisión que tengan en cuenta la alineación perceptual como criterio de evaluación.

Discusión

Los resultados obtenidos a lo largo de esta tesis permiten abrir una reflexión amplia sobre las tensiones y paradojas que emergen en la relación entre bioinspiración, rendimiento técnico y alineamiento perceptual en modelos de visión artificial. Aunque a primera vista podría parecer que estos tres aspectos deberían converger de manera natural —pues se presupone que emular a la biología debería conducir a sistemas robustos y perceptualmente humanos—, la evidencia recopilada muestra que la realidad es mucho más matizada.

En primer lugar, se observa un claro desacoplamiento entre la plausibilidad biológica y el comportamiento perceptual humano. La integración de mecanismos como la normalización divisiva en arquitecturas de segmentación produjo una mejora tangible en la robustez frente a degradaciones visuales, confirmando que la bioinspiración puede tener un valor instrumental en términos de generalización y resistencia a perturbaciones. Sin embargo, esa misma bioinspiración no resultó suficiente para garantizar que los modelos reprodujeran fenómenos psicofísicos básicos, como la sensibilidad al contraste o el enmascaramiento contextual. Dicho de otra forma, se puede incrementar la funcionalidad sin necesariamente aproximarse a la fenomenología perceptual humana. Este hallazgo desafía la intuición inicial de que la implementación de módulos inspirados en la neurociencia conllevaría automáticamente un mayor alineamiento, y sugiere que la traducción de mecanismos biológicos a sistemas artificiales exige considerar no solo la arquitectura, sino también el contexto estadístico y la dinámica del aprendizaje.

En segundo lugar, la investigación confirma que el alineamiento humano es una propiedad emergente y multifactorial. No puede atribuirse exclusivamente a un único componente del modelo, ya sea la arquitectura, el conjunto de datos o el objetivo de entrenamiento. Más bien, surge de la interacción inseparable entre múltiples elementos: las estadísticas del entorno de entrenamiento, los objetivos de optimización seleccionados, la regularización aplicada, la escala del modelo y, en el caso de los sistemas multimodales, la naturaleza de la supervisión lingüística. Esta perspectiva resuena con el marco clásico de los niveles de análisis de Marr, donde se distinguen el nivel computacional (qué se hace y por qué), el nivel algorítmico (cómo se hace) y el nivel de implementación (con qué mecanismos se lleva a cabo). En la práctica, los resultados de esta tesis muestran que estos niveles no se mantienen aisla-

dos, sino que están profundamente entrelazados. Un cambio en la tarea o en los objetivos de optimización repercute en la forma en que se configuran las representaciones internas y, en consecuencia, en el alineamiento perceptual. Esto dificulta establecer relaciones lineales de causa-efecto y obliga a adoptar un enfoque sistémico para comprender el comportamiento de los modelos.

Otro hallazgo crucial es el fenómeno contraintuitivo según el cual una optimización excesiva puede alejar a los modelos de la percepción humana. A medida que se incrementa la capacidad del modelo, se prolonga el entrenamiento o se aplican técnicas avanzadas de regularización, las redes tienden a alcanzar un rendimiento superior en benchmarks clásicos de clasificación o segmentación, pero pierden similitud con la percepción humana en tareas de bajo nivel. Esta relación en forma de U invertida plantea un dilema fundamental: maximizar la precisión técnica puede implicar un coste en términos de alineamiento perceptual. En otras palabras, perseguir obsesivamente el estado del arte en métricas convencionales no solo no garantiza, sino que incluso puede obstaculizar, la aproximación a la percepción humana. Este hallazgo abre un debate de fondo sobre cuáles deberían ser los objetivos prioritarios en el desarrollo de sistemas de visión artificial: ¿deberíamos centrarnos únicamente en la eficacia medida en conjuntos de datos, o deberíamos dar mayor peso a la similitud con la cognición humana, incluso a costa de perder algunos puntos porcentuales de precisión?

El papel del lenguaje en los modelos multimodales constituye otro aspecto digno de discusión. Los resultados muestran que la supervisión lingüística, lejos de ser un complemento neutral, introduce un sesgo estructural en las representaciones internas de los modelos. Bajo esta influencia, las redes tienden a desplazar sus representaciones hacia abstracciones semánticas y formas globales, lo que incrementa la robustez y la capacidad de categorización en niveles altos de procesamiento, pero reduce la sensibilidad a fenómenos de bajo nivel como las texturas locales o las variaciones finas de contraste. Este hallazgo cuestiona la noción de que la multimodalidad constituya una vía directa hacia un mayor alineamiento con la percepción humana. Si bien el lenguaje enriquece las representaciones y potencia la capacidad de razonamiento, también modifica la "ecología perceptual" de los modelos, alejándolos de las características que son propias de la percepción visual temprana en humanos. En cierto sentido, lo que se gana en abstracción semántica se pierde en fidelidad perceptual básica, lo que plantea interrogantes sobre cómo equilibrar ambas dimensiones en futuros diseños de

modelos multimodales.

Finalmente, es importante resaltar que los resultados de esta tesis también invitan a replantear la relación entre robustez, interpretabilidad y alineamiento. Mientras que la robustez técnica puede lograrse a través de mecanismos bioinspirados o mediante el aumento de la escala del modelo, el alineamiento perceptual requiere marcos de evaluación más sofisticados, basados en la psicofísica y en comparaciones conductuales. La contribución de este trabajo no radica únicamente en haber mostrado divergencias entre los modelos y los humanos, sino en haber proporcionado metodologías sistemáticas que permiten medir, con cierto grado de objetividad, en qué puntos se aproximan y en qué puntos se separan. Esta capacidad de evaluación es clave, ya que sin ella el alineamiento perceptual quedaría relegado a una categoría anecdótica o subjetiva, mientras que aquí se consolida como un objetivo empírico y cuantificable.

Conclusiones

Al evaluar el desarrollo de esta tesis en relación con los objetivos inicialmente planteados, puede afirmarse que el trabajo ha logrado dar respuesta a cada uno de ellos, aunque los hallazgos obtenidos invitan a reflexionar más allá de lo estrictamente propuesto. El primer objetivo consistía en explorar si la incorporación de mecanismos bioinspirados, en particular la normalización divisiva, podía mejorar el rendimiento de las redes neuronales profundas en tareas de segmentación. Los resultados obtenidos confirman que esta estrategia efectivamente aporta un valor añadido: las arquitecturas enriquecidas con DN demostraron una mayor robustez frente a degradaciones visuales como niebla, baja iluminación o variaciones de contraste. Esto constituye una prueba empírica de que la bioinspiración puede contribuir de forma directa a aumentar la generalización y la estabilidad de los modelos. Sin embargo, también quedó claro que la mejora en rendimiento no se traduce automáticamente en un mayor alineamiento perceptual. Dicho de otro modo, la robustez técnica y la similitud con el comportamiento humano son dimensiones relacionadas pero independientes, y no basta con incorporar un mecanismo neuronal conocido para garantizar que una red artificial se comporte como un observador humano.

El segundo objetivo planteaba el desarrollo de metodologías que permitiesen evaluar de manera más rica y granular el grado de alineamiento entre

modelos y percepción humana. En este aspecto, la tesis aporta una contribución sustancial. Se diseñaron y aplicaron distintos marcos de evaluación inspirados en la psicofísica, tales como el Decálogo de fenómenos de bajo nivel, la adaptación de las elipses de MacAdam para estimar la discriminación cromática, la evaluación de la función de sensibilidad al contraste en modelos multimodales y el análisis del alineamiento a distintos niveles de abstracción en CLIP. Estos instrumentos permiten no solo diagnosticar de manera más precisa las similitudes y divergencias entre visión artificial y humana, sino también comparar arquitecturas heterogéneas bajo un mismo prisma evaluativo. En este sentido, el trabajo no se limita a identificar fallos o aciertos en modelos concretos, sino que establece un marco metodológico que puede ser reutilizado y ampliado en investigaciones futuras.

El tercer objetivo buscaba esclarecer qué factores son determinantes en la emergencia del alineamiento perceptual. La evidencia reunida demuestra que este fenómeno es multifactorial y que, lejos de depender exclusivamente de la arquitectura, está condicionado por la interacción entre múltiples elementos. La naturaleza de los datos de entrenamiento y los objetivos de optimización resultaron ser más influyentes que la mera elección de bloques arquitectónicos. El caso de la discriminación cromática es paradigmático: lo que más acercó a los modelos a la percepción humana no fue la introducción de mecanismos de normalización divisiva, sino la riqueza cromática del entorno visual contenido en los datos. De forma más amplia, el trabajo muestra que el alineamiento perceptual es un producto emergente de la interacción entre arquitectura, datos, objetivos y supervisión, y que ningún factor aislado resulta suficiente para explicarlo por completo.

En conjunto, la tesis ofrece aportaciones en tres planos complementarios. Desde el punto de vista empírico, aporta resultados concretos sobre la robustez en segmentación, la discriminación cromática y el comportamiento de modelos multimodales, iluminando así diferentes facetas de la relación entre visión humana y artificial. En el plano metodológico, introduce marcos de evaluación novedosos inspirados en la psicofísica, que abren la posibilidad de establecer criterios más finos y rigurosos para medir la alineación. Y en el plano conceptual, invita a replantear la clásica separación de niveles de análisis propuesta por Marr, mostrando que en la práctica del aprendizaje profundo estos niveles —tarea, algoritmo e implementación— aparecen entrelazados y no pueden estudiarse de manera independiente sin perder de vista su mutua interdependencia.

Las conclusiones generales de la tesis, por tanto, no se limitan a validar los objetivos iniciales, sino que ponen de manifiesto la necesidad de replantear los criterios de diseño y evaluación de los modelos de visión artificial. La robustez, la precisión y el alineamiento perceptual no son metas que se alcancen por los mismos caminos, y comprender sus tensiones y compromisos es esencial para avanzar hacia sistemas verdaderamente humanocéntricos.

Perspectivas futuras

Los hallazgos de esta tesis doctoral abren múltiples vías de investigación que pueden orientar el trabajo futuro en el ámbito de la visión artificial y el alineamiento perceptual. Una de las direcciones más prometedoras consiste en el diseño de objetivos de entrenamiento híbridos que no se limiten a maximizar la precisión en benchmarks tradicionales, sino que integren explícitamente criterios de semejanza con la percepción humana. Esto podría materializarse a través de pérdidas perceptuales derivadas de juicios humanos, bases de datos anotadas con evaluaciones psicofísicas o, más recientemente, mediante técnicas de optimización directa de preferencias, como la Direct Preference Optimization que ya se ha explorado en el campo del lenguaje natural. La incorporación de estos criterios permitiría examinar si es posible superar el dilema, identificado en esta tesis, entre la precisión técnica y la similitud perceptual, y abriría la puerta a modelos que no solo acierten más, sino que también "vean" de un modo más parecido al humano.

Otra línea de desarrollo natural es la extensión del estudio de mecanismos bioinspirados a arquitecturas más recientes y dominantes en el panorama actual. En particular, los Vision Transformers y los modelos multimodales de gran escala ofrecen un terreno fértil para experimentar con la integración de normalización divisiva u otros cálculos inspirados en la neurociencia. La pregunta aquí no es únicamente si tales mecanismos pueden mejorar la robustez en tareas específicas, sino también si su incorporación en arquitecturas caracterizadas por el autoatentido y la multimodalidad introduce formas más humanas de procesamiento perceptual. En este sentido, evaluar la utilidad de la bioinspiración en un contexto donde el sesgo del lenguaje y la abstracción semántica juegan un papel tan central permitirá comprender mejor los límites y el alcance de la transferencia de principios biológicos al aprendizaje profundo.

Del mismo modo, resulta esencial seguir avanzando en el diseño de métricas de alineamiento perceptual que sean independientes de la estrategia de lec-

tura de las representaciones internas de los modelos. La experiencia de esta tesis muestra que, en muchos casos, la comparación capa a capa no es suficiente para capturar la complejidad del alineamiento. Sería necesario proponer métricas inspiradas en paradigmas psicofísicos clásicos —sensibilidad al contraste, discriminación cromática, ilusiones visuales— pero formuladas de manera que puedan aplicarse a modelos como los MLLMs, donde no siempre es posible acceder de manera transparente a las activaciones internas. Estas métricas ofrecerían medidas comparables y robustas incluso en sistemas de caja negra, y permitirían establecer estándares más homogéneos en la evaluación de modelos de distinta naturaleza.

Además, el marco de evaluación debería expandirse hacia fenómenos perceptivos de nivel medio y alto, que hasta ahora han recibido menos atención. Aspectos como la percepción de la forma global, la integración contextual o incluso la sensibilidad a ilusiones visuales podrían constituir pruebas valiosas para analizar hasta qué punto los modelos logran capturar las regularidades de la percepción humana más allá de los dominios de bajo nivel. Incorporar estas dimensiones ayudaría a cerrar la brecha entre la evaluación de características sensoriales básicas y la comprensión de procesos cognitivos más abstractos, consolidando el alineamiento perceptual como un objetivo transversal en todas las capas y niveles de representación.

Finalmente, los resultados aquí presentados refuerzan la necesidad de que el alineamiento perceptual pase a formar parte de los estándares de evaluación en visión artificial. Para aplicaciones críticas en las que los modelos interactúan directamente con seres humanos —como la conducción autónoma, la medicina asistida por IA o los sistemas de seguridad—, no basta con alcanzar un alto rendimiento en un conjunto de datos estático. Es indispensable que las métricas de éxito reflejen también la similitud con los procesos perceptuales humanos, garantizando que los sistemas artificiales operen de un modo más seguro, interpretable y confiable. En este sentido, la tesis ofrece tanto evidencias empíricas como herramientas metodológicas que pueden servir de base para futuras iniciativas orientadas a institucionalizar el alineamiento perceptual como criterio de referencia. Avanzar en esta dirección no solo contribuiría a mejorar la calidad científica del campo, sino que también fortalecería la confianza social en los sistemas de inteligencia artificial que forman parte cada vez más activa de nuestro entorno cotidiano.

Justificación de la unidad temática de los artículos

El conjunto de artículos que conforman esta tesis doctoral responde a una misma preocupación central: comprender hasta qué punto los modelos de visión artificial se aproximan a la percepción humana y bajo qué condiciones emerge o se debilita dicho alineamiento. Aunque cada publicación aborda esta cuestión desde un ángulo específico —ya sea el diseño de modelos bioinspirados, la elaboración de marcos de evaluación psicofísica o el análisis comparativo de arquitecturas y factores de entrenamiento—, todas comparten el hilo conductor de explorar la interacción entre rendimiento técnico, plausibilidad biológica y semejanza perceptual.

La primera línea de trabajos se centró en introducir y evaluar mecanismos bioinspirados, como la normalización divisiva, en arquitecturas de segmentación. Estos estudios no solo aportaron evidencias de que la bioinspiración puede mejorar la robustez y la generalización en condiciones adversas, sino que también establecieron la base experimental para contrastar si tales mecanismos inducen o no comportamientos perceptualmente humanos.

Una segunda serie de publicaciones se orientó a la creación de marcos de evaluación perceptual, inspirados en la psicofísica, con el fin de disponer de herramientas que permitieran medir el alineamiento de manera sistemática. El Decálogo de fenómenos de bajo nivel, la adaptación de las elipses de MacAdam y las pruebas de sensibilidad al contraste en modelos multimodales constituyen aportaciones metodológicas que trascienden los modelos concretos y ofrecen un repertorio aplicable a futuros trabajos en la disciplina.

Finalmente, un tercer bloque de artículos abordó la comparación sistemática entre diferentes arquitecturas —desde CNNs y Vision Transformers hasta CLIP y grandes modelos multimodales— con el objetivo de identificar qué factores determinan en mayor medida la proximidad a la percepción humana. Estos análisis permitieron mostrar que la alineación no es exclusiva de un diseño arquitectónico particular, sino que depende de la interacción multifactorial entre datos, objetivos de entrenamiento y supervisión.

En conjunto, los distintos artículos no constituyen contribuciones aisladas, sino piezas complementarias de una investigación unificada. Cada publicación aborda un aspecto parcial del problema, pero todas convergen en una conclusión común: el alineamiento perceptual debe entenderse como una propiedad emergente que requiere tanto innovaciones en el diseño de modelos como marcos de evaluación inspirados en la percepción humana.

Esta coherencia temática justifica el formato de compendio y refuerza el valor integrador de la tesis, que ofrece una visión global sobre cómo acercar la visión artificial a la humana desde perspectivas empíricas, metodológicas y conceptuales.

Abstract

This thesis addresses a fundamental problem in contemporary computer vision: the question of perceptual alignment, specifically investigating the extent to which deep neural networks perceive and interpret the visual world in a manner comparable to human observers. While deep learning has catalyzed revolutionary advances in the field, achieving state-of-the-art performance across a wide array of tasks, including image segmentation, object classification, and complex multimodal reasoning, the connection between these artificial systems and the mechanisms underlying human visual perception remains notably limited. Despite impressive accuracy metrics, deep neural networks often exhibit behaviors that diverge from human perception, particularly under conditions of visual ambiguity or in the presence of subtle contextual cues. This work investigates the interplay between biological plausibility, computational performance, and perceptual alignment, employing a combination of bio-inspired architectural modifications and psychophysically grounded evaluation protocols to rigorously quantify the similarities and divergences between machine and human vision.

The first part of the thesis focuses on biologically inspired computational mechanisms, examining their capacity to enhance robustness while maintaining computational efficiency. Specifically, we investigate the integration of divisive normalization (DN), a canonical computation observed in the early visual cortex, into state-of-the-art segmentation architectures, including variants of the widely used U-Net. Divisive normalization serves as a canonical gain-control mechanism that modulates neural responses based on local contrast, and it has been implicated in numerous low-level perceptual phenomena observed in human vision. Experimental results demonstrate that models incorporating DN exhibit increased robustness under adverse environmental conditions, such as fog, low lighting, or reduced contrast, achieving improved segmentation performance with only minimal increases in model complexity

or parameter count. However, when these biologically inspired models are evaluated using the Decalogue, a rigorously designed battery of psychophysical tests assessing low-level visual phenomena, including contrast sensitivity and contextual masking, they fail to reproduce human-like perceptual behaviors. These findings suggest a nuanced conclusion: while biologically inspired computations can improve robustness and generalization, they do not inherently induce human-like perceptual characteristics. This underscores the distinction between improving task performance and achieving genuine perceptual alignment with human observers.

The second part of the thesis develops systematic methodologies to measure perceptual alignment beyond traditional accuracy metrics, moving toward behaviorally and psychophysically informed evaluation frameworks. This includes developing the Decalogue for low-level phenomena, devising novel procedures for assessing chromatic discrimination via MacAdam ellipses, evaluating contrast sensitivity function (CSF) responses in multimodal language models (MLLMs), and establishing a framework to quantify abstraction levels in vision-language models such as CLIP. The empirical results obtained through these methodologies reveal critical insights into the factors that influence alignment. For instance, neural networks trained on richer chromatic distributions generate discrimination ellipses that more closely approximate human color perception, highlighting the importance of the visual environment and data diversity. CSF evaluations reveal that even advanced MLLMs exhibit marked limitations in reproducing basic human sensitivities to spatial frequency, suggesting persistent gaps in low-level perceptual fidelity. In CLIP, alignment is found to vary across network layers: early layers, which encode primarily texture-based information, exhibit moderate alignment with human perception, whereas later layers, influenced by linguistic supervision, increasingly abstract visual representations toward semantic concepts. This abstraction enhances model robustness and task generalization but diminishes alignment with low-level human perceptual behaviors.

The third part of the thesis investigates the broader determinants of perceptual alignment. Through systematic analyses across convolutional neural networks (CNNs), Vision Transformers, CLIP, and multimodal language models, the work demonstrates that alignment is a multifactorial property emerging from complex interactions between architectural design, optimization objectives, statistical properties of the training data, duration of training, and reading strategies. Interestingly, the relationship between task per-

formance and perceptual alignment is non-monotonic: increasing model capacity or optimizing solely for accuracy can paradoxically reduce alignment with human perception, resulting in an inverted U-shaped relationship between accuracy and perceptual similarity. Additionally, linguistic supervision biases models toward global shape representations at the expense of local texture information, emphasizing that the type of task and supervision can play a more substantial role than architectural choices alone. These findings suggest that perceptual alignment is more strongly constrained by the combination of data, supervision, and task demands than by modifications to network architecture.

Conceptually, this thesis contributes to a deeper understanding of the interplay between performance, biological inspiration, and perceptual alignment, highlighting that improvements in accuracy or biologically motivated design do not necessarily translate to human-like perceptual behavior. Methodologically, it introduces systematic evaluation frameworks inspired by psychophysics, which can be applied to both vision-only and multimodal models to assess alignment rigorously. Empirically, it clarifies how factors such as early visual computations, chromatic environmental richness, optimization regimes, and language-based supervision interact to influence the degree of similarity between artificial and human perception.

In conclusion, this thesis advances the understanding of how artificial neural networks perceive visual stimuli and delineates the conditions under which they diverge from human visual experience. It provides strong evidence that bridging the gap between computational performance and perceptual alignment requires moving beyond architectural inspiration, toward evaluation frameworks and design principles that are explicitly informed by human behavioral and psychophysical data. These contributions lay the foundation for future research on biologically inspired, robust architectures and establish perceptual alignment as a critical, complementary objective to accuracy in the development and evaluation of computer vision systems. By integrating insights from neuroscience, psychophysics, and machine learning, the work positions perceptual alignment as a central consideration for designing artificial vision systems capable of functioning in real-world, human-centered environments.

Resumen

Esta tesis aborda un problema fundamental en la visión por computadora contemporánea: la cuestión de la alineación perceptual, investigando específicamente en qué medida las redes neuronales profundas perciben e interpretan el mundo visual de una manera comparable a los observadores humanos. Si bien el aprendizaje profundo ha provocado avances revolucionarios en el campo—alcanzando un rendimiento de vanguardia en tareas como segmentación de imágenes, clasificación de objetos y razonamiento multimodal complejo—la conexión entre estos sistemas artificiales y los mecanismos subyacentes a la percepción visual humana sigue siendo notablemente limitada. A pesar de métricas de precisión impresionantes, las redes neuronales profundas a menudo exhiben comportamientos que divergen de la percepción humana, especialmente en condiciones de ambigüedad visual o en presencia de señales contextuales sutiles. Este trabajo investiga la interacción entre plausibilidad biológica, rendimiento computacional y alineación perceptual, utilizando una combinación de modificaciones arquitectónicas inspiradas en la biología y protocolos de evaluación fundamentados en la psicofísica para cuantificar rigurosamente las similitudes y divergencias entre la visión humana y la artificial.

La primera parte de la tesis se centra en los mecanismos computacionales inspirados en la biología, examinando su capacidad para mejorar la robustez sin comprometer la eficiencia computacional. Específicamente, se investiga la integración de la normalización divisiva (ND), un cálculo canónico observado en la corteza visual temprana, en arquitecturas de segmentación de última generación, incluyendo variantes de la ampliamente utilizada U-Net. La normalización divisiva actúa como un mecanismo de control de ganancia que modula la respuesta neuronal en función del contraste local y se ha implicado en numerosos fenómenos perceptuales de bajo nivel observados en la visión humana. Los resultados experimentales muestran que los modelos que in-

corporan ND presentan una mayor robustez frente a condiciones ambientales adversas, como niebla, baja iluminación o contraste reducido, mejorando el desempeño de segmentación con solo un aumento mínimo en la complejidad o en la cantidad de parámetros del modelo. Sin embargo, cuando estos modelos inspirados en la biología se evalúan utilizando el Decálogo, un conjunto rigurosamente diseñado de pruebas psicofísicas que evalúan fenómenos visuales de bajo nivel, incluyendo sensibilidad al contraste y enmascaramiento contextual, no logran reproducir comportamientos perceptuales similares a los humanos. Estos hallazgos sugieren una conclusión matizada: aunque los cálculos inspirados en la biología pueden mejorar la robustez y la generalización, no inducen necesariamente características perceptuales humanas, lo que subraya la diferencia entre mejorar el rendimiento de la tarea y alcanzar una alineación perceptual genuina con los observadores humanos.

La segunda parte de la tesis desarrolla metodologías sistemáticas para medir la alineación perceptual más allá de las métricas tradicionales de precisión, avanzando hacia marcos de evaluación fundamentados en el comportamiento y la psicofísica. Esto incluye el desarrollo del Decálogo para fenómenos de bajo nivel, el diseño de nuevos procedimientos para evaluar la discriminación cromática mediante las elipses de MacAdam, la evaluación de las funciones de sensibilidad al contraste (CSF) en modelos de lenguaje multimodal (MLLM) y el establecimiento de un marco para cuantificar los niveles de abstracción en modelos de visión y lenguaje como CLIP. Los resultados empíricos obtenidos mediante estas metodologías revelan información clave sobre los factores que influyen en la alineación. Por ejemplo, las redes neuronales entrenadas con distribuciones cromáticas más ricas generan elipses de discriminación que se aproximan más a la percepción humana del color, lo que destaca la importancia del entorno visual y la diversidad de los datos. Las evaluaciones de CSF muestran que incluso los MLLM más avanzados presentan limitaciones significativas para reproducir sensibilidades humanas básicas a la frecuencia espacial, indicando brechas persistentes en la fidelidad perceptual de bajo nivel. En CLIP, la alineación varía a lo largo de las capas de la red: las capas iniciales, que codifican información principalmente basada en texturas, muestran una alineación moderada con la percepción humana, mientras que las capas posteriores, influenciadas por la supervisión lingüística, abstraen las representaciones visuales hacia conceptos semánticos. Esta abstracción mejora la robustez y la generalización de la tarea, pero disminuye la alineación con los comportamientos perceptuales humanos de bajo

nivel.

La tercera parte de la tesis investiga los determinantes más amplios de la alineación perceptual. A través de análisis sistemáticos de redes neuronales convolucionales (CNN), transformadores de visión, CLIP y modelos de lenguaje multimodal, se demuestra que la alineación es una propiedad multifactorial que emerge de la interacción compleja entre el diseño arquitectónico, los objetivos de optimización, las propiedades estadísticas de los datos de entrenamiento, la duración del entrenamiento y las estrategias de lectura. De manera interesante, la relación entre el rendimiento en la tarea y la alineación perceptual es no monótona: aumentar la capacidad del modelo o optimizar exclusivamente para la precisión puede reducir, paradójicamente, la alineación con la percepción humana, resultando en una relación en forma de U invertida entre precisión y similitud perceptual. Además, la supervisión lingüística sesga los modelos hacia representaciones globales de forma a expensas de la información de textura local, enfatizando que el tipo de tarea y la supervisión pueden tener un papel más sustancial que las elecciones arquitectónicas. Estos hallazgos sugieren que la alineación perceptual está más fuertemente determinada por la combinación de datos, supervisión y demandas de la tarea que por modificaciones arquitectónicas.

Conceptualmente, esta tesis contribuye a una comprensión más profunda de la interacción entre rendimiento, inspiración biológica y alineación perceptual, destacando que las mejoras en precisión o en el diseño inspirado biológicamente no se traducen necesariamente en comportamientos perceptuales humanos. Metodológicamente, introduce marcos sistemáticos de evaluación inspirados en la psicofísica, aplicables tanto a modelos solo de visión como multimodales para evaluar la alineación de manera rigurosa. Empíricamente, clarifica cómo factores como cálculos visuales tempranos, riqueza cromática del entorno, regímenes de optimización y supervisión lingüística interactúan para influir en el grado de similitud entre percepción artificial y humana.

En conclusión, esta tesis avanza en la comprensión de cómo las redes neuronales artificiales perciben estímulos visuales y delimita las condiciones bajo las cuales divergen de la experiencia visual humana. Proporciona evidencia sólida de que cerrar la brecha entre rendimiento computacional y alineación perceptual requiere ir más allá de la inspiración arquitectónica, hacia marcos de evaluación y principios de diseño explícitamente informados

por datos conductuales y psicofísicos humanos. Estas contribuciones sientan las bases para investigaciones futuras sobre arquitecturas robustas e inspiradas biológicamente y establecen la alineación perceptual como un objetivo crítico y complementario a la precisión en el desarrollo y evaluación de sistemas de visión por computadora. Al integrar conocimientos de neurociencia, psicofísica y aprendizaje automático, el trabajo posiciona la alineación perceptual como una consideración central para el diseño de sistemas de visión artificial capaces de operar en entornos del mundo real centrados en el ser humano.

Resum

Esta tesi aborda un problema fonamental en la visió per a ordinador contemporània: la qüestió de l'alineació perceptual, investigant específicament fins a quin punt les xarxes neuronals profundes perceben i interpreten el món visual d'una manera comparable als observadors humans. Tot i que l'aprenentatge profund ha provocat avanços revolucionaris en el camp, aconseguint un rendiment de punta en tasques com segmentació d'imatges, classificació d'objectes i raonament multimodal complex, la connexió entre aquests sistemes artificials i els mecanismes subjacents a la percepció visual humana continua sent notablement limitada. Malgrat les mètriques de precisió impressionants, les xarxes neuronals profundes sovint mostren comportaments que divergeixen de la percepció humana, especialment en condicions d'ambigüitat visual o en presència de senyals contextuals subtils. Aquest treball investiga la interacció entre la plausibilitat biològica, el rendiment computacional i l'alineació perceptual, utilitzant una combinació de modificacions arquitectòniques inspirades en la biologia i protocols d'avaluació fonamentats en la psicofísica per a quantificar rigorosament les similituds i divergències entre la visió humana i l'artificial.

La primera part de la tesi se centra en els mecanismes computacionals inspirats en la biologia, examinant la seua capacitat per a millorar la robustesa sense comprometre l'eficiència computacional. Concretament, s'investiga la integració de la normalització divisiva (ND), un càlcul canònic observat en la cortexa visual primerenca, en arquitectures de segmentació de última generació, incloent variants de la àmpliament utilitzada U-Net. La normalització divisiva actua com un mecanisme de control de guany que modula la resposta neuronal en funció del contrast local, i s'ha implicat en nombrosos fenòmens perceptuals de baix nivell observats en la visió humana. Els resultats experimentals mostren que els models que incorporen ND presenten una major robustesa davant condicions ambientals adverses, com boira, baixa il·luminació

o contrast reduït, millorant el rendiment de segmentació amb només un augment mínim en la complexitat o en el nombre de paràmetres del model. No obstant això, quan aquests models inspirats en la biologia s'avaluen utilitzant el Decàleg, un conjunt rigorosament dissenyat de proves psicofísiques que avaluen fenòmens visuals de baix nivell, incloent la sensibilitat al contrast i l'enmascarament contextual, no aconsegueixen reproduir comportaments perceptuals similars als humans. Aquests resultats suggereixen una conclusió matisada: tot i que els càlculs inspirats en la biologia poden millorar la robustesa i la generalització, no induïxen necessàriament característiques perceptuals humanes, destacant la diferència entre millorar el rendiment de la tasca i aconseguir una alineació perceptual genuïna amb els observadors humans.

La segona part de la tesi desenvolupa metodologies sistemàtiques per a mesurar l'alineació perceptual més enllà de les mètriques tradicionals de precisió, avançant cap a marcs d'avaluació fonamentats en el comportament i la psicofísica. Això inclou el desenvolupament del Decàleg per a fenòmens de baix nivell, el disseny de nous procediments per a avaluar la discriminació cromàtica mitjançant els el·lipses de MacAdam, l'avaluació de les funcions de sensibilitat al contrast (CSF) en models de llenguatge multimodal (MLLM) i l'establiment d'un marc per a quantificar els nivells d'abstracció en models de visió i llenguatge com CLIP. Els resultats empírics obtinguts mitjançant aquestes metodologies revelen informació clau sobre els factors que influeixen en l'alineació. Per exemple, les xarxes neuronals entrenades amb distribucions cromàtiques més riques generen el·lipses de discriminació que s'aproximen més a la percepció humana del color, destacant la importància de l'entorn visual i la diversitat de dades. Les avaluacions de CSF mostren que fins i tot els MLLM més avançats presenten limitacions significatives per a reproduir sensibilitats humanes bàsiques a la freqüència espacial, indicant bretxes persistents en la fidelitat perceptual de baix nivell. En CLIP, l'alineació varia al llarg de les capes de la xarxa: les capes inicials, que codifiquen informació principalment basada en textures, mostren una alineació moderada amb la percepció humana, mentre que les capes posteriors, influenciades per la supervisió lingüística, abstrauen les representacions visuals cap a conceptes semàntics. Aquesta abstracció millora la robustesa i la generalització de la tasca, però disminueix l'alineació amb els comportaments perceptuals humans de baix nivell.

La tercera part de la tesi investiga els determinants més generals de

l'alineació perceptual. Mitjançant anàlisis sistemàtics de xarxes neuronals convolucionals (CNN), transformadors de visió, CLIP i models de llenguatge multimodal, es demostra que l'alineació és una propietat multifactorial que emergeix de la interacció complexa entre el disseny arquitectònic, els objectius d'optimització, les propietats estadístiques de les dades d'entrenament, la durada de l'entrenament i les estratègies de lectura. De manera interessant, la relació entre el rendiment en la tasca i l'alineació perceptual es no monòtona: augmentar la capacitat del model o optimitzar exclusivament per a la precisió pot reduir, de manera paradoxal, l'alineació amb la percepció humana, resultant en una relació en forma de U invertida entre precisió i similitud perceptual. A més, la supervisió lingüística sesga els models cap a representacions globals de forma a costa de la informació de textura local, emfatitzant que el tipus de tasca i la supervisió poden tenir un paper més substancial que les decisions arquitectòniques. Aquests resultats suggereixen que l'alineació perceptual està més fortament determinada per la combinació de dades, supervisió i exigències de la tasca que per modificacions arquitectòniques.

Conceptualment, aquesta tesi contribueix a una comprensió més profunda de la interacció entre rendiment, inspiració biològica i alineació perceptual, destacant que les millores en precisió o en el disseny inspirat en la biologia no es tradueixen necessàriament en comportaments perceptuals humans. Metodològicament, introdueix marcs sistemàtics d'avaluació inspirats en la psicofísica, aplicables tant a models només de visió com multimodals per avaluar l'alineació de manera rigorosa. Empíricament, clarifica com factors com càlculs visuals primerencs, riquesa cromàtica de l'entorn, regims d'optimització i supervisió lingüística interactuen per influir en el grau de similitud entre percepció artificial i humana.

En conclusió, aquesta tesi avança en la comprensió de com les xarxes neuronals artificials perceben estímuls visuals i delimita les condicions sota les quals divergeixen de l'experiència visual humana. Proporciona evidència sòlida que tancar la bretxa entre rendiment computacional i alineació perceptual requereix anar més enllà de la inspiració arquitectònica, cap a marcs d'avaluació i principis de disseny explícitament informats per dades conductuals i psicofísiques humanes. Aquestes contribucions establixen les bases per a investigacions futures sobre arquitectures robustes i inspirades en la biologia i posicionen l'alineació perceptual com un objectiu crític i complementari a la precisió en el desenvolupament i avaluació de sistemes de visió per a or-

dinador. Integrant coneixements de neurociència, psicofísica i aprenentatge automàtic, el treball situa l'alineació perceptual com una consideració central per al disseny de sistemes de visió artificial capaços d'operar en entorns del món real centrats en l'ésser humà.

Part I

Introduction, Objectives and Thesis Structure

1

Introduction

1.1 General Context

From its inception, deep learning, based on artificial neural networks, has been profoundly inspired by the architecture and information processing of the human brain [McCulloch and Pitts, 1943, Rosenblatt, 1958, Fukushima, 1980]. This connection is particularly evident in the field of computer vision, where convolutional neural networks (CNNs) were directly inspired by biological vision systems. In particular, their design was influenced by the pioneering studies of Hubel and Wiesel in the late 1950s, who characterized the receptive fields of neurons in the primary visual cortex [Hubel et al., 1959, Hubel and Wiesel, 1962]. These studies, for which they won the Nobel Prize in 1981, revealed how early visual neurons respond to oriented edges and spatial patterns in a convolutional way, an idea that has lately been included in the hierarchical structure of CNNs. These ideas of hierarchical structures, sensitivity to local visual features and progressive complexity across layers were some of the core ideas behind the neocognitron model [Fukushima, 1980], considered the direct precursor of modern CNNs.

However, it was not until several decades later that training deep neural networks became truly effective, thanks to the error backpropa-

gation algorithm. Although the principle of gradient descent is older [Cauchy et al., 1847], backpropagation was first formalized for neural networks in the 1970s [Werbos, 1974] and later popularized in the 1980s [Rumelhart et al., 1986]. This algorithm enabled efficient weight adjustment via gradient descent, laying the foundation for supervised learning in deep architectures. Early examples such as LeNet [LeCun et al., 1989] demonstrated the feasibility of applying convolutional neural networks to visual recognition tasks, such as automatic digit recognition, still with the limitations imposed by the computational constraints of that time.

The true breakthrough of deep learning in computer vision came in 2012, with the introduction of AlexNet [Krizhevsky et al., 2012], the first deep neural network to win the ImageNet image classification challenge [Deng et al., 2009] with a significantly lower error rate than any previous method. AlexNet marked a turning point in the field, combining a hierarchical convolutional architecture with modern training strategies such as Rectified Linear Unit (ReLU) activations [Nair and Hinton, 2010, Agarap, 2018], GPU-parallelization training, and dropout regularization [Hinton et al., 2012]. In addition to its hierarchical structure inspired by the receptive fields discovered by Hubel and Wiesel, AlexNet also incorporated another brain-inspired computation, the Local Response Normalization. This computation is a mechanism of local activity normalization, where the normalization of a single neuron takes also into account its surroundings. It emulates the Divisive Normalization model observed in the human brain [Carandini and Heeger, 1994, Carandini et al., 2005, Carandini and Heeger, 2012]. Indeed, the authors identified this normalization mechanism as one of the two most crucial model architectural components, alongside the ReLu nonlinearity. This highlights how biologically inspired components can contribute to computational plausibility and to improve performance in practical visual tasks, a dual benefit that this thesis will investigate.

In more recent years, the dominant architecture in computer vision has shifted from convolutional models to those based on self-attention mechanisms, the Transformers. Originally designed for natural language processing [Vaswani et al., 2017], Transformers rely on a different principle: instead of using localized receptive fields as in CNNs, each input element can attend to all others through an attention mechanism that dynamically weights their relevance. This idea, when applied to vision, led to the development of the

Vision Transformer (ViT) [Dosovitskiy et al., 2020], which splits images into patches and processes them using blocks of self-attention. While ViTs depart from the convolutional principle of localized receptive fields, they build on the biologically inspired idea of attention, allowing the selective integration of information across space. This introduces a different but still biologically plausible mechanism, complementing the localized processing of CNNs. Unlike the classic biologically inspired design of CNNs, ViTs do not impose the same kind of progressive spatial compression and explicit hierarchical structure characteristic of both CNNs and biological vision. Instead, they offer new architectures that surpass CNNs in computer vision benchmarks, with internal attention mechanisms that can also be examined for their biological plausibility.

Thanks to the maturity of these architectures, from CNNs to ViTs and hybrid variants, deep learning has reached outstanding levels of performance across a wide range of tasks in vision and many other fields. Supervised models trained on massive labelled datasets such as ImageNet [Deng et al., 2009] or Microsoft Common Objects in Context, known as COCO [Lin et al., 2014], currently lead benchmarks in object classification, semantic segmentation, detection, and instance recognition, in some cases even surpassing human performance. These models have demonstrated impressive ability to learn useful representations, scale with model size, and generalize to new tasks via transfer learning or fine-tuning techniques [Zhai et al., 2022]. Simultaneously, self-supervised approaches, such as Self-Distillation with No-labels (known as DINO) [Caron et al., 2021], and multimodal approaches that integrate language, such as Contrasting Language-Image Pre-training (CLIP) [Radford et al., 2021a], have expanded the frontiers of computer vision, showing that it is possible to learn high-quality visual representations without explicit labels or using textual supervision. As a result, deep learning has become the *de facto* standard in artificial vision, displacing traditional methods and powering applications as diverse as autonomous driving, computational medicine, or robotics.

However, reaching these levels of performance has come at a conceptual cost. In many cases, the biologically inspired principles that originally guided the design of these architectures have been progressively replaced by empirical strategies and designs focused on optimizing performance on specific data. A clear example is what happened with AlexNet and the use of the Local Response Normalization. While AlexNet authors highlighted this compo-

nent as crucial, the next most famous model, the Visual Geometry Group (VGG) [Simonyan and Zisserman, 2014], found that it was not worth it for their specific architecture and data. Since then, Local Response Normalization and similar mechanisms have largely disappeared from standard deep learning architectures. This illustrates a broader trend: while certain biologically inspired ideas remain in modern models, such as attention, which draw on concepts from visual attention, many recent architectural choices and training strategies are guided mainly by empirical performance rather than neuroscientific plausibility. Importantly, excelling at a task does not necessarily mean that these models perceive the world in ways comparable to humans. Perceptual alignment is a separate dimension, which does not require strict biological realism but speaks to whether models make decisions in ways that resemble human perception. Understanding this distinction, and its consequences, motivates a deeper analysis of both architectural choices and perceptual outcomes. Indeed, numerous studies have shown a growing gap between human perceptual behavior and the internal strategies employed by these models to solve visual tasks.

Several warning signs illustrate this misalignment. First, it has been observed that deep learning models tend to exploit statistical shortcuts in the data, basing their decisions on spurious correlations or trivial features, rather than attending to more shape-biased, abstract or invariant properties as the human brain does. A few examples shown in figure 1.1: Deep learning models are used to classify images of cows perfectly, but only when they are in the typical green grass landscape. If the cow is located in a "strange" background, as a beach, models tend to fail much more in their classification, meaning that they use the background as a trivial feature for the classification [Beery et al., 2018, Geirhos et al., 2020]. Moreover, deep learning models have been shown to classify images based on the object textures, while humans mainly classify objects based on shape [Landau et al., 1988]. This difference is extremely visible in the shape-texture cue conflict images, i.e. modified images with shape and textures from different objects, such as an image with the shape of a cat but with the texture of an elephant skin. While humans use to classify this type of images according to the class of the shape objects, a cat, deep learning models classify them according to the texture objects, an elephant [Geirhos et al., 2018a]. While these strategies allow deep learning models to maximize their accuracy in controlled settings, they severely limit their ability to generalize out of distribution, to new unseen

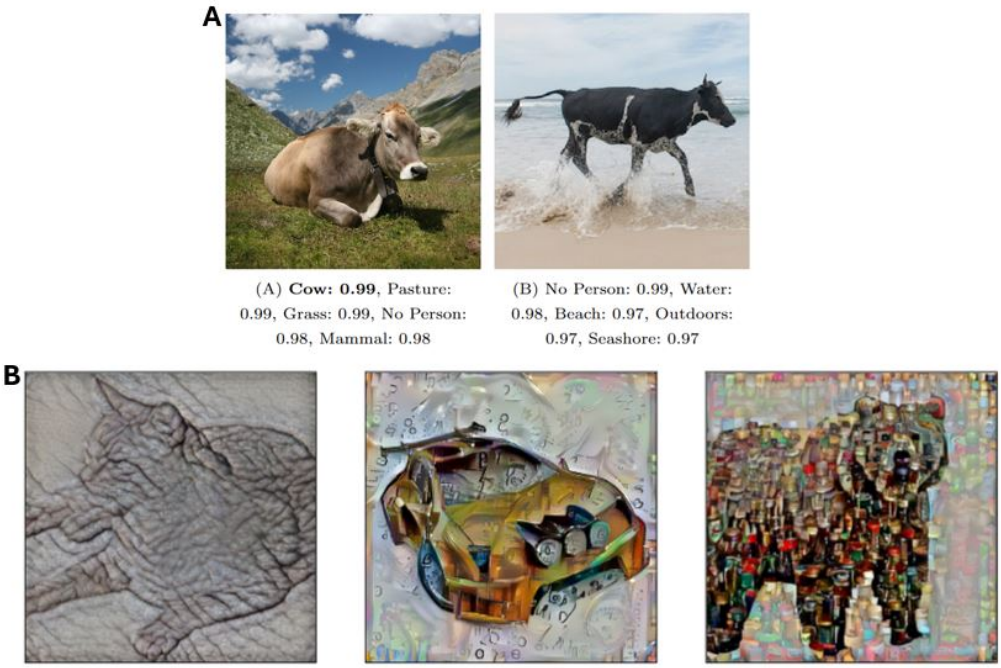


Figure 1.1: **Deep learning models classification short-cuts:** Different works have shown that deep learning models classify images using short-cuts or following non-human feature decisions. Particularly, panel A shows how a model fails to detect the cow present in the image when it is located in a "strange" or non-usual background for a cow, such as the beach [Beery et al., 2018]. Pabel B shows some of the texture-shape cue conflict images that, in general, humans classify according to their shape (cat, car and bear from left to right), while deep learning models tend to classify according to their texture (elephant, clock and bottle) [Geirhos et al., 2018a].Figures from [Beery et al., 2018, Geirhos et al., 2018a].

data [Geirhos et al., 2018b].

Second, deep learning models are notably sensitive to minor image perturbations, such as noise, blur, compression, or lighting changes that barely affect human judgments, the so-called adversarial attacks [Szegedy et al., 2013, Kurakin et al., 2018, Wichmann and Geirhos, 2023]. In this case, as shown in figure 1.2, a non-visible noise for humans can completely change the prediction classification of a deep learning model. This lack of robustness increases the doubt on their suitability for real-world deployment and suggests that they fail to capture the perceptual constancies that charac-

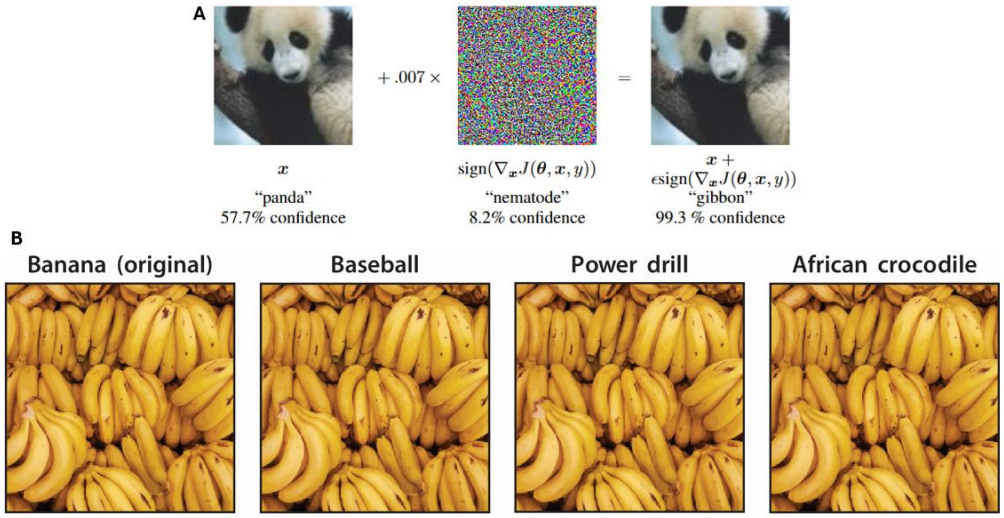


Figure 1.2: **Deep learning models affected by non-visible noise for humans:** Deep learning models completely change their image classification prediction when noise, non-visible for humans, is added to the images. Particularly, panel A shows how an example from the original adversarial attack work, where an original image is classified as "panda", but after adding a small quantity of white noise, which leaves the image unchanged from a human point of view, the model prediction completely changes to "gibbon" [Szegedy et al., 2013]. Panel B shows another example of adversarial attacks where visually similar images to humans are completely misclassified by models when unnoticeable noise is added, changing the prediction from the correct "banana" to "baseball", "power drill" or "African crocodile" [Wichmann and Geirhos, 2023]. Figures from [Szegedy et al., 2013, Wichmann and Geirhos, 2023].

terize biological vision. Finally, various studies have shown that the internal representations constructed by these models do not always reflect the functional organization of the human visual system. In particular, they show a lack of global semantic correspondence [Peterson et al., 2018, Roads and Love, 2021] as well as some hierarchical missalignment with human perception [Xu and Vaziri-Pashkam, 2021, Muttenthaler et al., 2024]. These limit the deep learning model's ability to perform clustering and different tasks in the same way humans do.

Taken together, these observations point to the central problem motivating this thesis: success in quantitative benchmarks does not guarantee per-

ceptual alignment with humans. Bridging this gap requires designing models that take inspiration from the visual system and developing systematic methods to evaluate how closely their internal representations and behavioral outputs align with human perception.

1.2 Motivation

Understanding and replicating human visual perception is a long-standing objective in neuroscience and an increasingly relevant topic within artificial intelligence, particularly in efforts to build more robust and interpretable vision systems [Kriegeskorte, 2015, Dapello et al., 2022, Liu et al., 2023, Muttenthaler et al., 2022, Sucholutsky et al., 2023]. While many artificial intelligence models prioritize task performance, investigating how and when they align with human perception can offer valuable insights for both improving these systems and improve our understanding of the human brain. Despite the remarkable success of deep neural networks in vision-related tasks and many other areas, fundamental questions remain about whether these systems perceive the world in ways that are even remotely comparable to human perception. In this context, the concept of human alignment, i.e. the degree to which artificial systems process information in functionally and representationally human-like ways, has gained growing attention [Geirhos et al., 2018a, Peterson et al., 2018, Dapello et al., 2022, Muttenthaler et al., 2022, Liu et al., 2023, Sucholutsky et al., 2023, Malo et al., 2022, Hernandez-Camara et al., 2022, Hernández-Cámara et al., 2024b, Hernandez-Camara et al., 2025b, Hernandez-Camara et al., 2025a], particularly in domains like vision where perceptual correspondence is critical.

However, why study the human alignment of deep learning models? From a scientific perspective, human alignment provides a valuable framework for probing the principles underlying human perception itself. Artificial models offer a controlled environment to test hypotheses about visual coding, representation, and generalization. If a network trained on natural images spontaneously reproduces known human perceptual behaviors this suggests that such behaviors may emerge naturally from the statistics of the environment and the constraints of the task, without the need for hardcoded biological mechanisms [Barlow et al., 1961, Yamins and DiCarlo, 2016, Richards et al., 2019]. Inversely, discrepancies

between model behavior and human perception can shed light on what makes biological vision unique, offering insights into the computational goals, architectural constraints, and inductive biases that shaped the human visual brain. More broadly, this connects to Marr's classic framework of levels of analysis [Marr and Poggio, 1976, Marr, 2010]. Deep learning models blur the distinction between computational principles, algorithmic strategies, and implementation details. For example, models trained under similar computational objectives can nevertheless diverge from humans depending on architectural design or training statistics, raising the question of whether computational goals can ever be studied independently from architectural constraints. Recent work emphasized that alignment depends on the interplay between these levels, with biological plausibility and task performance often pulling in different directions [Geirhos et al., 2020, Muttenthaler et al., 2022, Sucholutsky et al., 2023]. In this context, analyzing which biological and engineering elements contribute to emergent behaviors can clarify how design choices, spanning architecture, optimization, and data, shape the emergence of human-like perceptual properties in artificial systems.

From an engineering point of view, human alignment is not only desirable, but it may be essential. The human brain is probably the most robust and general-purpose visual system we know. It excels in low-contrast environments, easily resists perturbations or noise, and adapts almost effortlessly to changes in context or illumination [Gibson, 2014, Geirhos et al., 2018b, Dodge and Karam, 2017]. Getting these properties into artificial systems could lead to models that are significantly more robust, better able to generalize beyond their training distribution, and ultimately more interpretable for human users [Sucholutsky and Griffiths, 2023]. In applications where safety, trust, and human interaction are needed, such as autonomous driving, medical diagnostics, or assistive technology, models that make decisions "in ways humans can understand" are far more valuable than black-box systems that are merely optimized for raw accuracy. Therefore, improving the alignment with human perceptual principles is not only a scientific challenge, but a practical necessity.

Despite its importance, visual human alignment is rarely evaluated systematically in current deep learning practice. Standard benchmarks in computer vision typically focus on task performance, such as classification accuracy, but provide little insight into either the internal representations learned by the model or whether the strategies it uses to arrive at its predictions are

comparable to those employed by human observers [Geirhos et al., 2018a, Geirhos et al., 2020, Hepburn et al., 2022, Hernández-Cámara et al., 2024b, Hernandez-Camara et al., 2025c]. In this thesis, we explore both levels of human alignment: behavioral alignment, assessed through the similarity between model and human responses on perceptual tasks; and representational alignment, examined via layer-wise analyses of internal embeddings and their correspondence with perceptual similarity. A model may achieve high scores by exploiting non-robust features, spurious correlations, or data-specific biases, while entirely missing the core invariances and perceptual structures that guide human vision. In short, performance is not the same as perceptual similarity. This gives rise to the central questions that motivate the present thesis:

How can we build models that not only perform well, but also see like humans? Can we measure the human alignment of deep learning models? If so, how do different deep learning design choices affect performance and human alignment?

Answering these questions requires a dual effort: improving model architectures through biologically inspired mechanisms, and developing methods to evaluate and understand their perceptual alignment. More broadly, it requires an interdisciplinary perspective, bridging computational neuroscience, vision science, and artificial intelligence, to ensure that the models we design are not only effective but also human-aligned. These considerations set the stage for the specific objectives of this thesis, which are presented in the next chapter, together with the overall structure of the manuscript.

2

Objectives and Thesis Structure

2.1 Objectives

The main goal of this thesis is to bridge the gap between artificial and biological vision by combining bio-inspired model design with systematic evaluation of human alignment. This entails three specific objectives:

1. **Integrating biologically inspired computations into deep learning architectures:**

This objective focuses on Divisive Normalization, a canonical computation of the early visual cortex, reinterpreted as a functional non-linearity in modern networks. By embedding DN modules into deep segmentation models, the thesis explores whether this mechanism can improve robustness to distortions, domain shifts, and other challenges where human vision excels. This objective reflects the broader hypothesis that biological computations, when transplanted into artificial systems, can enhance both their robustness and plausibility.

2. **Developing and applying methodologies for measuring human**

alignment.

A second objective is to design systematic ways of evaluating how closely artificial models resemble human perceptual behavior. This includes the development of psychophysics-inspired tasks, layer-wise analyses, and evaluations across different levels of abstraction, from low-level distortions to high-level semantic categorization. By combining behavioral comparisons with layer-wise representational analyses, these methodologies aim to provide a more comprehensive assessment of human alignment than standard benchmarks allow.

3. **Analyzing the factors that shape alignment in deep models.**

The third objective is to disentangle the contributions of architectural design choices, training data, and optimization strategies to the emergence of human-like behavior in deep models. By comparing models across different configurations and systematically varying these factors, the thesis seeks to identify which aspects most strongly determine perceptual alignment. This allows us to move beyond case-by-case observations toward a clearer understanding of the principles that drive human-like properties in artificial networks.

By addressing these three objectives, the thesis aims to make both practical and conceptual contributions. On the practical side, it introduces improved architectures and novel methodologies for evaluating perceptual alignment. On the conceptual side, it offers insights into when and why deep networks resemble the human visual system, helping advance the broader dialogue between neuroscience and artificial intelligence.

2.2 Thesis Structure

To address the objectives outlined above, this thesis follows a dual research strategy aimed to bridging the gap between artificial and biological vision systems. On one hand, it explores how to construct bio-inspired neural architectures by incorporating mechanisms observed in the human visual system into modern deep learning models. On the other hand, it seeks not only to evaluate how aligned these models are with human perception, but also to understand which factors from deep learning models most strongly shape this alignment. By combining model design with systematic evalua-

tion and analysis, the thesis aims to contribute both practical advances and conceptual insights into how and why neural networks sometimes resemble (or fail to resemble) the human visual system.

The first pillar of this approach focuses on bio-inspired model design. In particular, the thesis introduces the Divisive Normalization (DN) [Carandini et al., 2005, Carandini and Heeger, 2012], a canonical computation in the early visual cortex, as a key architectural component in deep segmentation networks. By including DN into deep learning architectures, the goal is to enhance the model's robustness to naturalistic distortions, which human vision handles easily.

The second pillar focuses on the perceptual evaluation and analysis. Beyond conventional metrics like accuracy, this thesis develops and applies novel methodologies inspired by vision science to assess the model's alignment with human perception. These include a comprehensive layer-by-layer analysis of internal representations, enabling fine-grained comparison across models and abstraction levels, as well as a psychophysics-inspired method for specific perceptual domains. For example, the thesis examines low-level properties such as color discrimination (e.g., MacAdam ellipses), contrast sensitivity in multimodal models, robustness to noise perturbations, and higher-level semantic similarity judgments. The models are evaluated in terms of output similarity and representational alignment, with behavioral data across multiple perceptual tasks and complexity levels, from low-level noise perturbations to high-level semantic categorization, providing a multi-layered understanding of how human-like each model truly is. In addition to evaluating alignment, this thesis also investigates which components of deep learning models most significantly influence it, helping to disentangle whether perceptual similarity arises from biological mechanisms, statistical regularities in the data, or task-induced inductive biases.

These two pillars define the structure of the manuscript, which is divided into four main parts:

- Part I: Context, motivation and thesis structure.
- Part II: Construction of bio-inspired models for robust visual processing.
- Part III: Evaluation of perceptual alignment across tasks, representations, and architectures.

- Part IV: Discussion of the implications on the relation between deep models and human perception.

Given the publication of multiple papers studies during the PhD period, this thesis adopts the format of a compendium. It includes four journal articles that constitute the core scientific contributions, six conference publications that provide complementary analyses, and one commentary reflecting on computational levels of analysis in relation to biological plausibility. Together, these contributions provide a dual perspective: on one hand, how biologically inspired computations such as Divisive Normalization can improve robustness in deep models; and on the other, how we can measure and understand their degree of alignment with human perception.

Part II

Building Bio-Inspired Deep Learning Models

3

Deep Learning Models with Divisive Normalization

3.1 Introduction

One of the key objectives of this thesis is to design deep learning models that not only perform well but also process visual information in ways that are closer to how the human visual system works. As discussed in the first part of this thesis, current deep learning models often rely on shortcuts, such as textures or undesired correlations, that lead to limited generalization and weak robustness. This contrasts with human perception, which is remarkably stable and invariant under similar variations.

One reason for this discrepancy could be the absence of biologically inspired mechanisms in modern deep learning architectures. At the same time, it is also a possible solution to address this gap. While early neural networks were more inspired by neuroscience, many of the current models shifted away from those foundations in pursuit of pure task performance. One way to reconnect with biological vision is to reintroduce computations that play a fundamental role in how the brain processes visual information.

This chapter focuses on one such computation: *Divisive Normalization*

(DN). Divisive Normalization is a canonical operation observed in the early stages of the visual cortex [Heeger, 1992, Carandini and Heeger, 1994]. Here, we investigate how DN can be integrated into deep neural networks for image segmentation, and whether doing so improves their robustness to naturalistic image degradations like blur, fog, or reduced contrast. We begin by introducing DN from a biological perspective and then explain how we incorporated this mechanism into deep learning architectures. Finally, we present a set of experiments, originally published in two of the four journal papers of this thesis [Hernández-Cámara et al., 2023, Hernández-Cámara et al., 2025], that evaluate how DN affects performance under various distortions.

3.2 Divisive Normalization in Human Vision

Divisive Normalization is one of the most widely observed computations in the early stages of the visual cortex [Heeger, 1992, Carandini and Heeger, 1994], particularly in areas such as V1 and the Lateral Geniculate Nucleus. However, it is also present in other brain regions such as the auditory channel [Rabinowitz et al., 2011]. The DN is a non-linearity that describes how the response of a neuron is not just determined by the input it receives, but is also modulated or scaled by the activity of surrounding neurons. This introduces a local gain control mechanism in which strong responses are attenuated if nearby activations are also strong, and, at the same time, weaker signals may be enhanced in low-activity regions.

Mathematically, DN was first formalized in the early 1990s as a non-linear transformation applied to the output of a linear filter bank [Heeger, 1992]. For a neuron tuned to the $i-th$ feature and $p-th$ spatial location, the DN y_{ip} response is defined as:

$$y_{i,p} = \frac{z_{i,p}}{(\beta_i + \sum_{j,p'} \gamma_{ijpp'} |z_{j,p'}|^{\alpha_{ij}})^{\epsilon_{ij}}} \tag{3.1}$$

Here, $z_{i,p}$ is the linear response of the $i-th$ feature and $p-th$ spatial location neuron, and the parameters α, ϵ and β control the non-linearity. In particular, the relation between β and $\gamma_{ijpp'}$ determines the inhibition strength. When $\beta_i/\gamma_{ijpp'} \gg 1$ the transformation is almost linear, whereas for $\beta_i/\gamma_{ijpp'} \ll 1$ its effect is highly non-linear. The exponents also control the norm used in the normalization pool.

The interaction kernel, γ, determines which surrounding neurons and how much they contribute to the normalization. In its most general form, the kernel can have whatever arbitrary structure, including neurons across space and feature channels [Carandini and Heeger, 1994]. For example, if the weights are uniform, the DN resembles a mean-normalization; if they are sparse or biased, it can selective suppress certain features or locations. Different kernel configurations have been proposed: some ignore spatial interactions and focus only in channel-wise interactions, either in a dense [Ballé et al., 2016, Hepburn et al., 2020, Ballé et al., 2015] or convolutional [Miller et al., 2021] combination of features; while others take into account spatial interactions under specific constrains such as uniform weights [Ren et al., 2016], circular (ring) neighborhoods [Giraldo and Schwartz, 2019, Pan et al., 2021] or symmetries interaction patterns in space [Burg et al., 2021]). In neuroscience, spatial interactions are often modelled as convolutional patterns, i.e. Gaussian weights over a local neighbourhood, although more complex or learned kernels can be used in computational models [Watson and Solomon, 1997, Martinez-Garcia et al., 2018, Martinez et al., 2019].

Over time, Divisive Normalization has evolved from a purely physiological model to an applicable computational tool across domains. Although it was initially developed to describe neural responses in the early visual system [Heeger, 1992], DN was later applied to image processing applications. For example, it contributed to improving compression standards like JPEG and MPEG [Epifanio et al., 2003, Malo et al., 2005, Ballé et al., 2016, Islam et al., 2021]. It has also been used in tasks such as image enhancement [Gutierrez et al., 2005] or subjective image quality assessment [Laparra et al., 2010, Hepburn et al., 2020, Ma et al., 2017, Bowen et al., 2022, Vila-Tomás et al., 2024], where its ability to normalize perceptual variations was beneficial. From a statistical perspective, DN has been shown to increase independence between responses, supporting more efficient representations [Schwartz and Simoncelli, 2001, Cekic et al., 2022, Malo and Gutiérrez, 2006, Malo and Laparra, 2010, Ballé et al., 2015].

The integration of DN into modern deep learning was possible due to the rise of automatic differentiation. This enabled the implementation of the DN as a differentiable module and optimize it jointly with the rest of the network, the so-called *Generalized Divisive Normalization* [Ballé et al., 2016]. Early models with this *Generalized Divisive Normalization* simplified the spatial interactions to focus only on channel-wise ker-

nels, but more recent approaches have recovered the full spatial and feature interactions, bringing the model closer to its biological origins. These advancements have enabled the use of the DN in many different machine learning tasks, such as compression algorithms [Ballé et al., 2016], perceptual distance metrics [Hepburn et al., 2020, Vila-Tomás et al., 2024], and classification models, where it has been shown to improve robustness to noise, adversarial perturbations, and domain shifts [Coen-Cagli and Schwartz, 2013, Giraldo and Schwartz, 2019, Miller et al., 2021].

3.3 Implementing DN in Deep Learning Architectures

While most prior work has focused on classification, here we extend the use of DN to a different and highly structured task: semantic image segmentation. The goal in this task is to assign a meaningful class label to each pixel in the image. This involves both fine-grained local feature extraction and global contextual understanding, two processes where biologically inspired normalization could provide benefits.

To evaluate whether the use of the DN is also beneficial in image segmentation tasks, we implement segmentation models with and without Divisive Normalization. By building it as a functional layer in an automatic differentiation framework, we can use it as a simple normalization layer in deep learning architectures. In particular, we focus on the U-Net architecture [Ronneberger et al., 2015], which has become a standard backbone in image segmentation tasks [Azad et al., 2024] due to its simplicity, effectiveness and widespread use. As illustrated in figure 3.1, we implement the DN at the beginning of each of the encoder blocks, where the network extracts and processes the visual information. This allows the DN to modulate local features from the early layers. The addition of DN layers introduces a minimal increase in model parameters, just a 1.8% increase from 2.749.902 to 2.798.482 trainable parameters for the standard and DN-improved U-Nets [Hernández-Cámara et al., 2023].

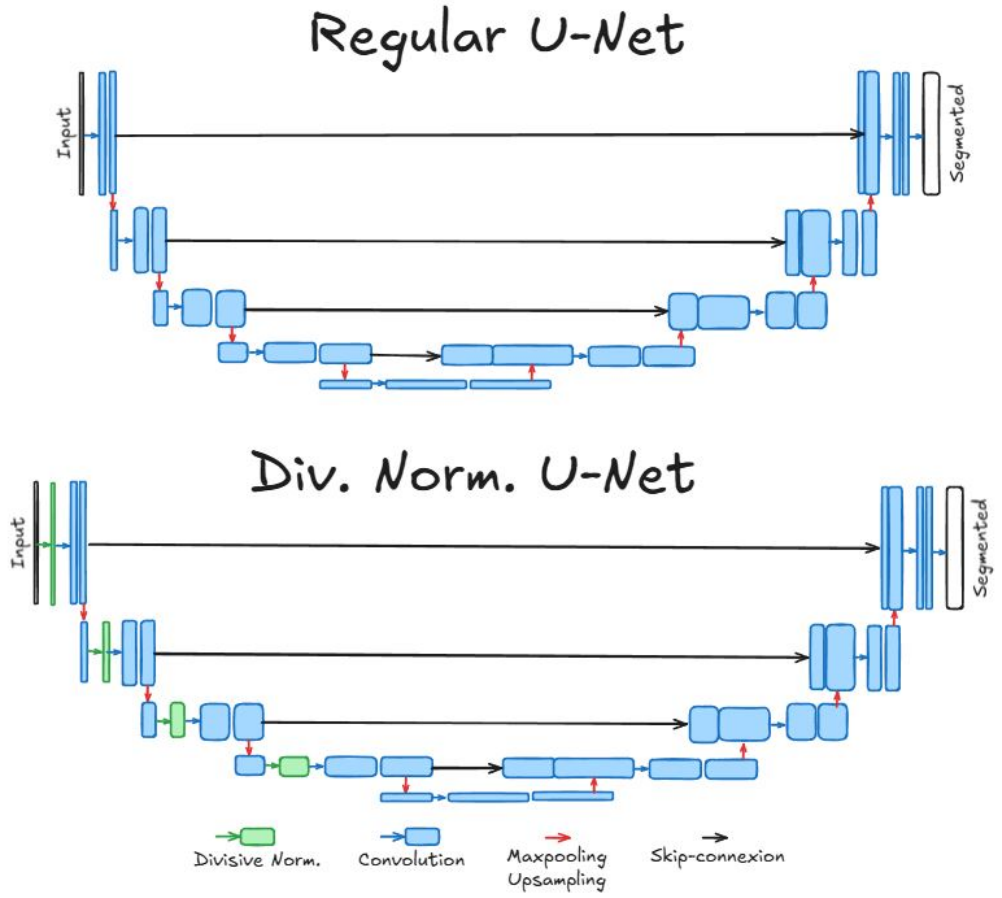


Figure 3.1: **U-Net for segmentation with and without DN layers:** Comparison between the baseline standard U-Net architecture (top) and the modified U-Net with four Divisive Normalization (DN) layers (bottom). DN layers are shown in green and are added at the start of each encoder block. Both models share the same structure regarding convolutional layers, skip connections (black arrows), and pooling/upsampling operations (red arrows). The legend indicates the layer types used in the diagram. Figure from [Hernández-Cámara et al., 2025].

3.4 Experimental Setup

To evaluate the impact of Divisive Normalization on segmentation performance and robustness, we consider an autonomous driving scenario. This domain offers a particularly challenging setting, as it involves scenes captured

under highly variable conditions: from bright daylight to nighttime, across diverse weather phenomena such as fog, rain, or snow, and with different lighting and spectral light conditions. These factors lead to images with high dynamic ranges that make segmentation extremely sensitive to changes that are often imperceptible to humans.

To rigorously assess the effect of DN, we train and evaluate segmentation models with and without Divisive Normalization under identical training conditions. All models are trained until convergence using different real-world and simulated datasets, and their performance is evaluated under multiple types of data degradations, including fog and luminance, contrast, and illumination-based distortions. For each dataset and model configuration, we perform 10 independent training runs with different seeds. We measure the model's performance using the standard Intersection-over-Union (IoU) metric (also known as Jaccard's index) across the different test sets, which is widely adopted in semantic segmentation benchmarks [Real and Vargas, 1996, Cordts et al., 2016, Lin et al., 2014]. It varies between 0 for a totally wrong prediction and 1 for a perfect prediction, and it measures how much the classes of the model prediction overlap with the real classes. This setup allows us a controlled comparison of robustness and to isolate the contribution of DN to performance under realistic variable visual conditions.

In particular, we use two real-world datasets: Cityscapes [Cordts et al., 2016] and Nighttime Driving-test datasets [Dai and Van Gool, 2018], which include images in day and night-time, respectively, and always in good weather conditions. For the simulated data, we used: FoggyCityscapes [Sakaridis et al., 2018], GTA-V [Richter et al., 2016], and a custom dataset generated with the CARLA simulator [Dosovitskiy et al., 2017, IPL,]. Foggy Cityscapes is a version of Cityscapes with three severity levels of synthetic fog, while the GTA-V and CARLA datasets contain images from a videogame and a city-driving simulator, respectively, in many different weather and daytime conditions. Figure 3.2 shows a representative image of each dataset.

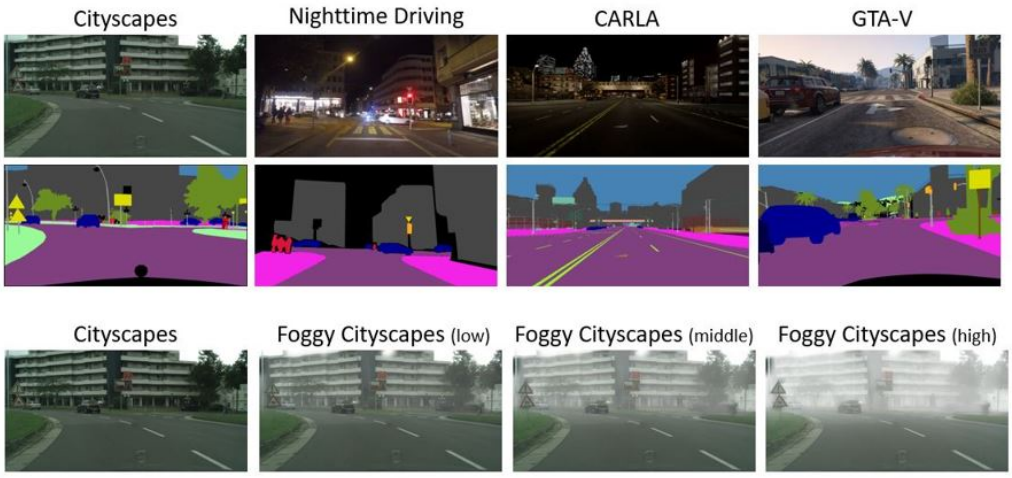


Figure 3.2: **Representative images of the different datasets used to train and evaluate the segmentation models:** All datasets correspond to autonomous driving scenarios and share common semantic segmentation labels. Cityscapes and Nighttime Driving are based on real images, while CARLA and GTA-V are synthetic. Foggy Cityscapes includes synthetic fog onto real images from the Cityscapes dataset. Figure adapted from [Hernández-Cámara et al., 2025].

3.5 Segmentation and Robustness Results

Using the datasets described above, we trained two types of segmentation models: a baseline U-Net without Divisive Normalization (DN), and a modified U-Net including four DN layers, as shown in Figure 3.1. We called them no-DN and 4-DN models, respectively. We train both models on the training sets of the Cityscapes and CARLA datasets, and evaluate them in their respective test sets, as well as on all the other additional test datasets: Foggy Cityscapes (3 fog levels), Nighttime Driving, GTA-V. Table 3.1 summarizes the mean Intersection-over-Union (IoU) for the different training and test configurations.

There are many important results to notice from this table: first and foremost, across all the training and test combinations, the models with Divisive Normalization always improve the baseline that do not implement it. Second, as expected, models perform best on the dataset they were trained on. This effect is even more extreme for the models trained in CARLA, highlighting that training solely on synthetic data can be problematic when transferring

Table 3.1: Mean IoU over ten runs for the models trained on Cityscapes and CARLA training sets and evaluated in the test sets of: Cityscapes, Foggy Cityscapes (3 fog levels), Nighttime Driving, CARLA and GTA-V. Improvements due to the inclusion of DN layers relative to the no-DN baseline are indicated in parentheses. Table from [Hernández-Cámara et al., 2025].

Test dataset	Cityscapes train		CARLA train	
	no-DN	4-DN	no-DN	4-DN
Cityscapes	0.75	0.77 (2.7%)	0.50	0.54 (8.0%)
Foggy (low)	0.65	0.70 (7.7%)	0.46	0.52 (13.0%)
Foggy (middle)	0.54	0.62 (14.8%)	0.44	0.50 (13.6%)
Foggy (high)	0.40	0.48 (22.5%)	0.40	0.46 (15.0%)
Nighttime	0.24	0.29 (20.8%)	0.31	0.33 (6.5%)
CARLA	0.51	0.54 (5.9%)	0.90	0.91 (1.1%)
GTA-V	0.55	0.61 (10.9%)	0.62	0.65 (4.8%)

it to real-world applications. In these in-domain scenarios, the benefit off the DN is smaller, but still positive. However, the effect of the Divisive Normalization becomes more important when the test data distribution differs from the training one, i.e. train in real and test on synthetic or *vice versa*. Third, all models get their lowest result in the Nighttime Driving dataset. This is expected due to the extremely low luminance of the images and the daytime of the real training data. Interestingly, in this most challenging scenario, is where the models trained with real images really benefit from the DN, with improvements exceeding 20% IoU increase. This shows the importance of including DN layers in extreme low-light conditions. Fourth, again as expected, as fog severity increases in the Foggy Cityscapes test sets, segmentation performance gets reduced. However, the models without DN are more sensitive and show larger IoU drops. At the same time, the improvements due to the DN grow as visual condition degrades, i.e. as fog increases.

Overall, these results show that while Divisive Normalization always improves models' results, its advantages become even more important under challenging or out-of-distribution test conditions, such as low visibility, strong illumination changes, or domain shifts between synthetic and real data.

To further explore the benefits of Divisive Normalization in segmentation models, we need to extend the analysis beyond fixed test datasets. A

key limitation of existing datasets is their lack of controlled and continuous variations in visual features, such as luminance, contrast and hue and saturation illumination. To overcome this limitation and extend the range of our experiments, we selected 100 images from Cityscapes and CARLA test sets and, systematically, modified them along five relevant visual dimensions: (1) mean luminance, (2) achromatic contrast, (3) chromatic contrast, and two aspects of spectral illumination, (4) hue angle, and (5) saturation. Figure 3.3 shows a representative image of the applied modifications.

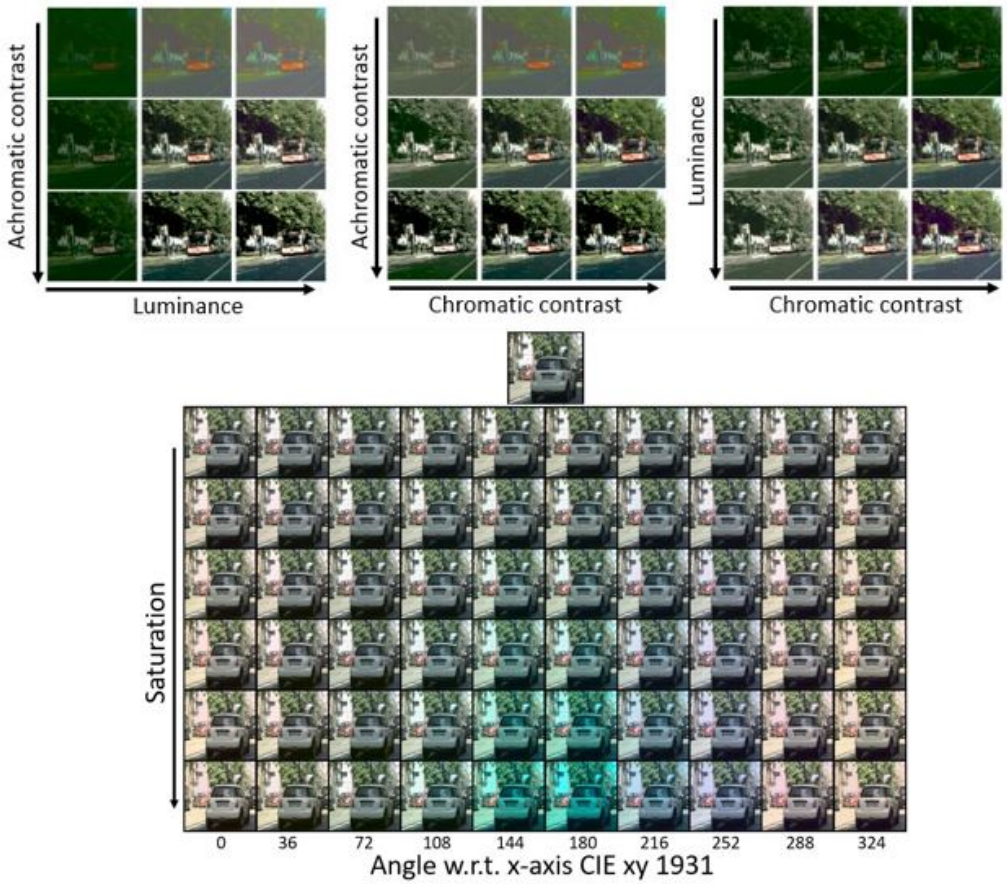


Figure 3.3: **Controlled modification of luminance, contrasts and spectral illumination.** Top: Examples of changes to mean luminance, achromatic contrast and chromatic contrast, modified in pairs while the other dimension remains fixed to its original value. Bottom: Variations in spectral illumination through hue rotation angle and saturation. Figure adapted from [Hernández-Cámara et al., 2025].

Figure 3.4 shows the relative IoU improvements (in percentage) due to the DN across the different combinations of luminance, achromatic contrast, and chromatic contrast. Each matrix corresponds to a specific chromatic contrast level, with luminance and achromatic contrast varying along the two axes. The top row shows the gains for models trained on the Cityscapes real images, and the bottom row corresponds to models trained on the CARLA synthetic images. Note that, as before, the use of DN improves segmentation accuracy. For the models trained on real images, the largest benefits happen under low luminance and low contrast scenarios, reinforcing the idea that the DN is especially beneficial in night-like or foggy images. In contrast, models trained on CARLA synthetic images obtain their peak gains in a different region of the parameters space, specifically at higher luminances and achromatic contrasts. This is due to the statistical differences in luminance and contrasts between the CARLA and real images. This, again, highlights the risk of using exclusively synthetic data that may not follow the distribution of real images.

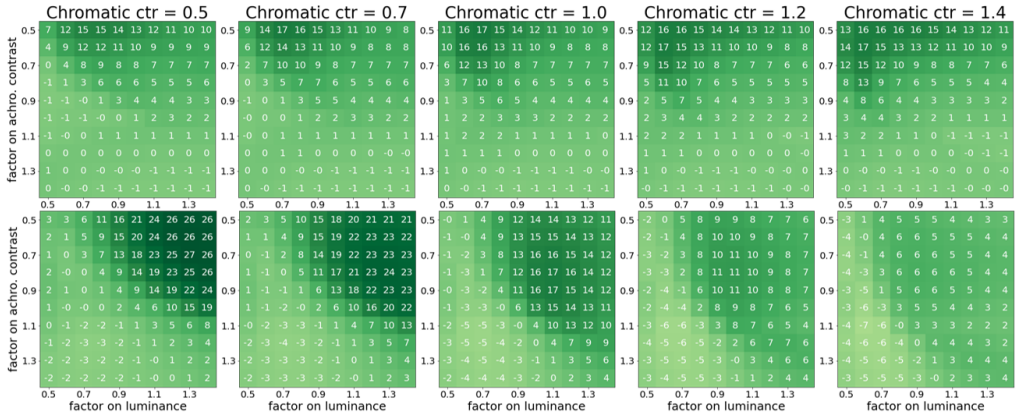


Figure 3.4: **Relative IoU gains (in percentage) of DN-augmented models over no-DN standard U-Net across luminance and contrast variations:** Each matrix corresponds to a specific chromatic contrast level, showing gains over combinations of achromatic contrast and mean luminance. Top: models trained on Cityscapes (real images). Bottom: models trained on CARLA (synthetic images). Figure from [Hernández-Cámara et al., 2025].

Similarly, figure 3.5 shows the DN benefits (in percentage) as a function of the changes in the image illuminants, i.e. the hue angle and the saturation. Models trained on real images (top row) consistently benefit from DN across

most illuminant conditions, i.e. obtain positive increments, except for a small region of highly saturated blue illuminants, which are rarely encountered in the real world. As before, models trained on CARLA (bottom row) show a different gain distribution. Nevertheless, DN still provide an important improvement, particularly under low-saturation illumination.

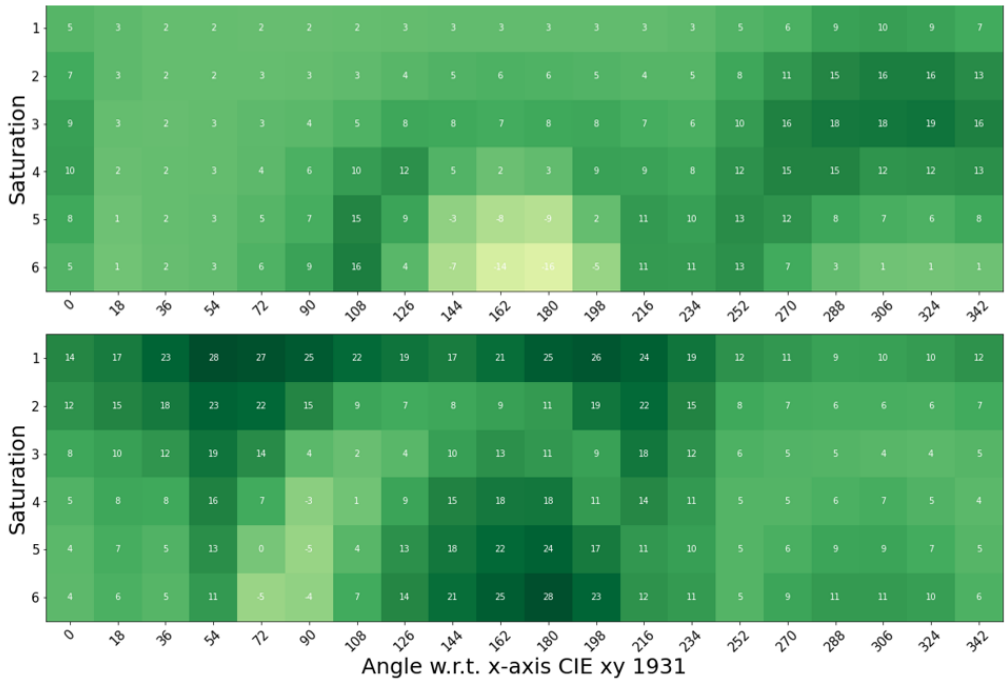
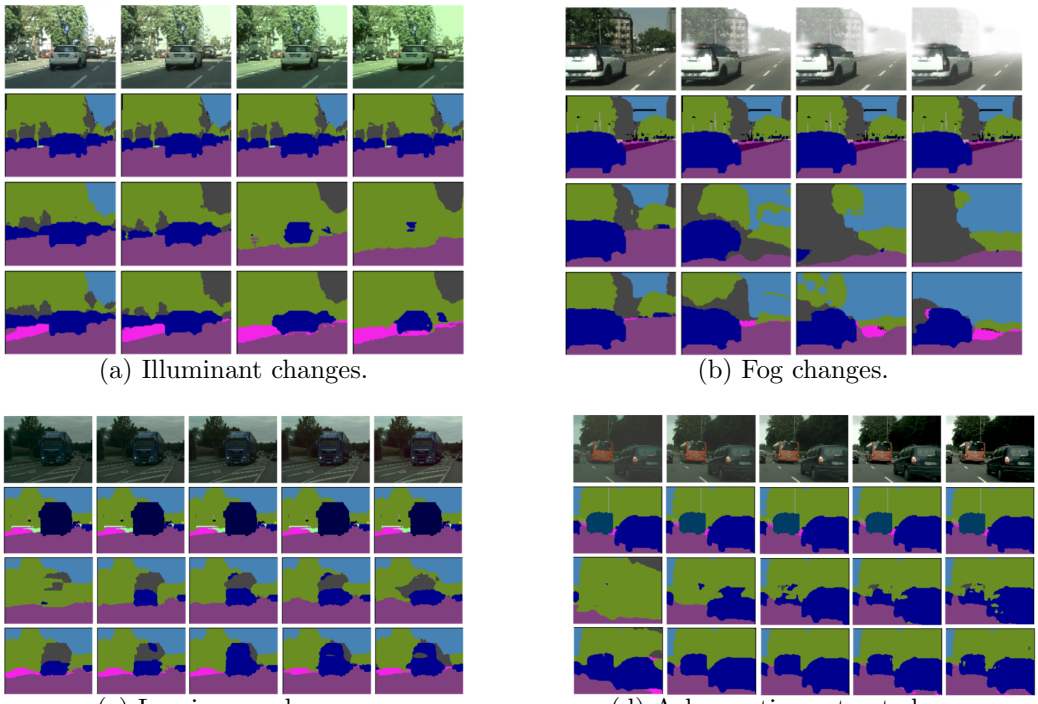


Figure 3.5: **Relative IoU gains (in percentage) of DN-augmented models over no-DN standard U-Net across illuminant variations:** Each matrix shows how segmentation performance improves depending on the hue angle and saturation of the input image's illuminant. Top: models trained on Cityscapes. Bottom: models trained on CARLA. Figure from [Hernández-Cámara et al., 2025].

Together, these results demonstrate that introducing Divisive Normalization into segmentation deep neural networks improves their performance under a variety of environmental conditions. Its results are better not only across diverse natural and synthetic datasets, but also under systematically controlled variations in five key visual dimensions: mean luminance, achromatic contrast, chromatic contrast, hue, and saturation. Our results also



(a) Illuminant changes.

(b) Fog changes.

(c) Luminance changes.

(d) Achromatic contrast changes.

Figure 3.6: **Predictions of standard and DN-augmented U-Net models under controlled visual variations:** Each panel compares predictions from the standard U-Net (no DN) and the DN-enhanced model. In each panel, the first row shows the input images; the second shows the segmentation ground truth; the third and fourth rows show the predictions from the standard and DN-enhanced models, respectively. Panel 3.6a and 3.6b: increasing saturation and fog intensity (left to right). Panel 3.6c and 3.6d: varying luminance and contrast from the original (center), increasing (right) or reducing (left). Figure from [Hernández-Cámara et al., 2025].

show that the benefit of Divisive Normalization is even more crucial in extreme conditions, such as low luminance (night images) or low contrast (heavy fog). In these scenarios, DN significantly improves the model robustness, reducing its sensitivity to this environmental variability. These findings are illustrated in figure 3.6, which compares the predictions from models with and without DN under some of the considered visual distortions. While standard U-Net model predictions get worse quickly with the different distortions, the U-Net model with DN gets more consistent predictions, being more robust to these changes.

4

Understanding the Robustness of Bio-Inspired Models

In the previous chapter, we empirically demonstrated that including Divisive Normalization into deep segmentation neural networks consistently results in better segmentation performance under a variety of natural visual distortions, including contrast, luminance or illuminants variations. However, this raises a deeper fundamental question: **Why does this happen? Why does DN lead to such improvements?**

In this chapter, we explore the reasons for the observed benefit in robustness gains. First, we investigate the reason in terms of the model's invariance under the applied image variations. Second, we analyze the internal behaviour of the Divisive Normalization layers to check where their behaviour effectively adapts to different inputs and surroundings and how it influences the feature maps extracted by the models.

4.1 Quantitative Measures of Invariance

First, we begin testing the hypothesis that the improved segmentation performance of DN-augmented models is due to an increased invariance to

input variations. Intuitively, the adaptability of Divisive Normalization may help the models to preserve their internal representations unchanged even when environmental factors alter the images. In other words, images that have been put away in the input domain due to changes in the environment would remain close in the inner representation of models augmented with Divisive Normalization. In contrast, models without DN layers will keep their inner representations further away.

To test this hypothesis quantitatively, we measured the degree of overlap (the IoU) between the model's predictions to the original and distorted input images. Analyzing how these representations change across different models and conditions, we aim to determine whether DN indeed improves the model's representational stability through invariance under these image perturbations.

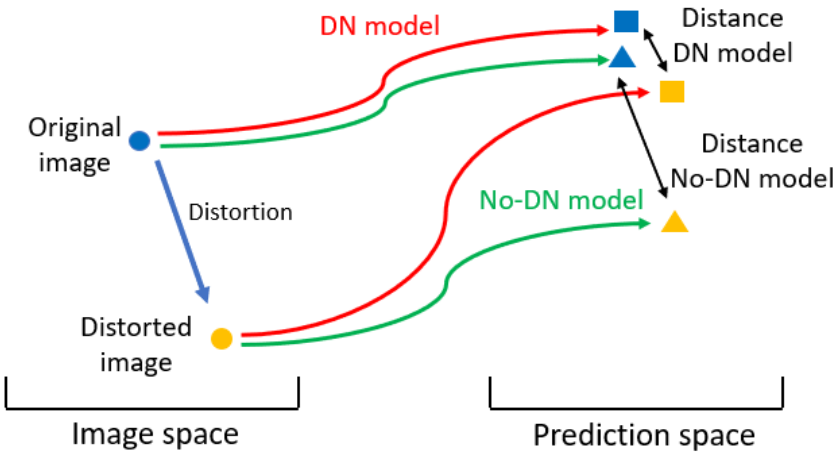


Figure 4.1: **Model invariance quantification:** If a model is invariant to a given image distortion, inputs that are far apart in pixel space due to that distortion should be mapped to nearby points in the model's prediction space. In this context, if Divisive Normalization promotes invariance, the distance between the predictions for the original and distorted images should be smaller in the DN-augmented model than in the standard model. Figure from [Hernández-Cámara et al., 2025].

To assess this, we compute the predictions of both the baseline standard U-Net and the 4-DN U-Net models for each original image and its corresponding distorted version (under fog, luminance, contrast and illuminant changes). Then, we calculate the IoU metric between the prediction for the

original image and the prediction for the modified image. This IoU score is a proxy for the "distance" between the two outputs in the model prediction space, i.e. a measure of the length of the black arrows in the inner domain shown in the figure 4.1: a bigger IoU implies more overlap and a shorter black arrow.

Table 4.1: Overlap (IoU) between original and modified images predictions under different perturbations, for models with and without Divisive Normalization. Higher values indicate greater representational stability and invariance. Table from [Hernández-Cámara et al., 2025].

Img. Variation	Standard U-Net Preds. IoU	DN U-Net Preds. IoU
Low fog	0.766	0.826
Middle fog	0.621	0.714
High fog	0.478	0.609
Achrom ctr = 0.6	0.649	0.786
Achrom ctr = 1.4	0.793	0.832
Luminance = 0.6	0.739	0.781
Luminance = 1.4	0.848	0.886
Chrom ctr = 0.6	0.842	0.844
Chrom ctr = 1.4	0.848	0.887
Angle 0	0.464	0.800
Angle 72	0.751	0.826
Angle 162	0.517	0.562
Angle 252	0.657	0.772
Angle 343	0.365	0.753

As shown in Table 4.1, across all tested image modifications (fog, luminance shifts, achromatic and chromatic contrast variations and illuminant hue and saturation changes), the models that implement DN layers consistently get higher prediction overlaps than standard models without DN. This means that the DN models produce more stable outputs when they face environmental changes, and therefore confirms that they are more invariant under these perturbations.

4.2 Where does the invariance come from? Adaptive nonlinearities

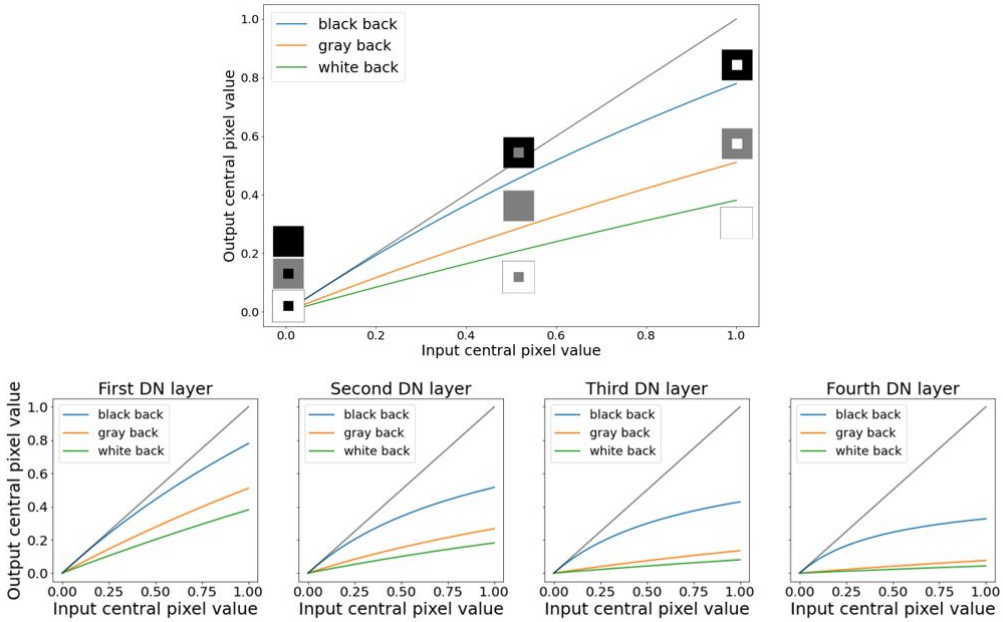


Figure 4.2: **Divisive Normalization non-linearities:** Top: Illustration of the first DN layer effect on input patches with varying central and surrounding pixels. Bottom: Response curves from the four DN model layers showing how the output of the central pixel depends on its own value and its surroundings. It illustrates that, effectively, the DN output deviates from a linear identity mapping depending on the surround. Figure adapted from [Hernández-Cámara et al., 2023, Hernández-Cámara et al., 2025].

We argue that the observed invariance in the DN models comes from the adaptivity of Divisive Normalization, specifically from how the response of each neuron is modulated by the activity of its neighbours. To illustrate this idea, we visualize the behaviour of the DN layers within a trained model. In particular, we build input patches corresponding to the 3×3 spatial neighborhood used in the kernel of the DN denominator, γ in equation 3.1. We vary the value of the central pixel from 0 to 1 (or across the dynamic range of the input feature), while holding the surrounding pixels constant at representative values: 0, 0.5, or 1.0, corresponding to dark, mid-gray, or

bright surrounds, respectively. These input patches are then passed through each of the four DN layers in the model, and we record how the central pixel's value is transformed as a function of its own intensity and the values of its neighbours. The resulting response functions are shown in Figure 4.2.

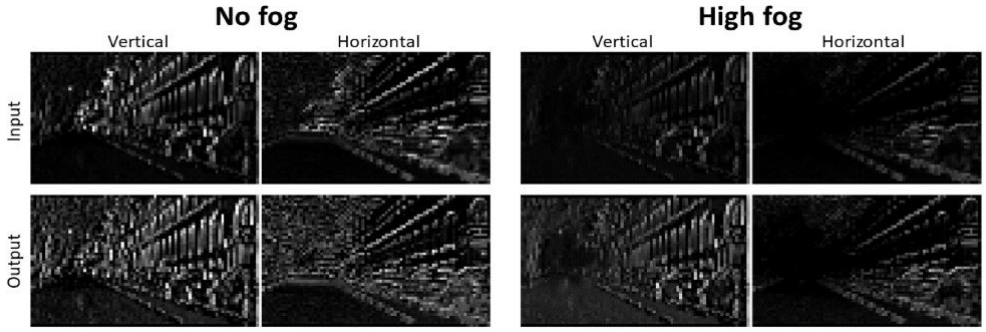


Figure 4.3: **Effect of Divisive Normalization on internal feature maps:** Visual comparison of the input and output of the second DN layer applied to two channels tuned to vertical and horizontal edges. Each column shows a good-weather image and its foggy counterpart. Top: input feature maps to the DN layer. Bottom: DN outputs showing substantial recovery of edge responses, particularly under fog. Figure from [Hernández-Cámara et al., 2023].

Interestingly, altought the γ kernel weights were randomly initialized, the training modified them to exhibit the characteristic Divisive Normalization behaviour. The resulting nonlinearity exhibits a saturating response: low-intensity values are amplified while high-intensity values are compressed. Furthermore, the presence of high values in the neighbours moderates the responses, while the presence of low values amplifies them. Notably, the nonlinearity becomes more pronounced in deeper layers, reflected by the deviation from the identity line. This progressive non-linear effect increase is due to a decrease in the β/γ ratio with the layer depth.

Finally, to understand how these adaptive nonlinearities affect the model image processing, we visualize a real example of their impact on the model feature maps. Figure 4.3 shows the effect of the second DN layer on two of the model feature maps tuned to detect vertical and horizontal edges, key features in object detection. In the first row, we observe the input feature maps to the DN layer. In the high fog condition, the central part of the feature map loses almost all its details, becoming black without any edge detection. Note

that without the DN layers, this would be the information transmitted to the following layer. In the second row we show the output feature maps after the DN is applied. They clearly show a significant recovery of the edge responses, even in the foggy condition. Therefore, the DN has a corrective effect in both clean and foggy images, although the effect is especially relevant in the high fog scenario, where it helps to restore perceptually important details that otherwise would be lost.

Part III

Evaluation of Human Perceptual Alignment

5

Alignment with Low-Level Psychophysics

5.1 Introduction to Alignment

In the previous part of the thesis, we showed that including biologically inspired computations, such as the Divisive Normalization, into deep learning models improves their performance and robustness. We also found that these improvements are due to an obtained invariance to undesired variability in the input data. In this part, we shift the focus to the second key objective of the thesis: analyzing the alignment of deep learning models and human visual perception. As stated in Part I, there are several motivations for studying this alignment.

On the one hand, from a neuroscientific perspective, deep learning models currently represent some of the best computational models of the brain. Therefore, analyzing whether their behaviour aligns with human perception can provide a better understanding of the principles underlying biological vision. For example, if a neural network trained on natural images spontaneously reproduces some human perceptual behaviours, this suggests that such behaviours may emerge naturally from the data statistics and the de-

mands and constraints of the task.

On the other hand, from a computational perspective, human alignment is not only desirable but also essential in some problems. The human brain is exceptionally robust and able to generelize across distortions, degradations and unseen data and conditions. Mimicking the human brain can help deep learning models to get these properties easily, an ongoing objective in deep learning. Moreover, there are many deep learning applications where human alignment is a requirement. For instance, in image compression, the main objective is to compress and decompress an image as much as possible while maintaining the most perceptual similarity of the reconstructed image, and this demands good models of human perception. Subjective image quality is the computer vision part that focuses on this objective, training and developing new algorithms which mimic human perception.

5.2 Low-level pshichophysics alignment

To investigate how deep learning models align with human perception, we begin with low-level visual processing phenomena. This choice is motivated because low-level psychophysics provides an ideal test case: the perceptual phenomena involved have been rigorously documented over decades, and their underlying mechanisms are relatively well understood. In contrast to higher-level tasks such as object recognition or scene interpretation, where alignment can also be evaluated, low-level phenomena are grounded in direct stimulus-response relationships. This allows for easy isolation and measurement of every phenomenon, without confounding effects from semantic or contextual complexity [Rust and Movshon, 2005, Martinez et al., 2019].

Moreover, these low-level effects are not just theoretically well-defined: they are also visually clear. Properties such as contrast sensitivity, brightness induction, or visual masking produce effects that can be seen directly in carefully designed stimuli. This makes the evaluation of human alignment (or misalignment) more interpretable, and even qualitative measures are possible. In addition, these low-level tests allow us to probe a model's internal representations without requiring task-specific training or fine-tuning, providing a clearer view of the model's perceptual biases.

In this chapter, we evaluate a collection of ten foundational visual phenomena, collectively referred to as the *Decalogue of Low-Level Perceptual*

Properties, which were formalized and presented in our CIP2022 contribution [Malo et al., 2022, Hernandez-Camara et al., 2022] and lately extended [Vila-Tomás et al., 2025]. It provides a structured framework for assessing how well a given computational model reproduces foundational psychophysical behaviours observed in human vision.

5.2.1 The Decalogue of Low-Level Visual Properties

To systematically assess how closely deep learning models replicate human low-level perception, we rely on a curated set of ten foundational psychophysical phenomena. These effects span across different dimensions of early vision, such as contrast, spatial frequency, luminance, color, and contextual interactions, and have been extensively validated through decades of human experimentation.

Each property is tested using controlled visual stimuli designed to test a well-defined perceptual response in human observers. Our goal is to qualitatively evaluate whether the models exhibit human-like behaviours. Below, we describe each phenomenon included in the *Decalogue*:

- **1. Spectral Sensitivities:**

  Human sensitivity to light varies across wavelengths, as captured by cone and opponent channel responses. This test examines whether the model reproduces the shape of these spectral sensitivity functions when exposed to quasi-spectral stimuli, both in the achromatic and chromatic axes [Hurvich and Jameson, 1957].

- **2. Brightness and Color Response Saturation:**

  Perceived brightness and color intensity do not increase linearly with physical stimulus intensity. Instead, they follow a saturating curve. This test evaluates whether the model exhibits a similar non-linear response to increasing brightness and chromatic input [Wyszecki and Stiles, 2000].

- **3. Achromatic Contrast Sensitivity Function (CSF):**

  Humans are most sensitive to spatial frequencies in the range of 3–5 cycles per degree, with reduced sensitivity at both lower and higher frequencies. This test evaluates whether the model exhibits a similar

band-pass behaviour for achromatic stimuli, matching the human CSF profile [Campbell and Robson, 1968].

- **4. Chromatic Contrast Sensitivity Function (CSF):**

  In contrast, human sensitivity to chromatic patterns, such as red-green or blue-yellow, peaks at lower spatial frequencies, i.e. it is more low-pass. This test examines whether the model exhibits similar chromatic CSF characteristics [Mullen, 1985, Díez-Ajenjo et al., 2011].

- **5. Spatio-Chromatic Receptive Fields:**

  The human brain shows spatially structured, oriented, and color and frequency-selective neurons. This test evaluates whether the model's receptive fields resemble these structured properties using local delta-like stimuli [Hubel et al., 1959].

- **6. Nonlinear Contrast Response: Saturation:**

  Human perceptual response to physically increasing contrast is not linear; it saturates. This test assesses whether the model also exhibits contrast saturation effects as stimulus contrast increases [Georgeson and Sullivan, 1975].

- **7. Nonlinear Contrast Response: Frequency Order:**

  Human sensitivity not only saturates but also depends on the test frequency. This test evaluates whether models show a similar frequency-order than humans [Georgeson and Sullivan, 1975].

- **8. Context Effects: Energy:**

  Human perception of a stimulus is attenuated when it appears on a high-contrast background. This masking effect increases with the energy (contrast) of the background. This test evaluates whether the model exhibits similar contextual suppression with increasing background contrast [Daly, 1990].

- **9. Context Effects: Frequency:**

  Masking effects in human perception are strongest when the background and the test stimulus have similar spatial frequencies. This test evaluates whether the model displays comparable frequency-specific contextual modulation [Daly, 1990].

- **10. Context Effects: Orientation:**

  Analogous to frequency-based masking, humans show stronger suppression when the background and test stimuli are similarly oriented. This test assesses whether the model exhibits orientation-specific contextual effects consistent with human vision [Daly, 1990].

Together, these ten tests provide a rigorous and interpretable framework for assessing whether deep neural networks exhibit emergent properties of human early vision. In the next section, we apply them to a U-Net segmentation model improved with Divisive Normalization to test whether it reproduces these basic human perceptual phenomena.

5.3 U-Net Alignment with Low-Level Human Perception

In the previous part, we showed that the deep learning model with Divisive Normalization (DN) had good segmentation results and improved robustness under real-world conditions. Here, we ask a different question: does this bio-inspired model also exhibit perceptual behaviors aligned with low-level human vision? To address this, we evaluate the U-Net model with DN layers using the set of ten psychophysical tests described in the *Decalogue* framework.

We applied each of the ten tests to the U-Net model and measure its response at the end of the encoder part using all the information presented here. Note that where to perform this measurement and whether to compare all the feature maps or just some of them is a specific decision we made, but other options could be explored. Based on the model activations collected, we computed the model's perceptual responses and qualitatively compared them with known human behaviours.

As an example, figure 5.1 shows two representative results from the *Decalogue*. Specifically, it shows the results of the Contrast Sensitivity Functions (experiments 3 and 4) and the orientation-based contextual masking (experiment 10). None of these results reproduces human-like behaviour: In the CSF test, the U-Net model with DN shows a clear high-pass filter behaviour, missing the band-pass human CSF in the achromatic scenario. Regarding the chromatic CSFs, the model obtains almost flat CSFs with unexpected dom-

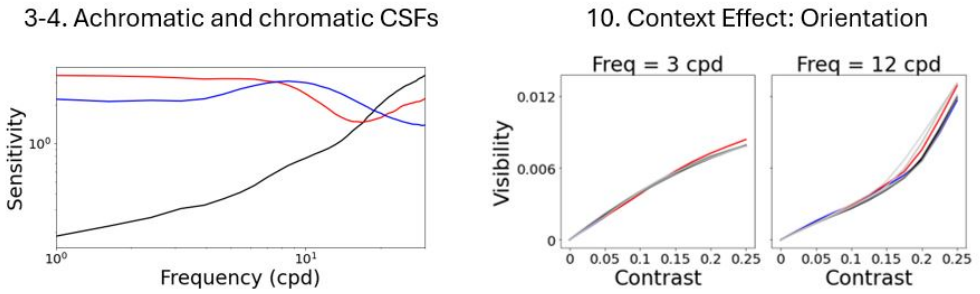


Figure 5.1: **Representative results from the *Decalogue* applied to the U-Net with Divisive Normalization:** Left: Achromatic and chromatic Contrast Sensitivity Functions (experiments 3 and 4) derived from model responses. Human observers exhibit band-pass sensitivity for achromatic contrast, peaking around 3–5 cycles per degree, and lower-frequency band-pass sensitivity for chromatic channels, with greater sensitivity in the blue channel. However, the U-Net shows high-pass behaviour for achromatic contrast and flat chromatic sensitivity curves with an unexpected dominance of the red channel. Right: Orientation-based contextual modulation (experiment 10). While human vision displays strong suppression when test and background stimuli are similarly oriented, the U-Net responses show negligible modulation across orientation conditions, suggesting a lack of orientation-selective contextual sensitivity. Figure adapted from [Hernandez-Camara et al., 2022, Vila-Tomás et al., 2025].

inance of the red channel at low frequencies. It again departs from human chromatic CSF, which shows a clear low-band filter with the yellow channel showing greater sensitivity. Similarly, in the orientation masking task, humans experience strong suppression when the test and background orientations are similar. However, the U-Net shows almost no modulation across orientations, suggesting a lack of orientation-selective context sensitivity.

When the model is evaluated in the full ten phenomena of the *Decalogue*, the results reveal a consistent pattern: despite its bio-inspired components (DN), its success in segmentation and its robustness, the U-Net with DN fails to reproduce the majority of the perceptual phenomena observed in human vision, as shown in table 5.1. Some partial alignment is observed, for instance in nonlinear contrast saturation (fact 6), frequency-dependent saturation (fact 7), energy masking (fact 8), and brightness/color saturation (fact 2). However, the model diverges significantly from human behaviour across

the majority of phenomena. In particular, it fails to capture key aspects of spectral sensitivities, both achromatic and chromatic contrast sensitivity bandwidths, and all three contextual modulation effects (orientation, frequency, and energy).

	Facts	UNet + DN
1	Spectral Sensitivities (achromatic and opponent)	✗✗
2	Brightness & Color Response Saturation	✗∼
3	Achromatic Contrast Sensitivity (Bandwidth)	✗
4	Chromatic Contrast Sensitivity (Bandwidth)	✗ ✗
5	Spatio-Chromatic Receptive Fields	✗✗∼
6	Nonlinear Contrast Response: Saturation	✗✔
7	Nonlinear Contrast Response: Frequency order	∼
8	Context effects: Energy	∼
9	Context effects: Frequency	✗
10	Context effects: Orientation	✗

Table 5.1: **Summary of the U-Net with Divisive Normalization (DN) performance across the ten *Decalogue* low-level psychophysical tests:** Each row corresponds to a fundamental visual phenomenon tested in the *Decalogue* framework. A green tick indicates the model qualitatively reproduces the human behavior; an orange ∼indicates partial or weak alignment; a red cross indicates the model fails to reproduce the expected perceptual response. Results show that, despite improved segmentation robustness, the U-Net with DN layers aligns poorly with most basic perceptual phenomena, particularly in contrast sensitivity and contextual modulation.

These results suggest that training for a high-level vision task such as segmentation does not necessarily guarantee the emergence of human-aligned low-level properties, even when including biologically inspired elements like Divisive Normalization. In contrast, models explicitly trained to match human perceptual judgments, such as PerceptNet [Hepburn et al., 2020], achieve stronger alignment across the full *Decalogue* test [Vila et al., 2022, Vila-Tomás et al., 2025].

This highlights a fundamental gap: success in robustness or accuracy on machine vision benchmarks does not imply alignment with human per-

ception. Bridging this gap may require not only architectural changes to include bio-inspired computations, but also training objectives, constraints, or inductive biases specifically designed to reflect human visual perception.

6

The Importance of Visual Environment

6.1 Color Discrimination and the Role of Visual Environment

In the previous chapter, we saw that, despite its bio-inspired design and improved robustness in segmentation, the U-Net model equipped with Divisive Normalization (DN) fails to reproduce most of the fundamental perceptual behaviours expected from human vision. However, a notable exception emerges in the case of color-related phenomena. Among the *Decalogue* tests, the model showed its highest (although partial) human alignment in tasks related to color saturation and nonlinear responses in the chromatic channels. This partial success, together with the important role that color may have for semantic segmentation, motivates a deeper investigation into how these models process color, and more specifically, how aligned they are with human color discrimination.

To evaluate the color perception, we make use of a classical human experiment from visual psychophysics: the so-called MacAdam ellipses [MacAdam, 1942]. These ellipses define regions of just-noticeable chromatic

difference in color space. That is, they represent how much a color must vary before a human observer perceives it as different. Interestingly, they depend on both the initial color and the direction of change within the color space. Therefore, to test whether the segmentation networks perceive color in a similar way compared to humans, we analyze their MacAdam ellipses. If the network "sees" color like a human, its representation of chromatic differences should provide similarly shaped and oriented ellipses.

To systematically evaluate this alignment, we conduct a controlled experiment comparing U-Net models trained under three distinct visual conditions: First, we consider models trained with natural images, representative of real-world terrestrial scenes. Second, we analyze models trained with underwater images, where the spectral distribution is heavily shifted toward blue-green wavelengths. Finally, we include a reference scenario with models trained with achromatic images, removing all color information from the training images.

For each of these three conditions, we train and evaluate two architectural variants of the same model: a standard U-Net architecture and a U-Net augmented with Divisive Normalization. This results in a total of six models, allowing us to disentangle the effects of architecture (with vs. without DN) from those of the training data (natural, underwater, or grayscale). This analysis was presented in our Frontiers in Psychology study [Hernández-Cámara et al., 2024a].

Comparing the MacAdam ellipses obtained from these networks to the human MacAdam ellipses, we aim to answer a central question: What matters more for achieving perceptual alignment in color vision: implementing biologically inspired computations, such as Divisive Normalization, or exposing models to visual environments with natural color statistics?

6.2 Assessing the MacAdam Ellipses of Deep Learning Models

To evaluate how well segmentation networks align with human color perception, we analyze their color discrimination behaviour using a methodology that allows us to compare with the classical MacAdam ellipses. As introduced before, MacAdam ellipses describe regions in the color space where chromatic variations are just barely noticeable to human observers. Their size, shape,

and orientation are well-characterized and highly consistent across observers, making them an ideal reference for evaluating perceptual alignment in artificial models.

To perform the analysis, we extend our evaluation beyond the segmentation models trained in part II. Specifically, we test both the standard U-Net architectures and the bio-inspired U-Net augmented with Divisive Normalization, trained on three different datasets with different spectral statistics. These include the Cityscapes dataset [Cordts et al., 2016], which represents real-day images with natural color distributions; the SUIM dataset [Islam et al., 2020], made of underwater images and therefore with dominant blue-green spectral bias due to water filtering; and an achromatic version of Oxford-IIIT Pets [Parkhi et al., 2012], without any color information. These datasets allow us to investigate how the spectral characteristics of the training data affect the model's color discrimination.

For each dataset, we train two models (with and without DN), resulting in six segmentation networks. We took the Cityscapes-trained models from part II [Hernández-Cámara et al., 2023, Hernández-Cámara et al., 2025], and we trained the other four models under equivalent conditions. Table 6.1 summarizes the segmentation performance (IoU) of each model on its respective test sets.

To estimate the MacAdam ellipses of each model, we follow a physically grounded procedure. For each test image, we assume it was generated from surfaces with fixed spectra reflectance illuminated by an equienergetic illuminant. Computing the image reflectance allows us to simulate illumination changes, modifying the spectral content of the light source while keeping the reflectances constant. This method allows us to generate consistent image variants across a controlled hue–saturation grid of illuminants. Figure 6.1 illustrates this procedure. Panel A shows synthetic patches of constant reflectances illuminated by different light sources with systematically varied illuminations. Panel B visualizes the hue-saturation sampling grid used for these illumination changes in the CIE 1931 xy chromaticity diagram. Panel C reproduces classical human MacAdam ellipses, which represent empirical tolerance thresholds in color space. Panel D shows an example scene created with the modified illuminants, demonstrating the application of this method across natural (top), underwater (middle), and achromatic (bottom) image domains.

	Natural illum.	Underwater	Achromatic
Standard U-Net	0.77 ± 0.02	0.66 ± 0.05	0.77 ± 0.02
U-Net + DN	0.78 ± 0.02	0.70 ± 0.04	0.80 ± 0.02

Table 6.1: **Segmentation results of standard and DN-augmented U-Net models trained in different datasets:** Test IoU results (mean $\pm$ standard deviation) of the models trained in the different environments when performing 300 evaluations over subsets of the test images. Results from [Hernández-Cámara et al., 2024a].

Once we have generated the modified scenes, we evaluate the segmentation performance of each model under these illuminant changes. For each architecture-dataset combination, we assess how the segmentation performance varies with the changes in hue and saturation, compared to the model's original performance in the unmodified images. We perform the evaluations over 300 random subsets of test images per model. Figure 6.2 shows the relative segmentation performance with regard to their original performance, and the associated 3%, 5% and 10% tolerance curves, that define each model's effective tolerance ellipse.

Note that the specific size of the tolerance regions depends on the (arbitrary) choice of the performance drop threshold used to define a "just noticeable" change (here 3%, 5%, or 10%). However, two observations are highly informative. First, the overall scale of the tolerance regions systematically varies across models trained in different visual environments, indicating that the models are more or less sensitive to chromatic variation depending on the data statistics. Second, the orientation and non-circular shape of these regions suggest that the segmentation models trained under different training distributions are anisotropic in color space, i.e. certain chromatic shifts produce more drastic changes than others. In this way, the model's behaviour mimics (or fails to mimic) the human MacAdam ellipses, offering a direct lens into the alignment of learned color representations with those of human perception.

To finally quantify the alignment between the model and human color discrimination ellipses, we compute the root mean squared error (RMSE) between the human MacAdam ellipses and the 5% tolerance regions of the models. Figure 6.3 shows histograms of the RMSE values across 300 evaluation subsets per model. We can see a clear pattern: the visual environment

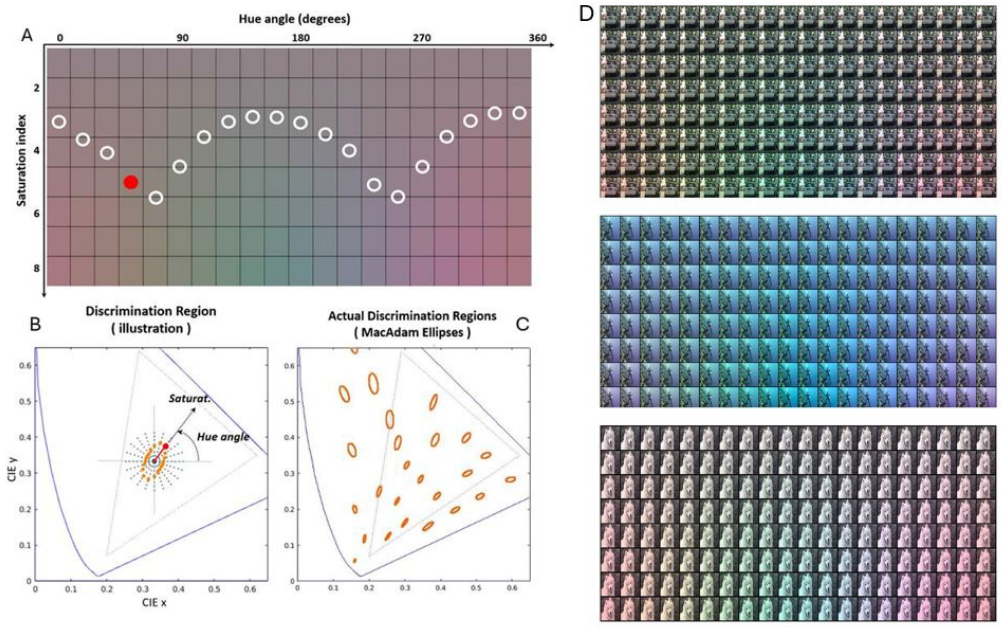


Figure 6.1: **Illustration of human color discrimination (tolerance to saturation for different hues) and our method to change the image illuminant:** Panel A shows patches of flat spectral reflectance illuminated by sources with spectral radiances selected to cover the 1931 CIE xy diagram. Black dots in the CIE xy chromatic diagram of panel B show the polar distribution of the chromaticity of the considered illuminants. The illuminants are organized as a function of hue and saturation, i.e. angle with respect to the x axis, and distance with respect to the white point respectively. For each hue (each column in the colored panel A), the Euclidean distance in the chromatic diagram required to induce certain perceptual departure from the white color of the same luminance is different. That is why the insensitivity region around the white (determined by the circles in panel A) is an ellipse with a certain orientation, marked by orange dots in panel B. The diagram in panel C displays the insensitivity regions for humans measured by MacAdam (1942) at a number of color locations over the chromatic diagram. Finally, panel D shows scenes with modified illumination starting from a different original image: natural (top), underwater (center), and achromatic (bottom). Figure adapted from [Hernández-Cámara et al., 2024a].

in which the model is trained plays a dominant role in shaping its perceptual alignment. Models trained on natural images consistently show lower RMSE values, i.e. better alignment, than those trained on underwater or achromatic

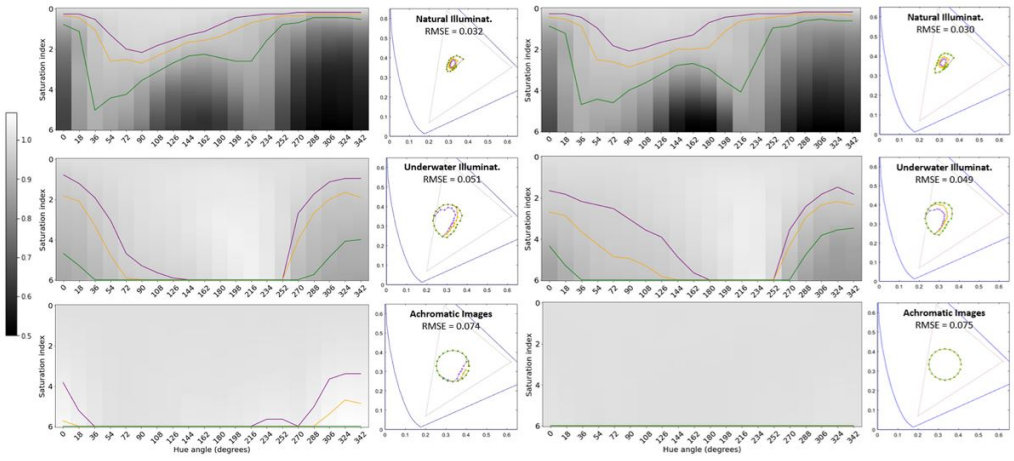


Figure 6.2: **Tolerance of segmentation performance to illuminant change for different environments:** The results of the natural, underwater, and achromatic environments are represented in the top, middle, and bottom rows respectively. Left results represent the standard U-Net models while right results represent the U-Net with DN layers. Gray level represents the segmentation performance under different illuminations with regard to the reference performance obtained for the original scenes. Darker values represent lower performance. The curves in purple, orange and green, represent variations of the performance of 3%, 5%, and 10%, respectively. These curves define tolerance regions for performance in the chromatic diagram. The RMSE values represent the distance between the average of these tolerance regions in the artificial system and the corresponding tolerance ellipse in humans. Figure adapted from [Hernández-Cámara et al., 2024a].

data, regardless of architectural differences. This observation is statistically confirmed by two-sample Kolmogorov–Smirnov tests [Hodges Jr, 1958]. The differences between environments are highly significant, with $p_{val} < 0.001$, as can be seen by the almost non-overlapping distributions across conditions. In contrast, architectural differences, with or without DN, do not produce significant differences in the underwater or achromatic conditions. Only in the natural training condition does the inclusion of DN produce a statistically significant but modest improvement in alignment.

These results demonstrate that, in the context of color discrimination, the statistical properties of the visual environment have a much greater impact on human alignment than architectural modifications alone, even when such modifications are biologically inspired.

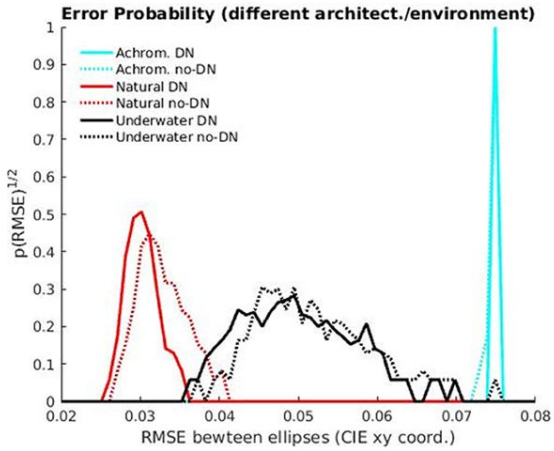


Figure 6.3: **Models' tolerance regions distances to MacAdam ellipses:** Histograms of the RMSE errors comparing the human MacAdam ellipses and the tolerance region of the models for 300 realizations with test subsets. Figure from [Hernández-Cámara et al., 2024a].

7

Analysis of the Factors that Influence the Alignment

7.1 Which factors affect the human alignment of a deep learning model?

In the previous chapter, we showed that the statistics of the visual environment play a more important role than the model architecture (bio-inspired or not) in achieving a higher human alignment, at least in the context of color discrimination. Particularly, we found that models trained on naturalistic images developed color sensitivity patterns closer to human thresholds than models trained on grayscale or underwater scenes, independently of whether the architecture was modified to include Divisive Normalization. This result raises a new fundamental question: **Which factors of deep learning models (architecture, training objective, data properties, regularization, ...) have the strongest influence on perceptual alignment with human vision?**

Understanding the impact of these deep learning key components is essential, especially if we aim to develop models that not only perform well in computer vision tasks but also process visual information in ways comparable to

human observers. In this chapter, we address this question through two complementary studies: First, we perform a systematic comparison of CNN-based architectures trained under different tasks and constraints, assessing how each condition affects perceptual alignment [Hernandez-Camara et al., 2025]. Second, we turn our attention to state-of-the-art current architecture in computer vision tasks, the Vision Transformers (ViTs). Here, we also investigate which factors of their radically different architecture affect their human alignment [Hernandez-Camara et al., 2025b].

These two analyses allow us to disentangle the effect of various deep learning factors on human alignment. Together, they offer insight into the emergence (or absence) of human-like perception in modern artificial vision systems.

7.2 Measuring Human Alignment via Image Quality Databases

Assessing the perceptual alignment of deep learning models requires systematic and quantifiable comparisons between model outputs and human judgments. In previous chapters, we used low-level psychophysical experimental data and color discrimination thresholds, i.e. the MacAdam ellipses, to evaluate this alignment. Another widely used, robust and well-established methodology to compute the human alignment is through image quality assessment (IQA) databases. These datasets offer a proxy for low- to mid-level perceptual evaluation as they provide human ratings regarding image distortions and distances, which can be compared to model-derived similarity metrics.

7.2.1 Image Quality Databases

IQA databases are composed of pairs of images: a high-quality reference image and several distorted versions of it, each affected by specific types and levels of degradation. These distortions usually include noise, blur, compression artifacts, color shifts, or contrast variations. Each distorted image is scored by human observers using Mean Opinion Scores (MOS), which represent the human perceived quality of the distorted image with respect to its reference. Higher MOS values indicate that the distortion is less noticeable (i.e., more similar to the reference), while lower scores reflect stronger

perceptual differences.

Among the most widely used IQA datasets, and the ones used in our studies, are:

- **TID-2013:** Contains 25 reference images and 3,000 distorted images, spanning 24 types of distortions at five severity levels each. Includes classic distortions like Gaussian noise, JPEG compression, and color saturation changes [Ponomarenko et al., 2015].
- **KADID-10k:** A large-scale dataset of 81 reference images and 10,125 distorted versions across 25 distortion types. It introduces more diversity in content and degradation sources [Lin et al., 2019].

7.2.2 Model-Based Perceptual Distance Computation

To evaluate a model's perceptual alignment using IQA databases, we compare the model's internal response to each pair of images, reference and distorted, and compute the distance between these internal representations. The idea is that if a model is highly aligned with humans, it should encode visual inputs in a human-like manner, and its internal distances should correlate with the perceptual similarity ratings from human observers.

Formally, let I_{ref} and I_{dist} be a pair of reference and distorted images, respectively. A deep learning model f processes these images, and we extract its internal representations at a certain layer $f_l(I_{ref})$ and $f_l(I_{dist})$. The perceptual distance at this layer, D_l, can then be computed as:

$$D_l(I_{ref}, I_{dist}) = ||f_l(I_{ref}) - f_l(I_{dist})||_2 \tag{7.1}$$

In equation 7.1, we define the perceptual distance D using the Euclidean distance, which is one of the most commonly used. However, other distances or similarity metrics can be used. For example, cosine distance, which measures the angle between the internal representation of the two images, is another common option. Regardless of the specific metric, the central hypothesis is that if a model "sees" like a human, the distances between its feature representations should be similar to the perceived dissimilarity reported by human raters. To quantify this alignment, we compute the Spearman Rank-Order Correlation Coefficient (SROCC) [Spearman, 1987] to measure the correlation between human MOS and model distances D. SROCC is

a non-parametric metric that measures the monotonic relationship between two variables, making it well-suited for comparing the perceptual judgments and the model outputs.

In summary, IQA databases provide a well-established, standardized and interpretable method to assess the human perceptual alignment of deep learning models. Comparing distances between model feature representations of distorted and reference images with human judgments, we can easily estimate how well the model captures human-like visual dissimilarity. This methodology serves as the foundation for the analyses in the next sections of this chapter, where we explore how specific deep learning components, such as training objectives, architecture types, and model depth, affect model alignment with human perception.

7.3 Factor Analysis in CNN-based Deep Learning Models

To investigate which components of deep learning models are most influential for human perceptual alignment, we first perform a systematic analysis of convolutional neural networks (CNNs), presented in our publication in Neural Networks [Hernandez-Camara et al., 2025]. In this study, we isolate a range of factors, architecture, training objective, feature extraction, and data statistics, and evaluate their effects on human alignment using the previously described methodology.

7.3.1 Experimental Design and Analyzed Factors

The goal of the analysis is to determine which design choices and training conditions lead to internal representations that correlate most strongly with human perceptual judgments from IQA databases. To do so, we evaluate a range of CNN models trained under controlled conditions, varying only one factor at a time while keeping the others fixed. Table 7.1 summarized the specific tested factors we analyze. They are grouped into three main categories:

- **Function (training objective):** We compare models trained with supervised, self-supervised, and unsupervised learning goals.
- **Architecture:** We evaluate models of different depths and structural

complexity, and we test several strategies to summarize internal feature representations, including mean, standard deviation, and Gram matrices.

- **Visual Environment (Training Data):** We analyze models trained on three datasets with distinct image statistics, resolution, and size.

In all experiments, the internal representations of the deep learning models are evaluated layer-by-layer, allowing us to analyze how alignment with human perception evolves across model depth. This is particularly relevant, as it mirrors how visual representations become increasingly abstract and task-specific in both artificial and biological systems. We also benchmark the human alignment of the deep learning models against established image quality assessment (IQA) algorithms, including Structural Similarity Index Measure (SSIM) [Wang et al., 2004], Learned Perceptual Image Patch Similarity (LPIPS) [Zhang et al., 2018], Deep Image Structure and Texture Similarity (DISTS) [Ding et al., 2020], and PerceptNet [Hepburn et al., 2020], that are specifically optimized to match human perceptual ratings from certain IQA databases.

7.3.2 Experimental Results: Role of Different Factors

Figure 7.1 shows the results of the first experiment, where we evaluate how human alignment (with TID-2013 and KADID-10K) varies with the model's training goal. We test AlexNet [Krizhevsky et al., 2012] models trained on ImageNet [Deng et al., 2009] under supervised and self-supervised settings, and ResNet [He et al., 2016] models trained on the Taskonomy dataset [Zamir et al., 2018] with various supervised, self-supervised and unsupervised objectives.

In the case of AlexNet trained on ImageNet, we find that most training objectives obtain higher alignment with human perception than SSIM. However, some have better perceptual alignment than others. In particular, the supervised model and the RotNet (self-supervised trained for rotation prediction) obtain the highest correlations with human opinion scores. However, it is important to note that the RotNet model weights were initialized from the supervised model ones. Regarding the other goals, the Jigsaw model shows strong correlations at early layers, but its alignment decreases with depth, and the other self-supervised models show lower correlations with human perception across layers.

FUNCTION	ARCHITECTURE		ENVIRONMENT
task	connections	read-out	training data
Supervised Self-supervised No supervised	AlexNet	Euclidean No concatenate	ImageNet-1K
Supervised Classif.	AlexNet VGG-16 ResNet-50 DenseNet-121 EfficientNet-B0 ConvNeXt ViT	Euclidean No concatenate	ImageNet-1K
Supervised Classif.	AlexNet	Euclidean Mean Means-Sigmas Gram No concatenate Concatenate	ImageNet-1K
Supervised Classif.	AlexNet	Euclidean No concatenate	ImageNet-1K Places-365 Cifar-10

Table 7.1: **Summary of Factors Analyzed for Human Alignment:** Each column corresponds to a design dimension evaluated in our study: *training function*, *architecture*, or *environment*. Each row lists the specific configurations explored. This structured comparison allows us to disentangle the contribution of each individual factor to the perceptual alignment of CNNs. Table from [Hernandez-Camara et al., 2025].

For the ResNet models trained on Taskonomy, we found that semantic goals (such as semantic segmentation or scene classification) together with some of the 2D tasks achieve the highest perceptual alignment, especially in deeper layers. In contrast, models trained on 3D tasks or low-dimensional geometric properties, such as depth estimation or surface normals, exhibit lower alignment. This suggests that semantic supervision benefits the mod-

els to get internal representations that better align with human perception. These results are consistent with findings from other studies that have observed the emergence of human-like phenomena, such as color categories or contrast sensitivity curves, in models trained for semantic tasks rather than for low-level tasks [Akbarinia et al., 2023, Akbarinia, 2025].

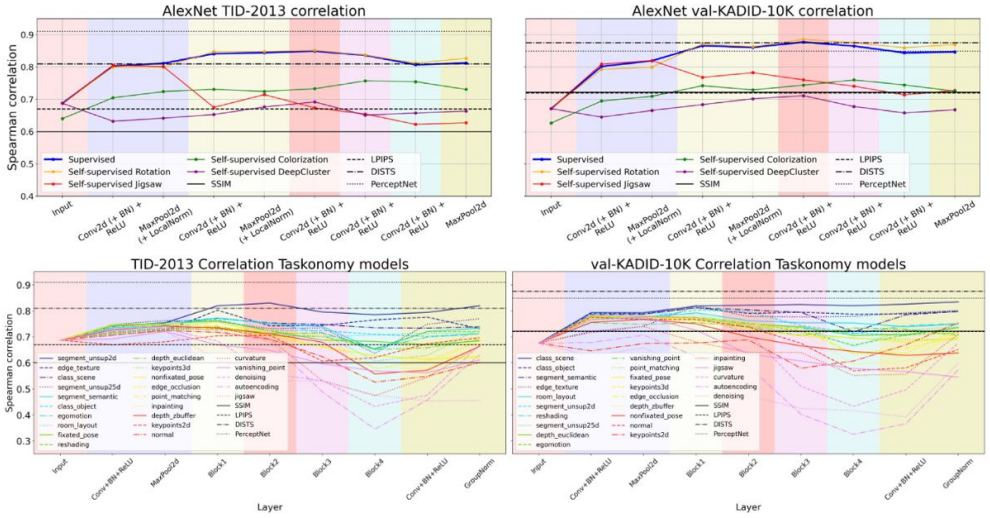


Figure 7.1: **Human alignment of CNNs trained with different objectives:** Spearman correlation between model-derived distances and human judgments across layers. Background colors denote the boundaries of model blocks. Top: Results for AlexNet trained on ImageNet with different supervised and self-supervised objectives. Bottom: Results for ResNet trained on Taskonomy with a range of visual goals. The analysis is repeated for two IQA databases: TID2013 (left) and KADID-10K (right). Models trained on semantic tasks consistently show higher alignment with human perception, especially in deeper layers. Figure adapted from [Hernandez-Camara et al., 2025]

Figure 7.2 shows the results of the second experiment, where we analyze how human alignment varies across models with different architectures. Again, we evaluate the Spearman correlations between model distances and human opinion scores for the TID2013 and KADID-10K databases. The architectures span a large range of complexity, from the relatively shallow AlexNet [Krizhevsky et al., 2012] to state-of-the-art Vision Transformers (ViT) [Dosovitskiy et al., 2020], and include VGG [Simonyan and Zisserman, 2014], ResNet [He et al., 2016], DenseNet

[Huang et al., 2017], or ConvNext [Liu et al., 2022]. All models were trained on the same supervised classification task using ImageNet [Deng et al., 2009] data. In addition to the architectural differences, we also evaluate the effect of how the image representation is read-out. Using the AlexNet model as a reference, we compare several strategies for summarizing the internal image feature maps: computing spatial means, combining means with standard deviations, or computing Gram matrices. These descriptors are applied either on a single layer or across concatenated outputs from previous layers. Several key findings emerge from these experiments. First, simple models correlate better with human perception. AlexNet and VGG-16 outperform deeper and more modern networks in perceptual alignment, despite having lower classification accuracy on ImageNet. Second, for these simplest models (AlexNet and VGG-16), correlation with human ratings increases with layer depth, peaks near the penultimate layer, and then drops in the final classification layers. Third, the read-out strategy significantly affects alignment, where the stronger the summarization, the worse the human alignment. The direct use of raw feature vectors (without any summarization) produces the best correlation. Among the tested strategies, the combination of spatial mean and standard deviation performs better than using means alone, and both outperform Gram matrices. Fourth, concatenating previous layers has limited benefit. However, it helps to obtain an upper limit of the correlation and improves the results of the last layers, whereas if one does not concatenate the features of the previous layer, the correlation goes down.

Finally, figure 7.3 shows the results of the third experiment, where we analyze how the training data affects perceptual alignment. Particularly, we fix the model architecture to AlexNet and the training objective to supervised classification, but vary the training datasets: ImageNet-1K [Deng et al., 2009] (1.2M diverse natural images), Places-365 [Zhou et al., 2014] (10M place scene images), and CIFAR-10 [Krizhevsky et al., 2009] (50K natural small low-resolution images of 10 categories). The results show that the ImageNet-trained model achieves the highest correlation with human perception, followed by Places-365, while CIFAR-10 obtains the lowest alignment. This confirms that training on larger and more naturalistic datasets benefits the human alignment, probably due to the richness and variety of visual statistics they include.

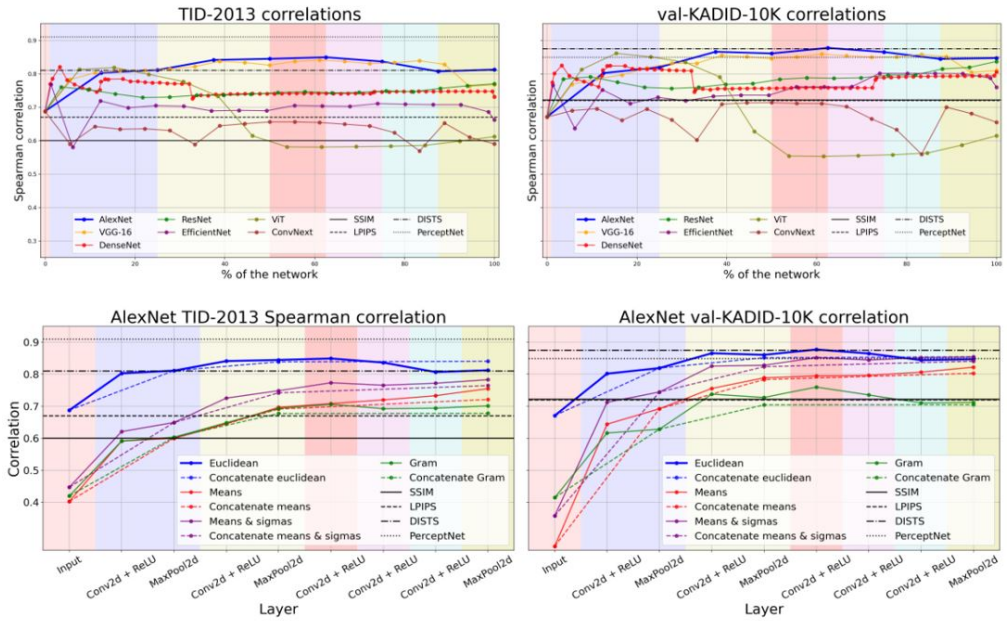


Figure 7.2: **Human alignment results across different model architectures and read-out strategies:** Top: Spearman correlation between model and human judgments for CNNs and a ViT trained for supervised classification on ImageNet. Simple models like AlexNet and VGG-16 align better with human perception than deeper modern architectures, despite lower classification accuracy. Bottom: Analysis of different read-out strategies applied to AlexNet. Raw feature vectors achieve the best alignment; Gram matrix summarization performs the worst. Background color indicates AlexNet blocks. Results are shown for TID2013 (left) and KADID-10K (right) IQA datasets. Figure adapted from [Hernandez-Camara et al., 2025].

7.4 Disentangling the Human Alignment in Vision Transformers

Following our previous study and motivated by the growing dominance of Vision Transformer (ViT) architectures [Dosovitskiy et al., 2020] in computer vision, we extend our analysis to investigate the perceptual alignment of ViTs. Note that Vision Transformers represented a significant architectural shift from convolutional neural networks (CNNs). Rather than relying on local receptive fields and hierarchical spatial composition, ViTs use global self-attention mechanisms that allow each image patch to directly interact with

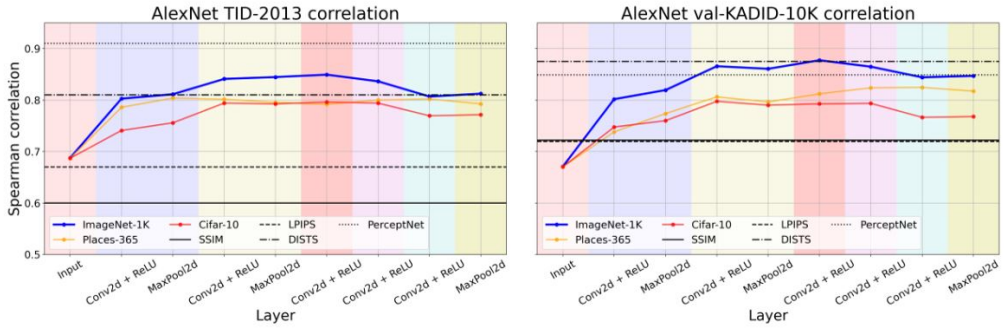


Figure 7.3: **Human alignment results for AlexNet models trained with different datasets:** Spearman correlation between model-predicted distances and human perceptual scores across model layers. All models were trained for supervised classification, but using different datasets. Models trained on ImageNet exhibit the highest alignment, followed by those trained on Places-365 and CIFAR-10. Results are shown for both the TID2013 (left) and KADID-10K (right) image quality databases. Figure from [Hernandez-Camara et al., 2025].

all others. While these models have achieved state-of-the-art performance across a wide range of vision tasks, their alignment with human perceptual judgments is still not well analyzed. In this study, we aim to examine whether ViTs develop human-like perceptual representations, and how this alignment is influenced by various training and architectural factors. This work builds on our CCN 2025 poster Do Vision Transformers See Like Humans? Evaluating their Perceptual Alignment [Hernandez-Camara et al., 2025b].

Particularly, to perform this analysis, we make use of a collection of pre-trained ViTs for the goal of image classification [Steiner et al., 2021]. This collection includes over 50,000 models trained under different configurations, allowing us to systematically explore how variations in architecture and training parameters affect perceptual alignment. We focus on the factors that vary in this ViTs collection, such as the model size (Tiny, Small, Base and Large), dataset size (number of unique images used during training), samples seen (how many times each image is seen during the training), intensity of data augmentation (strength of augmentations applied) and regularization (effect of techniques such as dropout and stochastic depth).

Figure 7.4 shows the alignment results for different ViT training configurations. In all cases, we compute the perceptual distance with TID-2013

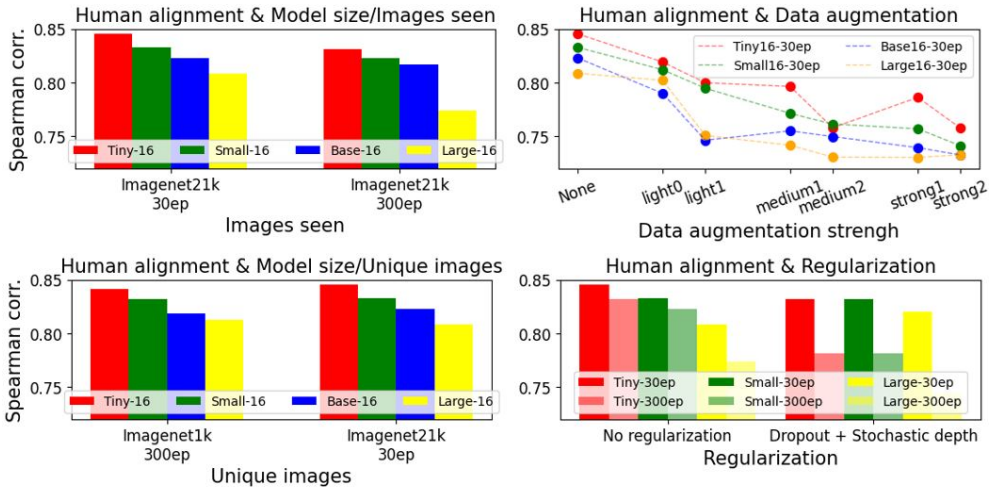


Figure 7.4: **Human alignment results of Vision Transformer (ViT) models across different training factors:** Plots show the Spearman correlation between model-based distances and human opinion scores from the TID2013 dataset. Top left: Alignment as a function of total samples seen (i.e., training duration) for fixed dataset size. Bottom left: Alignment across datasets of increasing diversity (from 1.3M to 13M unique images), for a fixed total compute. Top right: Impact of data augmentation strength. Bottom right: Impact of regularization (none vs. dropout + stochastic depth). Colors indicate different model sizes. Figure adapted from [Hernandez-Camara et al., 2025b].

using the final model representation before the classification head, i.e., the output of the transformer encoder. Several key insights emerge from this analysis. First, similarly to CNN, model size reduces the human alignment. Larger ViTs consistently showing lower perceptual alignment with human judgments, regardless of dataset or training configuration. Second, more training does not help to get higher alignment. Increasing the number of times each image is seen during training (i.e., training duration) leads to lower perceptual alignment. This effect is most evident in larger models, suggesting that prolonged optimization produces overfitting and pushes models away from human-like representations despite improving task performance. This supports prior findings on CNN, which showed that prolonged training and, therefore, more accuracy on the training objective can reduce alignment with human perception [Gomez-Villa et al., 2020, Li et al., 2022, Kumar et al., 2022, Hernandez-Camara et al., 2025]. Third, increasing the

dataset size has a limited effect. Increasing dataset diversity (from 1.3M to 13M unique images) has surprisingly almost no effect on the perceptual alignment when the total number of images seen is constant. This indicates that longer training (i.e., more samples seen) is a stronger factor than dataset variety. Fourth, stronger augmentations and regularization (dropout and stochastic depth) hurt human alignment. Although these techniques are effective for improving generalization and robustness, they systematically degrade human alignment across all models, particularly for the models with extensive training, i.e. the ones that have seen each image more times. This suggests that common strategies to improve model generalization may move models further from the human perception.

7.5 What Drives Human Alignment in Deep Models?

The results presented in this chapter allow us to obtain several conclusions regarding the factors that most strongly influence human alignment in deep learning models. Our findings indicate that getting a high perceptual alignment requires a deep understanding of how different deep learning components contribute to human-like behaviour and how they interact. Here we present a summary of our findings:

Training Objective Is Crucial

Across CNNs we observe that models trained for semantic objectives, such as classification or segmentation, exhibit higher alignment with human perception than models trained for geometric or low-level tasks. In particular, supervised models consistently outperform their self-supervised or unsupervised counterparts, especially in mid-to-late layers. This suggests that optimizing for meaningful, semantic labels drives the network to learn perceptually relevant features.

Simpler Architectures Align Better

Contrary to what one might expect, model performance on the training objective does not directly correlate with human alignment. Simpler CNN architectures like AlexNet and VGG-16 exhibit higher perceptual alignment than deeper and more complex architectures such as ResNets, DenseNets or ConvNeXts. Also, smaller ViTs align better with human perception than

larger models. This points to a trade-off between maximizing task accuracy and maintaining interpretable and human-like representations.

Layer Depth Matters

Human alignment tends to increase across early and intermediate layers of CNNs but declines in the deepest task-specific layers. This mirrors a functional hierarchy of the human brain, where early stages capture general-purpose visual features, while deeper layers become increasingly specialized. Therefore, alignment assessments should consider the layer being analyzed, rather than relying on a single readout.

Visual Environment Plays a Dominant Role

As shown here and in the previous chapter, the statistics of the training data can be more influential than the architectural choice. Models trained on natural scenes show higher alignment with IQA databases and developed color discrimination patterns more similar to human MacAdam ellipses than models trained on grayscale or underwater imagery. This reinforces the idea that exposure to naturalistic environments is key to achieving human-like perception.

Overtraining Reduces Alignment

In our analysis of Vision Transformers, we found that longer training, measured as the number of times each image was seen, reduced alignment with human judgments. This suggests that overfitting to the training objective causes the model to reduce its human alignment in favour of task-specific accuracy.

Standard ML Techniques Can Be Detrimental

Interestingly, widely used techniques to improve generalization in machine learning, such as data augmentation and regularization, consistently reduce perceptual alignment in ViT models. While they help models perform better on downstream tasks, they seem to push internal representations further away from human perception.

Readout Strategies Influence Alignment

Finally, we observe that how internal representations are extracted affects the alignment scores. Although it is secondary to training objectives or data properties, this factor still matters in the design of evaluation frameworks.

Taken together, these findings show a complex picture. While architec-

tural and computer vision innovations and techniques, such as ViTs, regularization or data augmentation, can boost models' performance and robustness, they are not sufficient on their own. Instead, we showed that they mainly reduce human alignment. We also found that the most decisive factors in driving human alignment appear to be the training objective, the data distribution and the architecture complexity. Importantly, these influences are not always additive, with many of them interacting in non-obvious ways.

To illustrate this non-trivial relationship, figure 7.5 shows the relation between model performance (on an objective task like ImageNet classification) and their perceptual alignment (measured using IQA databases). Clearly, the relation is not linear: after an initial increase, alignment peaks and then declines as models continue to optimize their task accuracy. This creates an inverted-U shape, indicating that training the models too much to get a high task-specific performance can reduce their similarity to human perception. This phenomenon is consistent with similar trends reported in autoencoder models and ResNets trained for reconstruction or contrastive objectives [Li et al., 2022, Kumar et al., 2022]. This shows that human alignment in deep learning is not an automatic by-product of better task performance or architectural complexity. Instead, it emerges from a careful balance of the right objectives, data distributions, and inductive biases.

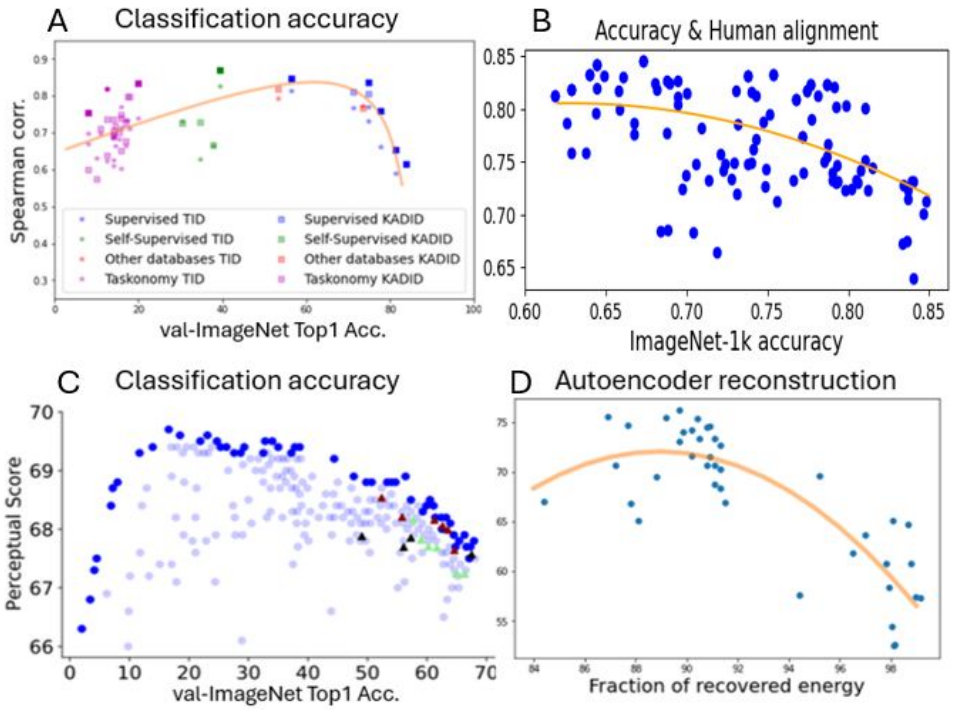


Figure 7.5: **Non-linear relation between image classification accuracy and human alignment:** Plots show the relation between human alignment and the model's accuracy on different tasks, the majority of which are ImageNet classification. Panel A: our results of correlation with the human opinion of distortion (both in TID-2013 and KADID-10K), depending on the validation ImageNet classification accuracy for all the different considered networks in [Hernández-Cámara et al., 2024a]. The points of the Pareto frontier have been highlighted in darker color. Panel B: our results of correlation with the human opinion of distortion in TID-2013, depending on the validation ImageNet classification accuracy for all the different considered networks in [Hernández-Cámara et al., 2024a]. Panel C: result from [Kumar et al., 2022], with the Pareto frontier showing an equivalent correlation between the validation ImageNet classification accuracy and the human behaviour (reproduced with permission of the authors). Panel D: results [Li et al., 2022] showing similarity with human Contrast Sensitivity Function (CSF) of autoencoders depending on the performance in several image reconstruction tasks (from data in tables 1-4 of [Li et al., 2022]). Figure adapted from [Hernandez-Camara et al., 2025, Hernandez-Camara et al., 2025b].

8

Alignment Evaluation for Multi-Modal Models

8.1 Multimodal Models Evaluation

In the previous chapter, we showed that the alignment between deep learning models and human perception is shaped by multiple interacting factors, including the model architecture, training objective, layer depth, or visual statistics of the training data. We also saw that the human alignment is not a fixed property, but that it varies across layers and evolves over the training. In particular, we found that as models become more optimized for task-specific objectives, they often become less aligned with human perception. These findings raise a key question regarding modern deep learning multimodal models: how do these multimodal models, trained jointly on vision and language, process visual information, and how aligned are their internal representations with human perception?

This question is especially relevant given the growing influence of multimodal vision-language models in both research and, obviously, in real-world applications. Models such as CLIP (Contrastive Language–Image Pretraining) [Radford et al., 2021b], and more recent Multimodal Large Language

Models (MLLMs) integrate visual and linguistic data to perform a wide variety of tasks, from image retrieval or classification to open-ended visual question answering. However, despite their impressive capabilities and their extended use, little is known about how these models process visual input in relation to human perception.

In this chapter, we explore the perceptual alignment of multimodal models, beginning with CLIP and extending to modern MLLMs. We organize the chapter into three complementary analyses: First, we examine how alignment with human perception varies across CLIP's layers and at multiple perceptual tasks of increasing abstraction. This analysis explores whether the depth at which alignment peaks depends on the nature of the perceptual phenomenon. Second, we study how CLIP's alignment evolves throughout training, revealing a different trend compared with vision-only models. Finally, we extend our analysis to modern MLLMs and introduce a psychophysics-inspired evaluation framework to assess contrast sensitivity without requiring explicit feature read-out, making the methodology suitable for end-to-end systems like conversational multimodal models.

Together, these studies offer new insights into how vision-language models process visual and textual information and provide novel tools for evaluating human alignment in increasingly complex and abstract systems.

8.2 CLIP Alignment at Different Abstraction Levels

Contrastive Language–Image Pretraining (CLIP) is a multimodal model trained to align visual and textual inputs through a contrastive loss function [Radford et al., 2021b]. Specifically, CLIP is trained to project paired images and text descriptions into a shared embedding space, such that corresponding pairs are mapped closer together. This training strategy enables CLIP to achieve strong zero-shot and few-shot performance on a variety of downstream tasks, including image classification and retrieval. Thanks to its scalability and extensive training data, CLIP has become one of the most widely adopted visual encoders in modern multimodal systems. However, beyond its performance on standard benchmarks, an important open question remains: to what extent do CLIP's internal visual representations align with human perception?

This question is particularly relevant because CLIP's training includes both visual and language information. The integration of linguistic supervision may shape how the model encodes perceptual features. Understanding whether this integration improves or harms alignment with human vision is crucial. As shown in the previous chapter, the layer at which alignment is measured plays a key role: we found that perceptual alignment often increases with depth, peaking in intermediate-to-deep layers before declining in task-specialized final layers. Motivated by this, we conducted a layer-by-layer analysis of CLIP to determine how human alignment varies across depth and abstraction levels. This study, presented at the ICLR 2024 Re-Align Workshop [Hernández-Cámara et al., 2024b], aims to reveal how human-like perception emerges (or fails to emerge) within different stages of the CLIP architecture. Importantly, human perceptual alignment can be evaluated at multiple levels of abstraction. For instance, measuring perceptual distances between an image and its slightly noisy version (TID2013 [Ponomarenko et al., 2015]) involves relatively low-level processing. In contrast, evaluating semantic similarity between different object categories (THINGS [Hebart et al., 2020]) focuses on higher-level visual representations. Figure 8.1 illustrates the range of abstraction levels used in our study, and the types of image variation and behavioural judgments that define each.

Figure 8.1: **Levels of abstraction in perceptual alignment tasks:** Visual examples of the three abstraction levels used to evaluate human-model perceptual alignment. Low abstraction (left): Pairwise similarity judgments for images with slight distortions from datasets such as TID2013 [Ponomarenko et al., 2015]. Middle abstraction (center): Triplet judgments between images of the same semantic class, differing in object count or composition (NIGHTS perceptual dataset [Fu et al., 2023]). High abstraction (right): Triplet odd-one-out judgments for semantically different classes (THINGS database [Hebart et al., 2020]). Figure from [Hernández-Cámara et al., 2024b].

To systematically explore these differences, we evaluated CLIP across

three distinct perceptual datasets, each corresponding to one of the abstraction levels. Figure 8.1 shows an example from each of the abstraction levels considered, which are listed below:

- **Low abstraction:** TID-2013 consists of reference images paired with distorted versions degraded by low-level perturbations such as noise, blur, compression, or color shifts. Human observers rated the perceived visual quality of each distorted image with regard to its reference using the Mean Opinion Scores (MOS). It provides a perfect ground truth to evaluate human alignment with low-level perceptual similarity judgments [Ponomarenko et al., 2015].

- **Mid abstraction:** NIGHTS perceptual triplets feature three images from the same object category. One image serves as a reference, and humans judge which of the other two is more visually similar to it. Differences include layout, pose, lighting, or quantity of objects, all intra-class but mid-level semantic variations [Fu et al., 2023].

- **High abstraction:** HINGS odd-one-out triplets contain three images from distinct semantic categories. Human participants are asked to identify the image that does not belong, the "odd one out", based on high-level conceptual or categorical similarity [Hebart et al., 2020].

For each dataset, we extracted image features from every layer of CLIP's visual encoder and computed model distances between image pairs or triplets. Alignment was then assessed by comparing model predictions with human judgments: using Spearman correlation for low-level (MOS-based) scores, and triplet accuracy for mid- and high-level tasks. Crucially, this analysis focuses only on CLIP's vision component, the image encoder, excluding the textual CLIP part. This ensures that the measurements reflect the structure of visual representations independently of language fusion.

We also evaluated how key design and training parameters of CLIP most strongly affect the human perceptual alignment. Specifically, we analyzed multiple architectural variants and training configurations: Architecture size and patch resolution (ViT-B/16, ViT-B/32, and ViT-L/14), activation function (standard softmax vs. SigLIP sigmoid [Zhai et al., 2023]), training language (English vs. Chinese captions [Yang et al., 2022]), and training domain (natural images vs. medical image datasets [Zhang et al., 2023]). Each

factor was varied independently while all other variables were held constant as shown in table 8.1. As a reference, we included IQA perceptual algorithms such as SSIM [Wang et al., 2004], LPIPS [Zhang et al., 2018], and PerceptNet [Hepburn et al., 2020], which serve as baselines for alignment across layers and abstraction levels.

ARCHITECTURE DESIGN		**TRAINING PROCEDURE**	
Model size	Last activation funct.	Languages	Data type
base-patch16 base-patch32 large-patch14 large-patch14-336	Softmax	English	Natural images
base-patch16	Softmax Sigmoid	English	Natural images
base-patch16	Softmax	English Chinese	Natural images
base-patch16	Softmax	English	Natural images Medical images

Table 8.1: **Analyzed factors for perceptual alignment in CLIP**: Design and training variables are grouped into two broad categories: *Architecture Design* and *Training Procedure*. Each column corresponds to a factor explored in the study, and each row lists the specific settings tested. All comparisons were conducted independently to isolate the effect of each variable on human alignment. Table from [Hernández-Cámara et al., 2024b].

We begin by analyzing how human alignment varies across different CLIP architectural configurations. In particular, we examine the effects of the number of layers (base vs. large), the patch size (e.g., ViT-B/16 vs. ViT-B/32), and the number of patches (ViT-L/14 vs. ViT-L/14-336). Figure 8.2 (top row) shows the results of this analysis across the three abstraction levels. Among the tested configurations, the base-patch32 variant consistently achieves the best perceptual alignment, especially at the low abstraction level. Two trends are clearly visible: (1) For fixed image sizes, larger patch sizes improve alignment (blue curve outperforms yellow), and (2) for fixed patch sizes, smaller images yield better results (green outperforms purple). Both observations suggest that there are optimal scales for spatial

analysis, and when the scene is partitioned into overly small regions, the model's encoding becomes less human-like. This supports prior work on the importance of scale in feature extraction and region-based representations [Lowe, 2004, Lazebnik et al., 2006].

Interestingly, alignment across depth also varies with abstraction. In low abstraction tasks, alignment increases sharply around the 10% depth mark. For mid abstraction, the alignment improvement appears closer to 30% depth, while for high abstraction, alignment remains low until around 50% of the model depth. This aligns with previous findings in both CNNs and ViTs, where feature complexity increases progressively through the network [Zeiler and Fergus, 2014, Ghiasi et al., 2022]. These results suggest that CLIP also encodes a hierarchy from low-level to high-level perceptual features.

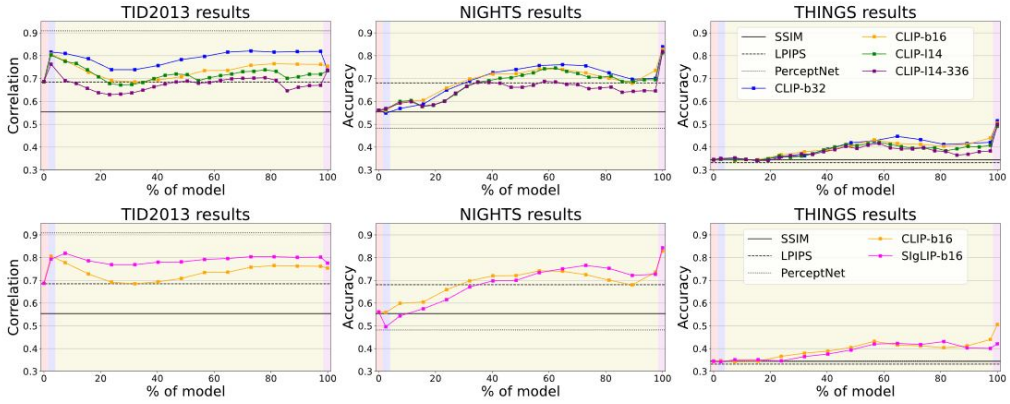


Figure 8.2: **CLIP alignment results across architecture variants:** Top row: Spearman or accuracy-based alignment scores (depending on abstraction level) for CLIP variants differing in depth, patch size, and number of patches. Bottom row: Alignment for CLIP (softmax-based) and SigLIP (sigmoid-based) across three abstraction levels. Each curve traces alignment scores through the visual encoder's layers. Figure adapted from [Hernández-Cámara et al., 2024b].

Next, we evaluate the effect of the activation function used during CLIP training. As shown in Figure 8.2 (bottom row), the SigLIP variant, trained with a sigmoid contrastive loss, achieves notably higher alignment than standard CLIP in the low abstraction regime (TID2013). In mid-level abstraction tasks, CLIP shows stronger alignment in earlier layers, but SigLIP performs

better in the deeper ones. At the high abstraction level, both models show similar behaviour until the final layers, where CLIP slightly outperforms SigLIP. One particularly notable observation is the sharp decrease in alignment at the SigLIP's first visual layer (embedding) in mid-level tasks, an effect not present in the low-level scenario. This may relate to how the activation non-linearity shapes the output space: in CLIP, softmax promotes confident, high-contrast scores; in contrast, SigLIP's sigmoid activation avoids output saturation, which may improve robustness and sensitivity to subtle differences, particularly relevant under noisy or out-of-distribution conditions.

We then turn to training procedure variations, beginning with the language used to caption the image-text pairs. In Figure 8.3 (top row), we compare CLIP models trained with English and Chinese captions. In the low abstraction setting, the model trained with Chinese captions exhibits slightly higher alignment, especially in early layers. However, this difference disappears at mid and high abstraction levels. It is important to note that these models come from different organizations and may differ in more than just language, including datasets, tokenization, training regimes, and augmentation strategies. Therefore, the observed differences may not be driven purely by language, but could reflect other hidden confounding features.

Finally, we compare models trained on different image domains. Figure 8.3 (bottom row) compares a CLIP model trained on natural image-text pairs (nat-CLIP) with one trained on medical data (med-CLIP). In the low abstraction, med-CLIP is substantially more aligned with human judgments than nat-CLIP. We hypothesize that this is due to the nature of medical images, which often contain fine-grained textures and subtle contrast variations, features particularly relevant for tasks like TID2013. Interestingly, this advantage diminishes with higher abstraction. In the mid and high levels, nat-CLIP becomes more aligned in the early layers, though med-CLIP regains superiority in the final layers. This suggests that medical data may help tune the network to better model local texture and noise sensitivity (important for low-level perception), but that natural images provide better grounding for broader semantic organization.

Taken together, these results show that the architecture, activation function, training language, and data domain all influence CLIP's alignment with human perception, but their effects vary across abstraction levels and net-

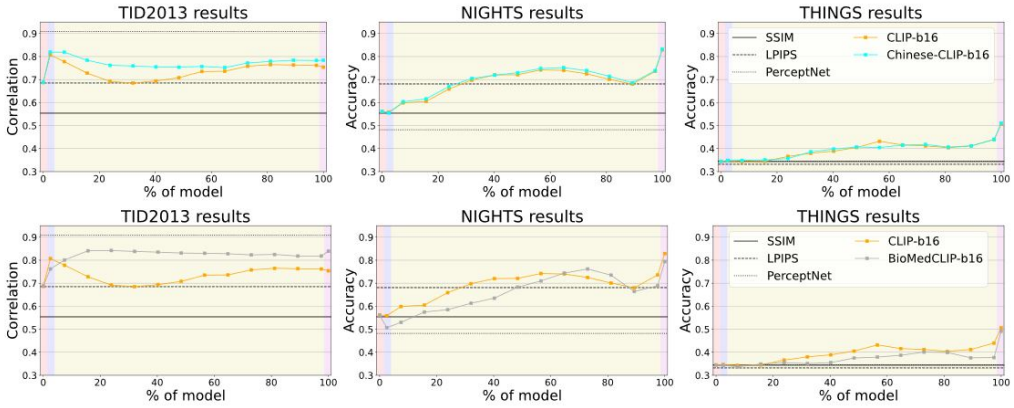


Figure 8.3: **CLIP alignment results across training procedures:** Top row: Alignment comparisons between CLIP models trained with English vs. Chinese captions. Bottom row: Alignment differences for models trained with natural image data (nat-CLIP) vs. medical imaging data (med-CLIP). Performance is shown separately for low-, mid-, and high-abstraction alignment tasks. Figure adapted from [Hernández-Cámara et al., 2024b].

work depth. The depth-alignment curves further reinforce the view that perceptual alignment is a layered, multiscale property that emerges in different parts of the network depending on the perceptual complexity of the task, model architecture and training data.

8.3 Evolution of Low-Level and Texture Human-CLIP Alignment

In the previous section, we explored how perceptual alignment in CLIP varies with network depth and the level of abstraction in the evaluated task. We now focus on a different but equally important question: how does this alignment evolve over training? That is, how does the human alignment of CLIP change across the course of optimization, and what internal mechanisms may explain its dynamics? Would it be similar to vision-only CNNs and ViTs which showed that highly trained models are less human aligned?

To investigate this, we track the evolution of a ViT-B/16 CLIP model throughout training using checkpoints from OpenCLIP [Cherti et al., 2023]. This work, presented at CCN 2025 [Hernandez-Camara et al., 2025c], eval-

uates the training dynamics of CLIP's perceptual alignment, from random weights initialization (epoch 0) to convergence (epoch 65), focusing on low-level vision, texture bias, and robustness.

Particularly, we tracked four metrics over the full CLIP training trajectory:

- **Classification Accuracy:** We evaluate the zero-shot classification accuracy on Cifar100 [Krizhevsky et al., 2009]. For each image, we compute the similarity between the image representation and textual representation of the possible classes. This serves as a proxy for the model's semantic abstraction capabilities.

- **Level Perceptual Alignment:** We measure human perceptual alignment using the TID2013 image quality assessment database [Ponomarenko et al., 2015]. For each epoch, we compute the similarity between each pair of original and distorted images in the model's embedding space. We then correlate these similarity scores with the human Mean Opinion Score (MOS) to quantify the alignment between model predictions and low-level human perception.

- **Texture Bias:** We quantify the model texture bias using the Geirhos Texture-Shape Bias dataset [Geirhos et al., 2018a]. For each conflict image (an image with a shape from one class and a texture from another), we compute the image's similarity with two textual descriptions corresponding to the shape and texture classes. The classification is based on which text has higher similarity, determining whether the model classifies by shape or texture.

- **Noise Sensitivity:** We evaluate the model's noise sensitivity by measuring the relative accuracy drop when Gaussian noise is introduced in the images. Specifically, we compare the zero-shot classification accuracy on the clean CIFAR-100 images with the accuracy on the corrupted CIFAR-100- C dataset with Gaussian noise [Hendrycks and Dietterich, 2019]. This relative accuracy drop reflects the model's sensitivity to image perturbations.

Figure 8.4 summarizes the evolution of these four metrics across training epochs. The results reveal a coherent pattern: Low-level perceptual alignment (TID2013 correlation) peaks in the early stages of training and then

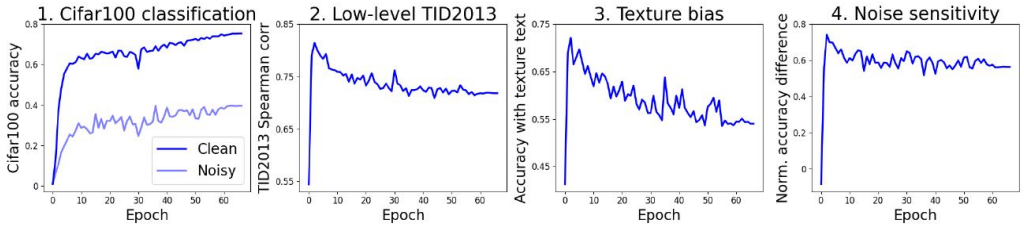


Figure 8.4: **Evolution of CLIP model metrics throughout training epochs:** Panel 1: Zero-shot classification accuracy on clean and noisy CIFAR-100. Panel 2: Correlation with human MOS scores in TID2013 (low-level perceptual alignment). Panel 3: Shape vs. texture classification preference on shape-texture cue stimuli (texture bias) [Geirhos et al., 2018a]. Panel 4: Accuracy drop under Gaussian noise. A peak in perceptual alignment, texture bias, and noise sensitivity is observed early in training, followed by a shift toward robustness and semantic performance. Figure from [Hernández-Cámara et al., 2024b].

gradually declines. Texture bias is strongest at the beginning of training and decreases over the epochs, indicating a transition from texture-based to more shape-based representations. Noise sensitivity follows the same trajectory as the previous metrics; early models are more sensitive to noise (larger accuracy drops), while later models become more robust. This trend aligns with the model's shift from texture-based to shape-based representations: as classification decisions become more dependent on global shape information, local pixel-level distortions, such as those introduced by noise, have less impact on the model performance. Finally, classification accuracy steadily improves throughout training, showing no drop or saturation during the training interval. Together, these trends point to a shared underlying mechanism: at early stages, CLIP learns fine-grained, low-level features that align closely with human perception and local distortions. However, as training proceeds, the model prioritizes more abstract, shape-based features to enhance semantic generalization and robustness, but this comes at the cost of perceptual alignment. This phenomenon is also reflected in the high correlation between alignment, texture bias, and noise sensitivity (Table 8.2).

These findings illustrate a fundamental trade-off between low-level alignment and high-level robustness in the training dynamics of CLIP. The model begins by relying on texture cues, which are aligned with human sensitivity to fine detail, but eventually shifts toward shape-based abstraction to improve

	Classification Accuracy	TID2013 Correlation	Texture Bias	Noise Sensitivity
Class. Accuracy	1			
TID2013 Corr	-0.795	1		
Texture Bias	-0.737	0.852	1	
Noise Sens.	-0.441	0.836	0.618	1

Table 8.2: **Pearson correlation between the four metrics across epochs of CLIP training:** Strong positive correlations between low-level perceptual alignment, texture bias, and noise sensitivity confirm a shared dynamic and help explain the model's transition from low-level to high-level representations. Table from [Hernandez-Camara et al., 2025c].

performance on semantic tasks. Interestingly, this contrasts with standard vision-only models. For example, CNNs have been shown to have a high texture bias even at the end of their training unless explicitly regularized to avoid it [Geirhos et al., 2018a]. In CLIP, however, the joint vision-language training appears to drive the model toward greater shape bias, in concordance with recent results suggesting that multimodal models are more shape-biased than their vision-only counterparts [Gavrikov et al., 2024]. Yet, while this improves robustness and semantic generalization, it moves the model further away from human-like low-level processing.

These observations raise a broader question about the alignment of even more advanced multimodal systems. As models evolve from dual-encoder architectures like CLIP to Multimodal Large Language Models (MLLMs) capable of open-ended visual reasoning and conversational interaction, their internal structure becomes increasingly opaque. Evaluating perceptual alignment in such models is especially challenging: traditional methods based on intermediate feature read-outs may no longer apply, as even many of these models are not open source. Moreover, from early chapters, we found that the read-out mechanism affects the human alignment score. To address these problems, in the next section we turn to an alternative evaluation strategy inspired by psychophysics, focusing on a fundamental visual property, contrast sensitivity, and asking whether modern MLLMs exhibit the same basic perceptual patterns that characterize human vision.

8.4 Evaluating Contrast Sensitivity Function of Multimodal Vision-Language Models

In the previous sections, we showed that human alignment in deep learning models depends on multiple factors, including architecture, training objectives, data statistics, and also, notably, the specific strategy used to extract internal representations. Moreover, we observed that perceptual alignment changes throughout the model hierarchy and training stages. For example, we found that multimodal CLIP models showed a reduction in their low-level alignment as their high-level semantic performance improved. However, all the models analyzed, from CNN to ViT and CLIP-like, still operate within a relatively vision-only or constrained vision-language encoding framework.

In contrast, modern Multimodal Large Language Models (MLLMs) represent a more flexible and complex class of architectures. These systems are not only trained to process visual inputs but are also optimized for conversational interactions, integrating vision and language to perform open-ended tasks through generative responses. However: Do these models still capture key low-level visual features, such as contrast and spatial frequency? To what extent do their generative decisions reflect sensitivity to early perceptual factors that are fundamental to biological vision?

Addressing these questions is non-trivial. Existing methods for evaluating low-level alignment, such as measuring contrast sensitivity, typically rely on access to internal features [Li et al., 2022, Cai et al., 2025] or require classifiers trained on model internal representations [Akbarinia et al., 2023]. These approaches assume a fixed read-out strategy and impose additional constraints that limit interpretability and applicability to complex models like MLLMs, which even in some cases are not open-access and therefore we cannot extracts their internal representations.

To overcome this, we introduce a new psychophysics-inspired methodology that enables the evaluation of contrast sensitivity in MLLMs without relying on internal representations as showed in figure 8.5. Instead of probing intermediate features, we treat the MLLM as a human observer and interact with it in natural language. Inspired by classical visual psychophysics, our method presents the model with a sequence of bandpass-filtered noise images at different spatial frequencies and contrast levels, and prompts it to detect the presence of a pattern using simple binary questions ("Is there a pattern

in the image?"). From the responses, we estimate psychometric functions for each frequency and compute the contrast thresholds, i.e. the needed contrast for a 50% detection probability. The inverse of these thresholds gives us the model's Contrast Sensitivity Function (CSF), following the same procedure used in human perceptual experiments.

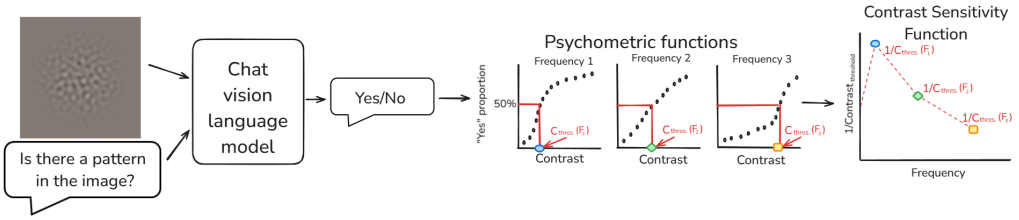


Figure 8.5: **Psychophysics-inspired methodology to evaluate CSF in MLLMs:** Models are shown noise-based grating stimuli across varying contrasts and spatial frequencies. Natural language prompts ask if a pattern is present. Binary "yes/no" answers allow us to build psychometric curves, from which we derive contrast thresholds and compute model CSFs, fully bypassing internal readout dependencies. Figure from [Hernandez-Camara et al., 2025a].

This methodology provides an end-to-end perceptual measure that bypasses assumptions about embedding distances, classification heads, or internal accessibility. Importantly, it does not commit to how contrast should be encoded within the model, offering a direct, interpretable view of how MLLMs respond to elementary visual properties.

We applied this approach to a diverse set of open-source MLLMs, including Qwen2.5VL-3B [Bai et al., 2025], InternVL2.5-4B [Chen et al., 2024], InternVL3-2B [Zhu et al., 2025], Qwen2.5VL-7B [Bai et al., 2025], LLaVA-1.5-7B [**?**], Magma-8B, [Yang et al., 2025], InternVL2.5-8B [Chen et al., 2024] and InternVL3-8B [Zhu et al., 2025], CLIP-b16-224 [Radford et al., 2021b] and SigLIP2-b16-224 [Tschannen et al., 2025]. Note that in this analysis, we do not plan to do a detailed analysis of all available models, but rather to present and prove the method with a selection of models that differ in architecture size, vision encoder backbones, and training objectives. The exhaustive analysis of the different variables that increase or reduce the alignment is beyond the scope of this work.

The resulting CSFs exhibit marked variability across models as shown in figure 8.6. Some, like Qwen2.5-VL and InstructBLIP, produce CSFs with the

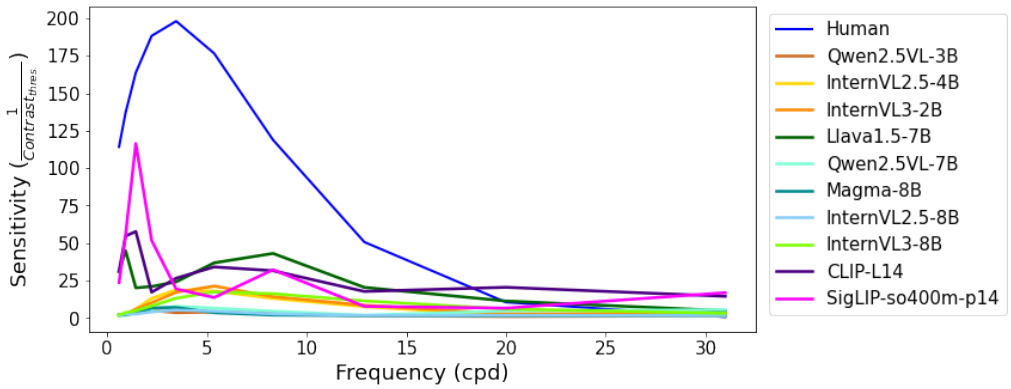


Figure 8.6: **Average contrast sensitivity functions (CSFs) for different MLLMs:** Each curve represents the average CSF across 25 prompt variants per spatial frequency (20 for the contrastive trained models). The human CSF (of the Standard Spatial Observer [Watson and Malo, 2002]) is included for reference as the dark blue line. Figure from [Hernandez-Camara et al., 2025a].

classic bandpass shape, peaking in mid-range spatial frequencies, closely resembling the human CSF [Campbell and Robson, 1968, Watson and Malo, 2002]. Others display flattened or low-pass profiles, suggesting diminished tuning to spatial frequency, a plausible outcome given that most of these models rely on vision backbones pretrained for classification, not spatial discrimination.

To quantify these comparisons, we compute both the Pearson correlation (ρ_P) [Galton, 1877] between each model's CSF and the human CSF to assess their shape alignment, and the root mean squared error (RMSE) as a measure of absolute sensitivity deviation. Table 8.3 summarizes the results. The results reveal important differences. Qwen2.5-VL exhibits the highest correlation with human CSF shape, peaking near the expected 3–6 cycles per degree. However, it underestimates contrast sensitivity, especially at low frequencies. Conversely, models like the contrastive CLIP and SigLIP or LLaVA-1.5 offer better alignment in absolute contrast sensitivity levels, but with flatter or more irregular profiles that deviate from human-like bandpass tuning. Others, such as Intern3, perform worse in both shape and sensitivity, displaying noisy or flat CSFs across the spatial frequency range.

These findings demonstrate that, while some MLLMs begin to exhibit human-like sensitivity to spatial frequencies, no model fully captures both the

Model	$\rho_{Pearson} \uparrow$	RMSE $\downarrow$
3B Scale		
Qwen 2.5VL (3B)	**0.86**	131.6
InternVL 2.5 (4B)	0.69	125.2
InternVL 3 (2B)	0.59	125.3
7B Scale		
Llava 1.5 (7B)	0.54	109.8
Qwen 2.5VL (7B)	0.31	130.6
Magma (8B)	0.76	131.2
InternVL 2.5 (8B)	0.84	131.6
InternVL 3 (8B)	0.37	126.4
Contrastive		
CLIP-b16-224	0.46	106.1
SigLIP-b16-224	0.45	**102.2**

Table 8.3: **Comparison between each model's average CSF and the human CSF:** Pearson correlation ($\rho_{Pearson}$) measures shape similarity; root mean squared error (RMSE) measures deviation in absolute sensitivity. These values highlight perceptual differences across models, with no expectation of human matching. Highest correlation and lowest error values are highlighted to illustrate models that most closely resemble human CSF shape and scale. Table from [Hernandez-Camara et al., 2025a].

shape and magnitude of the biological CSF. Moreover, because our method does not rely on internal metrics, these outcomes reflect the models' end-to-end capacity to "see" contrast through natural interaction, highlighting perceptual gaps that may otherwise be hidden by feature-based evaluations. In sum, this new proposed methodology offers a new avenue for probing low-level vision in multimodal systems. As MLLMs become central to human-facing Artificial Intelligence applications, understanding their visual processing fidelity, especially in foundational perceptual dimensions like contrast and frequency, is essential. Our results suggest that while modern MLLMs have made progress toward perceptual sensitivity, they still fall short of replicating basic human vision behaviour, particularly in tasks not explicitly targeted during training.

Part IV

Discussion and Conclusions

9

Discussion

This thesis had the dual goal of improving deep learning models through biologically inspired mechanisms and assessing their alignment with human visual perception. Through a combination of architectural bio-inspired improvements, perceptual benchmarking, and multimodal analysis, we discovered new insights into how modern artificial vision systems align, or fail to relate, with human vision. In this section, we reflect on the implications of our findings, the problems they reveal, and the broader questions they raise for the field of human-aligned artificial intelligence.

Bio-Inspired Computations Improve Performance, Not Alignment

A central outcome of this thesis is the systematic integration of Divisive Normalization (DN), a canonical computation of the early visual cortex [Carandini et al., 2005, Carandini and Heeger, 2012], into deep segmentation networks. Our results (see table 3.1 and robustness analyses in section 3.5) show that DN-augmented U-Nets outperform their baseline counterparts specially under challenging visual conditions, including fog, night, and global contrast shifts. These improvements highlight how biologically inspired mechanisms can provide artificial networks with greater invariance

to data variability, one of the distinctive features of human vision.

However, a deeper look at alignment reveals a striking limitation. Despite their robustness advantages, DN-augmented models did not exhibit significantly higher alignment with human perception in low-level psychophysical benchmarks. As demonstrated with our Decalogue framework, DN-augmented U-Nets fail to reproduce basic low-level human phenomena such as contrast sensitivity functions or context-dependent masking (see results in section 5.3). Interestingly, we observed partial alignment in color discrimination tasks, where models captured aspects of human-like MacAdam ellipses, though this is driven primarily by the training data rather than the architectural inclusion of DN (see chapter 6). Therefore, while bio-inspired components can boost robustness, perceptual alignment depends on a broader interplay of factors. This reveals a clear decoupling between biological plausibility (inspired by neuroscience) and perceptual alignment (inspired by behavioral data) highlighted in recent works [Hernández-Cámara et al., 2024a, Muttenthaler et al., 2024]. In other words, integrating computations known to exist in the human visual system does not automatically make a model behave perceptually like a human, and therefore, biological plausibility at Marr's computational level does not guarantee behavioral alignment at the algorithmic or representational level.

This decoupling raises an important point for the field: implementing computations known to exist in the brain can improve task performance and robustness without necessarily producing human-like perceptual behavior. In other words, neuroscience-inspired components can help networks become more resilient, but alignment with human judgments depends on broader interactions between architecture, training objectives, and data statistics. This observation both tempers strong claims of bio-inspiration as a shortcut to alignment and underscores the need for behavioral evaluation alongside architectural innovation.

Perceptual Alignment Is a Complex, Multi-Factorial Property

Our analyses across CNNs, ViTs, and CLIP-based models (see chapters 7 and first part of chapter 8) demonstrate that perceptual alignment is not determined by any single architectural or training choice. Instead, it emerges from the interaction of multiple factors: model architecture, training dataset and its statistics, optimization objective, regularization strategy, and

even the read-out method used to probe internal representations. For example, as shown in figure 7.4, Vision Transformers trained with larger datasets and stronger augmentation were less aligned with human perception, despite gains in accuracy. Similarly, in CLIP, alignment varied depending not only on the vision backbone size but also on the domain of the training data (figure 8.3).

Color perception provides a particularly clear illustration of this interplay. Our analyses of MacAdam ellipses and color discrimination thresholds (chapter 6) showed that alignment with human perception did not stem from architectural choices like DN, but from the statistics of the training data. Models trained on datasets with richer and more uniform chromatic distributions exhibited discrimination regions closer to human ellipses, whereas others displayed anisotropic or distorted sensitivity profiles. This example underscores how perceptual alignment often emerges from the interaction between environmental statistics and representational learning, rather than from architectural bio-inspiration alone. In Marr's terms [Marr and Poggio, 1976], the task-level constraints (dataset and statistics) shaped alignment more decisively than the implementation-level details (architecture).

This multi-factorial dependency challenges the assumption that perceptual alignment can be studied in isolation. In practice, the relevant influences span across Marr's three levels of analysis [Marr and Poggio, 1976]: implementation (architecture), algorithm (training objective), and task (dataset, labels). Our commentary work [Malo and Hernandez-Camara, 2024] and related discussions [Poggio, 2021] emphasize that these levels, while conceptually separable, are deeply entangled in practice. For instance, changing the dataset not only modifies the input statistics but also alters optimization dynamics and representational hierarchies, complicating causal attributions.

Too Much Optimization Hurts Alignment

A particularly striking insight is the non-linear relationship between model optimization and human alignment, which we summarize as an "inverted-U" pattern (figure 7.5). While initial improvements in training and scaling often increase human alignment, beyond a certain point additional optimization, whether through larger architectures, prolonged training, or stronger regularization, tends to reduce it.

This phenomenon was evident in both CNN-based and Transformer-

based models. For example, in Vision Transformers, longer training (i.e., more exposures per image) reduced correlation with human judgments from TID-2013, indicating overfitting away from human-like representations (figure 7.4). Similarly, in CLIP, longer contrastive optimization improved zero-shot accuracy but decreased alignment with human perceptual similarity (figure 8.4). These results resonate with other reports in the literature: networks that excel at their supervised objectives may exploit shortcuts or non-robust cues, diverging from human perceptual strategies [Geirhos et al., 2018a, Geirhos et al., 2020, Kumar et al., 2022].

Interestingly, the most human-like perceptual behavior often appeared in simpler models or earlier representational stages: AlexNet and VGG consistently showed higher correlations with human IQA judgments than more advanced ResNets or ConvNeXts (figure 7.2), and early layers of CLIP aligned more closely with human texture sensitivity before training pushed the model toward shape abstraction (8.2). This points to a fundamental trade-off: optimizing networks for accuracy and robustness may come at the cost of perceptual fidelity. This opens a challenge: whether future models can jointly achieve high task performance and strong human alignment, or whether these goals are inherently in conflict.

Low-Level Perception Is Not Emergent in Multimodal Models

While contrastively trained models such as CLIP already depart from conventional vision-only paradigms, modern multimodal large language models (MLLMs) extend this approach by integrating vision and language into generative, conversational frameworks. These systems are increasingly deployed as general-purpose "AI agents," raising the expectation that they can perceive and reason in ways analogous to humans.

However, our analysis of the contrast sensitivity function (CSF) in MLLMs (section 8.4) demonstrates that these models do not reproduce the human profile of sensitivity to spatial frequency and contrast. Although some, such as Qwen2.5-VL, approximate the shape of the human CSF qualitatively, none matched human thresholds quantitatively. This gap indicates that high-level generative competence and task generalization do not guarantee low-level perceptual alignment. In other words, conversational fluency and semantic reasoning in MLLMs are not supported by early visual mechanisms equivalent to those of the human visual system. Thus, low-level perceptual fidelity

should not be assumed as an emergent property of multimodal training; it must be explicitly engineered or evaluated.

The Role of Language in Shaping Representations

Our results also highlight the powerful inductive role of language supervision in multimodal representation learning. As shown in our analysis of CLIP training (section 8.3), visual features evolve from a texture-biased regime early in training, closer to human low-level judgments in IQA benchmarks, to increasingly shape-dominated representations aligned with semantic categorization. This observation is consistent with recent reports that multimodal contrastive training induces stronger shape biases than are typically found in CNNs or unimodal ViTs [Gavrikov et al., 2024, Geirhos et al., 2018a].

From a functional perspective, this language-driven abstraction improves robustness and semantic alignment: CLIP models are less fooled by texture-shape cue conflicts. However, this comes at a cost of low-level human alignment. The shift away from texture cues corresponds to a reduction in low-level perceptual alignment, as evidenced by weaker TID-2013 correlations (figure 8.4). Language supervision thus acts as a representational compressor, prioritizing global structure over fine detail, a trade-off that benefits semantic reasoning but undermines fidelity to human perceptual sensitivity.

Taken together, these results suggest that while language enhances semantic-level alignment with human categorization, it simultaneously drives a departure from the low-level coding strategies that characterize biological vision. This opens a new challenge ahead: can these two dimensions be achieved at the same time? Can multimodal models be both semantically and low-level perceptually human-like?

Layer Depth as a Window into Alignment

Across all experiments, we consistently observed that alignment with human perception varies substantially across layer depth. Early and intermediate layers of CNNs and CLIP models tend to correlate more strongly with human judgments, particularly in tasks such as image quality assessment (figures 7.2 and 8.2). In contrast, final task-specific layers often diverge from low-level human-like behavior, presumably due to over-specialization to the training objectives.

This observation aligns with previous findings in neuroscience

and computational modeling showing that early cortical areas encode lower-level perceptual features, while higher areas increasingly abstract toward task-relevant semantics [Yamins and DiCarlo, 2016, Khaligh-Razavi and Kriegeskorte, 2014]. In artificial systems, this suggests that perceptual alignment may reside in intermediate computations, even if the model's final output departs from low-level human behavior.

For researchers aiming to design interpretable or hybrid models, this highlights the importance of layer-wise evaluations. Rather than treating a model as a monolithic end-to-end predictor, analyzing alignment across depth allows us to identify which stages best approximate human-like processing. These insights could guide architectures where human-aligned components are deliberately incorporated at specific stages of computation.

Final Reflections and Theoretical Implications

Taken together, the findings of this thesis paint a complex picture of human alignment in artificial vision. Deep models can, under certain conditions, exhibit perceptual behaviors that parallel human judgments, but this alignment is context-dependent, and not a direct byproduct of performance optimization. Indeed, we showed that stronger optimization often reduces human alignment, creating a tension between engineering goals (accuracy, generalization, robustness) and perceptual correspondence with humans.

This raises a broader theoretical implication: human alignment is not an emergent guarantee of scale or training data. Instead, it must be explicitly considered in design, evaluation, and interpretation. Importantly, this thesis does not argue that human alignment should not be seen as universally necessary; there are domains where surpassing human-level is beneficial, such as super-resolution or medical imaging. However, in many other contexts such as compression, quality assessment, explainability, or safety-critical applications, human-aligned perception is indispensable.

Ultimately, achieving such alignment requires more than simply adding bio-inspired modules or training on larger datasets. It demands a closer integration of perceptual theory, behavioral evaluation, and architectural design principles. By identifying what drives models closer to or further away from human perception, this thesis contributes to both the practical design of more robust models and to the conceptual understanding of what it means for an artificial system to "see" like a human.

10

Conclusions and Future Work

10.1 Conclusions and Contributions

This thesis set out to bridge the gap between artificial and biological vision by pursuing three interconnected objectives: (1) to integrate biologically inspired computations into deep learning architectures, (2) to develop and apply methodologies for measuring human alignment, and (3) to analyze the factors that most strongly shape alignment in modern vision systems. Across multiple studies, we showed that progress on each of these fronts offer insights not only for improving artificial models but also for clarifying the computational principles that distinguish them from human perception.

Objective 1: Biologically Inspired Computations in Deep Learning Architectures

The first objective focused on testing whether canonical computations of early vision could improve deep models. We integrated Divisive Normalization into segmentation networks such as U-Net and showed that DN modules consistently improved performance under challenging visual conditions, including fog, low illumination, and altered contrast. These improvements

came with minimal parameter increase and offer high robustness gains, highlighting how biologically inspired mechanisms can enhance artificial networks.

However, improved robustness did not translate into perceptual alignment. DN-augmented models failed to reproduce most low-level psychophysical phenomena in the Decalogue evaluation, such as contrast sensitivity or contextual masking. This decoupling shows that biological plausibility at the architectural level does not automatically imply human-like perceptual behavior.

Objective 2: Methodologies for Evaluating Human Alignment

The second objective was to develop systematic tools for measuring perceptual alignment, moving beyond accuracy-driven benchmarks. To this end, the thesis introduced several methodological contributions:

- The Decalogue framework, a battery of psychophysics-inspired tests probing low-level perceptual phenomena in deep networks.
- A color discrimination paradigm based on MacAdam ellipses, which enabled a direct comparison of perceptual thresholds in models versus humans. This methodology revealed not only anisotropies in model color spaces but also provided a rigorous tool to test how architecture and training data shape perceptual thresholds.
- A contrast sensitivity function (CSF) evaluation for multimodal large language models (MLLMs), enabling the study of their visual sensitivity without reliance on internal read-out strategies.
- An abstraction-layer evaluation of CLIP, disentangling low-, mid-, and high-level alignment tasks to reveal where alignment emerges or decays during training.

Together, these tools enabled a more fine-grained, multi-level assessment of alignment, covering both behavioral similarity and representational correspondence across architectures, tasks, and modalities.

Objective 3: Factors Shaping Alignment in Deep Models

The third objective examined which design and training factors influence alignment. Across CNNs, ViTs, CLIP, and MLLMs, our results show that human alignment is a complex, multi-factorial property. It depends not only on architecture but also on dataset statistics, objective function, training duration, regularization, and the depth at which representations are read out.

Several key findings emerged from the works presented in this thesis:

- Simpler models and early layers often align better with human perception than deeper or more optimized networks.
- Alignment and performance are not monotonically related: in many cases, increasing optimization reduces perceptual similarity, producing an inverted-U relationship between accuracy and alignment.
- Color perception offered one of the clearest demonstrations of task-level dominance over architecture. Our MacAdam ellipse analyses showed that models trained on datasets with richer and more uniform chromatic distributions exhibited discrimination regions closer to human ellipses, while those with skewed statistics showed distorted sensitivity profiles. This confirms that training environment and input statistics can shape alignment more decisively than architectural design.
- Language supervision shifts representations toward semantic, shape-based encodings, improving high-level alignment with human categorization but weakening low-level alignment to perceptual phenomena such as contrast sensitivity.
- MLLMs, despite their impressive generalization, fail to reproduce core perceptual sensitivities such as the human CSF, underscoring the gap between high-level multimodal reasoning and basic perceptual fidelity.

Final Reflections

By addressing these three objectives, this thesis makes both practical contributions, new architectures and novel evaluation tools, and conceptual contributions, clarifying when and why deep networks resemble human vision. The findings highlight that human alignment is not an automatic byproduct

of scale or performance optimization but requires deliberate integration of perceptual theory, evaluation frameworks, and architectural design.

Ultimately, the work underscores a broader insight: to build models that are robust, interpretable, and human-aligned, we must look beyond benchmarks and embrace interdisciplinary strategies that bridge computer vision, computational neuroscience and vision science.

10.2 Future Research Directions

The work presented here opens several avenues for further exploration:

1. Bridging performance and alignment through targeted training objectives:

While we observed that training goals largely determine alignment, future work could explore hybrid objectives that explicitly optimize for both task performance and perceptual alignment. This may involve adding perceptual loss terms based on human behavior, or using human judgments to guide contrastive or reconstruction-based pertaining, using other known techniques for human alignment, such as Direct Preference Optimization (DPO) [Rafailov et al., 2023].

2. Scaling bio-inspired architectures across modalities:

Our integration of Divisive Normalization into CNNs demonstrated that biologically inspired modules can meaningfully improve robustness, but their impact in vision transformers and in multimodal models remains unexplored. A natural next step is to test whether DN, or related canonical computations, can enhance robustness and alignment in Vision Transformers, CLIP-like models, and multimodal large language models (MLLMs). Such work would connect architectural design more directly to the realities of current AI practice, where ViTs and multimodal systems are the dominant backbones.

3. Developing readout-independent alignment metrics:

Another contribution of this thesis was to show that alignment depends not only on architecture and data, but also on how one chooses to read out representations. Our CSF framework for MLLMs demonstrated the value of behavioral, readout-free evaluations. Extending this line of work, particularly through psychophysics-inspired paradigms, could yield cleaner and more

interpretable benchmarks for alignment, especially in complex or black-box systems [Dulay et al., 2024].

4. Alignment beyond early vision:

Most of our analyses focused on low-level perceptual properties such as contrast sensitivity, noise robustness, or color discrimination. However, many of the computational questions raised here extend to mid- and high-level phenomena, including attention, perceptual grouping, or object organization. These domains are better explored in current deep learning research but are equally important for bridging artificial and biological vision, and for building models that interact with humans in cognitively meaningful ways.

5. Alignment-aware model evaluation standards:

Finally, this thesis supports the idea that task accuracy should not be the unique metric of model success. Future evaluation standards, especially for models intended for human-facing tasks, should include alignment measures that reflect how well the model captures human-like perception or decision boundaries.

In summary, this thesis provides both empirical evidence and methodological tools to better understand the connection between deep learning and human vision. It shows that while bio-inspired components can improve machine performance and robustness, human alignment emerges from different factors, most notably, training objective and data statistics. By disentangling these factors, we move one step closer to building models that are accurate and aligned with how humans perceive the world.

Part V

Bibliography

Bibliography

[IPL,] Ipl-carla-dataset. `https://huggingface.co/datasets/isp-uv-es/IPL-CARLA-dataset`. Accessed: 2025-07-08.

[Agarap, 2018] Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

[Akbarinia, 2025] Akbarinia, A. (2025). Exploring the categorical nature of colour perception: Insights from artificial networks. *Neural Networks*, 181:106758.

[Akbarinia et al., 2023] Akbarinia, A., Morgenstern, Y., and Gegenfurtner, K. R. (2023). Contrast sensitivity function in deep networks. *Neural Networks*, 164:228–244.

[Azad et al., 2024] Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J. P., Adeli, E., and Merhof, D. (2024). Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Bai et al., 2025] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. (2025). Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

[Ballé et al., 2015] Ballé, J., Laparra, V., and Simoncelli, E. P. (2015). Density modeling of images using a generalized normalization transformation. *arXiv preprint arXiv:1511.06281*.

[Ballé et al., 2016] Ballé, J., Laparra, V., and Simoncelli, E. P. (2016). End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*.

[Barlow et al., 1961] Barlow, H. B. et al. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01):217–233.

[Beery et al., 2018] Beery, S., Van Horn, G., and Perona, P. (2018). Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473.

[Bowen et al., 2022] Bowen, E. F., Rodriguez, A. M., Sowinski, D. R., and Granger, R. (2022). Visual stream connectivity predicts assessments of image quality. *Journal of vision*, 22(11):4–4.

[Burg et al., 2021] Burg, M. F., Cadena, S. A., Denfield, G. H., Walker, E. Y., Tolias, A. S., Bethge, M., and Ecker, A. S. (2021). Learning divisive normalization in primary visual cortex. *PLoS computational biology*, 17(6):e1009028.

[Cai et al., 2025] Cai, Y., Yin, F., Hammou, D., and Mantiuk, R. (2025). Do computer vision foundation models learn the low-level characteristics of the human visual system? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20039–20048.

[Campbell and Robson, 1968] Campbell, F. W. and Robson, J. G. (1968). Application of fourier analysis to the visibility of gratings. *The Journal of physiology*, 197(3):551.

[Carandini et al., 2005] Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., and Rust, N. C. (2005). Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–10597.

[Carandini and Heeger, 1994] Carandini, M. and Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science*, 264(5163):1333–1336.

[Carandini and Heeger, 2012] Carandini, M. and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature reviews neuroscience*, 13(1):51–62.

[Caron et al., 2021] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

[Cauchy et al., 1847] Cauchy, A. et al. (1847). Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538.

[Cekic et al., 2022] Cekic, M., Bakiskan, C., and Madhow, U. (2022). Neuro-inspired deep neural networks with sparse, strong activations. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3843–3847. IEEE.

[Chen et al., 2024] Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al. (2024). Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

[Cherti et al., 2023] Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. (2023). Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829.

[Coen-Cagli and Schwartz, 2013] Coen-Cagli, R. and Schwartz, O. (2013). The impact on midlevel vision of statistically optimal divisive normalization in v1. *Journal of vision*, 13(8):13–13.

[Cordts et al., 2016] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223.

[Dai and Van Gool, 2018] Dai, D. and Van Gool, L. (2018). Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3819–3824. IEEE.

[Daly, 1990] Daly, S. J. (1990). Application of a noise-adaptive contrast sensitivity function to image data compression. *Optical Engineering*, 29(8):977–987.

[Dapello et al., 2022] Dapello, J., Kar, K., Schrimpf, M., Geary, R., Ferguson, M., Cox, D. D., and DiCarlo, J. J. (2022). Aligning model and macaque inferior temporal cortex representations improves model-to-human behavioral alignment and adversarial robustness. *BioRxiv*, pages 2022–07.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

[Díez-Ajenjo et al., 2011] Díez-Ajenjo, M. A., Capilla, P., and Luque, M. (2011). Red-green vs. blue-yellow spatio-temporal contrast sensitivity across the visual field. *Journal of Modern Optics*, 58(19-20):1736–1748.

[Ding et al., 2020] Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. (2020). Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581.

[Dodge and Karam, 2017] Dodge, S. and Karam, L. (2017). A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, pages 1–7. IEEE.

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[Dosovitskiy et al., 2017] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. (2017). Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR.

[Dulay et al., 2024] Dulay, J., Poltoratski, S., Hartmann, T. S., Anthony, S. E., and Scheirer, W. J. (2024). Informing machine perception with psychophysics. *Proceedings of the IEEE*, 112(2):88–96.

[Epifanio et al., 2003] Epifanio, I., Gutierrez, J., and Malo, J. (2003). Linear transform for simultaneous diagonalization of covariance and perceptual metric matrix in image coding. *Pattern Recognition*, 36(8):1799–1811.

[Fu et al., 2023] Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., and Isola, P. (2023). Dreamsim: Learning new dimensions of human visual similarity using synthetic data.

[Fukushima, 1980] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202.

[Galton, 1877] Galton, F. (1877). Typical laws of heredity. *Nature*, 15(389):512–514.

[Gavrikov et al., 2024] Gavrikov, P., Lukasik, J., Jung, S., Geirhos, R., Mirza, M. J., Keuper, M., and Keuper, J. (2024). Can we talk models into seeing the world differently? *arXiv preprint arXiv:2403.09193*.

[Geirhos et al., 2020] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

[Geirhos et al., 2018a] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018a). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*.

[Geirhos et al., 2018b] Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. (2018b). Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31.

[Georgeson and Sullivan, 1975] Georgeson, M. and Sullivan, G. (1975). Contrast constancy: deblurring in human vision by spatial frequency channels. *The Journal of physiology*, 252(3):627–656.

[Ghiasi et al., 2022] Ghiasi, A., Kazemi, H., Borgnia, E., Reich, S., Shu, M., Goldblum, M., Wilson, A. G., and Goldstein, T. (2022). What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727*.

[Gibson, 2014] Gibson, J. J. (2014). *The ecological approach to visual perception: classic edition*. Psychology press.

[Giraldo and Schwartz, 2019] Giraldo, L. G. S. and Schwartz, O. (2019). Integrating flexible normalization into midlevel representations of deep convolutional neural networks. *Neural computation*, 31(11):2138–2176.

[Gomez-Villa et al., 2020] Gomez-Villa, A., Martín, A., Vazquez-Corral, J., Bertalmío, M., and Malo, J. (2020). Color illusions also deceive cnns for low-level vision tasks: Analysis and implications. *Vision Research*, 176:156–174.

[Gutierrez et al., 2005] Gutierrez, J., Ferri, F. J., and Malo, J. (2005). Regularization operators for natural images based on nonlinear perception models. *IEEE Transactions on Image Processing*, 15(1):189–200.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Hebart et al., 2020] Hebart, M. N., Zheng, C. Y., Pereira, F., and Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11):1173–1185.

[Heeger, 1992] Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual neuroscience*, 9(2):181–197.

[Hendrycks and Dietterich, 2019] Hendrycks, D. and Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

[Hepburn et al., 2020] Hepburn, A., Laparra, V., Malo, J., McConville, R., and Santos-Rodriguez, R. (2020). Perceptnet: A human visual system inspired neural network for estimating perceptual distance. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 121–125. IEEE.

[Hepburn et al., 2022] Hepburn, A., Laparra, V., Santos-Rodriguez, R., Ballé, J., and Malo, J. (2022). On the relation between statistical learning and perceptual distances. *ICLR*.

[Hernández-Cámara et al., 2024a] Hernández-Cámara, P., Daudén-Oliver, P., Laparra, V., and Malo, J. (2024a). Alignment of color discrimination in humans and image segmentation networks. *Frontiers in Psychology*, 15:1415958.

[Hernandez-Camara et al., 2025a] Hernandez-Camara, P., Gomez-Villa, A., Jaen-Lorites, J. M., Vila-Tomas, J., Malo, J., and Laparra, V. (2025a). Contrast sensitivity function of multimodal vision-language models. In *8th Annual Conference on Cognitive Computational Neuroscience*.

[Hernandez-Camara et al., 2025b] Hernandez-Camara, P., Jaen-Lorites, J. M., Vila-Tomas, J., Laparra, V., and Malo, J. (2025b). Do vision transformers see like humans? evaluating their perceptual alignment. In *8th Annual Conference on Cognitive Computational Neuroscience*.

[Hernandez-Camara et al., 2025c] Hernandez-Camara, P., Jaen-Lorites, J. M., Vila-Tomas, J., Malo, J., and Laparra, V. (2025c). Evolution of low-level and texture human-clip alignment. In *8th Annual Conference on Cognitive Computational Neuroscience*.

[Hernandez-Camara et al., 2022] Hernandez-Camara, P., Vila, J., Li, Q., Laparra, V., and Malo, J. (2022). Basic psychophysics of deep networks trained for image segmentation. In *9th Iberian Conference on Perception*.

[Hernández-Cámara et al., 2025] Hernández-Cámara, P., Vila-Tomás, J., Dauden-Oliver, P., Alabau-Bosque, N., Laparra, V., and Malo, J. (2025). Why divisive normalization works in image segmentation? *Neurocomputing*, page 130569.

[Hernández-Cámara et al., 2023] Hernández-Cámara, P., Vila-Tomás, J., Laparra, V., and Malo, J. (2023). Neural networks with divisive normalization for image segmentation. *Pattern Recognition Letters*, 173:64–71.

[Hernandez-Camara et al., 2025] Hernandez-Camara, P., Vila-Tomas, J., Laparra, V., and Malo, J. (2025). Dissecting the effectiveness of deep features as metric of perceptual image quality. *Neural Networks*, 185:107189.

[Hernández-Cámara et al., 2024b] Hernández-Cámara, P., Vila-Tomás, J., Malo, J., and Laparra, V. (2024b). Measuring human-clip alignment at different abstraction levels. In *ICLR 2024 Workshop on Representational Alignment*.

[Hinton et al., 2012] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

[Hodges Jr, 1958] Hodges Jr, J. (1958). The significance probability of the smirnov two-sample test. *Arkiv för matematik*, 3(5):469–486.

[Huang et al., 2017] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

[Hubel and Wiesel, 1962] Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106.

[Hubel et al., 1959] Hubel, D. H., Wiesel, T. N., et al. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591.

[Hurvich and Jameson, 1957] Hurvich, L. M. and Jameson, D. (1957). An opponent-process theory of color vision. *Psychological review*, 64(6p1):384.

[Islam et al., 2021] Islam, K., Dang, L. M., Lee, S., and Moon, H. (2021). Image compression with recurrent neural network and generalized divisive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1875–1879.

[Islam et al., 2020] Islam, M. J., Edge, C., Xiao, Y., Luo, P., Mehtaz, M., Morse, C., Enan, S. S., and Sattar, J. (2020). Semantic segmentation of underwater imagery: Dataset and benchmark. In *2020 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 1769–1776. IEEE.

[Khaligh-Razavi and Kriegeskorte, 2014] Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915.

[Kriegeskorte, 2015] Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1(1):417–446.

[Krizhevsky et al., 2009] Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25.

[Kumar et al., 2022] Kumar, M., Houlsby, N., Kalchbrenner, N., and Cubuk, E. D. (2022). Do better imagenet classifiers assess perceptual similarity better? *arXiv preprint arXiv:2203.04946*.

[Kurakin et al., 2018] Kurakin, A., Goodfellow, I. J., and Bengio, S. (2018). Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC.

[Landau et al., 1988] Landau, B., Smith, L. B., and Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321.

[Laparra et al., 2010] Laparra, V., Muñoz-Marí, J., and Malo, J. (2010). Divisive normalization image quality metric revisited. *Journal of the optical society of America A*, 27(4):852–864.

[Lazebnik et al., 2006] Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 2169–2178. IEEE.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551.

[Li et al., 2022] Li, Q., Gomez-Villa, A., Bertalmío, M., and Malo, J. (2022). Contrast sensitivity functions in autoencoders. *Journal of Vision*, 22(6):8–8.

[Lin et al., 2019] Lin, H., Hosu, V., and Saupe, D. (2019). Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE.

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

[Liu et al., 2023] Liu, Z., Gan, E., and Tegmark, M. (2023). Seeing is believing: Brain-inspired modular training for mechanistic interpretability. *Entropy*, 26(1):41.

[Liu et al., 2022] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986.

[Lowe, 2004] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

[Ma et al., 2017] Ma, K., Liu, W., Zhang, K., Duanmu, Z., Wang, Z., and Zuo, W. (2017). End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213.

[MacAdam, 1942] MacAdam, D. L. (1942). Visual sensitivities to color differences in daylight. *Journal of the Optical Society of America*, 32(5):247–274.

[Malo et al., 2005] Malo, J., Epifanio, I., Navarro, R., and Simoncelli, E. P. (2005). Nonlinear image representation for efficient perceptual coding. *IEEE Transactions on Image Processing*, 15(1):68–80.

[Malo and Gutiérrez, 2006] Malo, J. and Gutiérrez, J. (2006). V1 non-linear properties emerge from local-to-global non-linear ica. *Network: Computation in Neural Systems*, 17(1):85–102.

[Malo and Hernandez-Camara, 2024] Malo, J. and Hernandez-Camara, P. (2024). A separate theory-on-top level may be inspiring, but it is neither separate nor enough. *Journal of Physiology*, 609(9):1919.

[Malo et al., 2022] Malo, J., Hernandez-Camara, P., Vila, J., Li, Q., and Laparra, V. (2022). A visual psychophysics decalogue to check the human nature of artificial networks. In *9th Iberian Conference on Perception*.

[Malo and Laparra, 2010] Malo, J. and Laparra, V. (2010). Psychophysically tuned divisive normalization approximately factorizes the pdf of natural images. *Neural computation*, 22(12):3179–3206.

[Marr, 2010] Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

[Marr and Poggio, 1976] Marr, D. and Poggio, T. (1976). From understanding computation to understanding neural circuitry. Technical report.

[Martinez et al., 2019] Martinez, M., Bertalmío, M., and Malo, J. (2019). In praise of artifice reloaded: caution with natural image databases in modeling vision. front neurosci. 2019.

[Martinez-Garcia et al., 2018] Martinez-Garcia, M., Cyriac, P., Batard, T., Bertalmío, M., and Malo, J. (2018). Derivatives and inverse of cascaded linear+ nonlinear neural models. *PloS one*, 13(10):e0201326.

[McCulloch and Pitts, 1943] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

[Miller et al., 2021] Miller, M., Chung, S., and Miller, K. D. (2021). Divisive feature normalization improves image recognition performance in alexnet. In *International Conference on Learning Representations*.

[Mullen, 1985] Mullen, K. T. (1985). The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *The Journal of physiology*, 359(1):381–400.

[Muttenthaler et al., 2022] Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A., and Kornblith, S. (2022). Human alignment of neural network representations. *arXiv preprint arXiv:2211.01201*.

[Muttenthaler et al., 2024] Muttenthaler, L., Greff, K., Born, F., Spitzer, B., Kornblith, S., Mozer, M. C., MÃžller, K.-R., Unterthiner, T., and Lampinen, A. K. (2024). Aligning machine and human visual representations across abstraction levels. *arXiv preprint arXiv:2409.06509*.

[Nair and Hinton, 2010] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

[Pan et al., 2021] Pan, X., Sánchez Giraldo, L. G., Kartal, E., and Schwartz, O. (2021). Brain-inspired weighted normalization for cnn image classification. *bioRxiv*.

[Parkhi et al., 2012] Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. (2012). Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE.

[Peterson et al., 2018] Peterson, J. C., Abbott, J. T., and Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, 42(8):2648–2669.

[Poggio, 2021] Poggio, T. (2021). From marr's vision to the problem of human intelligence. Technical report, Center for Brains, Minds and Machines (CBMM).

[Ponomarenko et al., 2015] Ponomarenko, N., Jin, L., Ieremeiev, O., Lukin, V., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., et al. (2015). Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77.

[Rabinowitz et al., 2011] Rabinowitz, N. C., Willmore, B. D., Schnupp, J. W., and King, A. J. (2011). Contrast gain control in auditory cortex. *Neuron*, 70(6):1178–1191.

[Radford et al., 2021a] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021a). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

[Radford et al., 2021b] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021b). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

[Rafailov et al., 2023] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

[Real and Vargas, 1996] Real, R. and Vargas, J. M. (1996). The probabilistic basis of jaccard's index of similarity. *Systematic biology*, 45(3):380–385.

[Ren et al., 2016] Ren, M., Liao, R., Urtasun, R., Sinz, F. H., and Zemel, R. S. (2016). Normalizing the normalizers: Comparing and extending network normalization schemes. *arXiv preprint arXiv:1611.04520*.

[Richards et al., 2019] Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770.

[Richter et al., 2016] Richter, S. R., Vineet, V., Roth, S., and Koltun, V. (2016). Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer.

[Roads and Love, 2021] Roads, B. D. and Love, B. C. (2021). Enriching imagenet with human similarity judgments and psychological embeddings. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 3547–3557.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

[Rust and Movshon, 2005] Rust, N. C. and Movshon, J. A. (2005). In praise of artifice. *Nature neuroscience*, 8(12):1647–1650.

[Sakaridis et al., 2018] Sakaridis, C., Dai, D., and Van Gool, L. (2018). Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992.

[Schwartz and Simoncelli, 2001] Schwartz, O. and Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature neuroscience*, 4(8):819–825.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[Spearman, 1987] Spearman, C. (1987). The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471.

[Steiner et al., 2021] Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. (2021). How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*.

[Sucholutsky and Griffiths, 2023] Sucholutsky, I. and Griffiths, T. (2023). Alignment with human representations supports robust few-shot learning. *Advances in Neural Information Processing Systems*, 36:73464–73479.

[Sucholutsky et al., 2023] Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., Grant, E., Groen, I., Achterberg, J., et al. (2023). Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

[Szegedy et al., 2013] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

[Tschannen et al., 2025] Tschannen, M., Gritsenko, A., Wang, X., Naeem, M. F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al. (2025). Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Vila et al., 2022] Vila, J., Hernandez-Camara, P., Li, Q., Laparra, V., and Malo, J. (2022). Basic psychophysics of deep networks trained for subjective image distortion. In *9th Iberian Conference on Perception*.

[Vila-Tomás et al., 2024] Vila-Tomás, J., Hernández-Cámara, P., Laparra, V., and Malo, J. (2024). Parametric enhancement of perceptnet: A human-inspired approach for image quality assessment. *arXiv preprint arXiv:2412.03210*.

[Vila-Tomás et al., 2025] Vila-Tomás, J., Hernández-Cámara, P., Li, Q., Laparra, V., and Malo, J. (2025). A turing test for artificial nets devoted to model human vision. *arXiv preprint arXiv:2502.00721*.

[Wang et al., 2004] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

[Watson and Malo, 2002] Watson, A. B. and Malo, J. (2002). Video quality measures based on the standard spatial observer. In *Proceedings. International Conference on Image Processing*, volume 3, pages III–III. IEEE.

[Watson and Solomon, 1997] Watson, A. B. and Solomon, J. A. (1997). Model of visual contrast gain control and pattern masking. *Journal of the optical society of America A*, 14(9):2379–2391.

[Werbos, 1974] Werbos, P. J. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. PhD thesis, Harvard University.

[Wichmann and Geirhos, 2023] Wichmann, F. A. and Geirhos, R. (2023). Are deep neural networks adequate behavioral models of human visual perception? *Annual review of vision science*, 9(1):501–524.

[Wyszecki and Stiles, 2000] Wyszecki, G. and Stiles, W. S. (2000). *Color science: concepts and methods, quantitative data and formulae*. John wiley & sons.

[Xu and Vaziri-Pashkam, 2021] Xu, Y. and Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature communications*, 12(1):2065.

[Yamins and DiCarlo, 2016] Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365.

[Yang et al., 2022] Yang, A., Pan, J., Lin, J., Men, R., Zhang, Y., Zhou, J., and Zhou, C. (2022). Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

[Yang et al., 2025] Yang, J., Tan, R., Wu, Q., Zheng, R., Peng, B., Liang, Y., Gu, Y., Cai, M., Ye, S., Jang, J., et al. (2025). Magma: A foundation model for multimodal ai agents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14203–14214.

[Zamir et al., 2018] Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., and Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722.

[Zeiler and Fergus, 2014] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

[Zhai et al., 2022] Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. (2022). Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113.

[Zhai et al., 2023] Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

[Zhang et al., 2018] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

[Zhang et al., 2023] Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al. (2023). Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2(3):6.

[Zhou et al., 2014] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27.

[Zhu et al., 2025] Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al. (2025). Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Part VI

Scientific Publications

Publications

The achievements and conclusions of this work have been published in the following papers in high-quality international journals. Below is a list of the journal publications, ordered according to their appearance in the thesis:

- Neural Networks with Divisive Normalization for Image Segmentation. Pablo Hernández-Cámara, Jorge Vila-Tomás, Valero Laparra, Jesús Malo. Pattern Recognition Letters, 173, 64-71, 2023.
  DOI: `https://doi.org/10.1016/j.patrec.2023.07.017`
  Described in chapter 3 and chapter 4.

- Why Divisive Normalization Works in Image Segmentation? Pablo Hernández-Cámara, Jorge Vila-Tomás, Paula Daudén-Oliver, Nuria Alabau-Bosque, Valero Laparra, Jesús Malo. Neurocomputing, 649, 130569, 2025.
  DOI: `https://doi.org/10.1016/j.neucom.2025.130569`
  Described in chapter 3 and chapter 4.

- Alignment of Color Discrimination in Humans and Image Segmentation Networks. Pablo Hernández-Cámara, Paula Daudén-Oliver, Jesús Malo. Frontiers in Psychology, 15, 415958, 2024.
  DOI: `https://doi.org/10.3389/fpsyg.2024.1415958`
  Described in chapter 6.

- Dissecting the Effectiveness of Deep Features as Metric of Perceptual Image Quality. Pablo Hernández-Cámara, Jorge Vila-Tomás, Valero Laparra, Jesús Malo. Neural Networks, 185, 107189, 2025.
  DOI: `https://doi.org/10.1016/j.neunet.2025.107189`
  Described in chapter 7.

In addition, other relevant contributions and intermediate results of this thesis have been presented at international conferences. Below is a list of

those conference publications, also ordered according to their appearance in the thesis.

- Basic Psychophysics of Deep Networks Trained for Image Segmentation. Pablo Hernández-Cámara, Jorge Vila-Tomás, Qiang Li, Valero Laparra, Jesus Malo. IX Iberian Conference on Perception. 2022. Oral presentation. Barcelona (Spain).
  Accessible in: `https://isp.uv.es/docs/Hernandez_et_al_CIP_22.pdf`
  Described in chapter 5.

- Measuring Human-CLIP Alignment at Different Abstraction Levels. Pablo Hernández-Cámara, Jorge Vila-Tomás, Jesus Malo, Valero Laparra. ICLR Workshop on Representational Alignment. 2024. Poster presentation. Vienna (Austria).
  Accessible in: `https://openreview.net/pdf?id=xQyhHjLGmj`
  Described in chapter 8.

- Correspondence Between Humans and CLIP at Different Abstraction Levels. Pablo Hernández-Cámara, Jorge Vila-Tomás, Jesus Malo, Valero Laparra. Fifth International Convention on the Mathematics Of Neuroscience and AI. 2024. Poster presentation. Rome (Italy).
  Accessible in: `https://isp.uv.es/docs/Neuromonster_24.pdf`
  Described in chapter 8.

- Do Vision Transformers See Like Humans? Evaluating their Perceptual Alignment. Pablo Hernández-Cámara, Jose Manuel Jaén-Lorites, Jorge Vila-Tomás, Valero Laparra, Jesus Malo. 8th Annual Conference on Cognitive Computational Neuroscience. 2025. Poster presentation. Amsterdam (Netherlands).
  Accessible in: `https://2025.ccneuro.org/poster/?id=YJsXpaWng7`
  Described in chapter 7.

- Evolution of Low-Level and Texture Human-CLIP Alignment. Pablo Hernández-Cámara, Jose Manuel Jaén-Lorites, Jorge Vila-Tomás, Jesus Malo, Valero Laparra. 8th Annual Conference on Cognitive Computational Neuroscience. 2025. Poster presentation. Amsterdam (Netherlands).
  Accessible in: `https://2025.ccneuro.org/poster/?id=wNFrCi3AO1`
  Described in chapter 8.

- Contrast Sensitivity Function of Multimodal Vision-Language Models. Pablo Hernández-Cámara, Alexandra Gomez-Villa, Jose Manuel Jaén-Lorites, Jorge Vila-Tomás, Jesus Malo, Valero Laparra. 8th Annual Conference on Cognitive Computational Neuroscience. 2025. Poster presentation. Amsterdam (Netherlands).
  Accessible in: `https://2025.ccneuro.org/poster/?id=m5YIx5sGg2`
  Described in chapter 8.

Finally, a theoretical commentary about some of the results was also published in an international journal:

- A separate theory-on-top level may be inspiring, but it is neither separate nor enough. Jesús Malo, Pablo Hernández-Cámara. Journal of Physiology, 602, 9, 2024.
  Accessible in: `https://doi.org/10.1113/JP279550#support-information-section`
  Described in chapter 9.

The four peer-reviewed journal papers, which constitute the core scientific contributions of this thesis, are included in the following pages:

Pattern Recognition Letters 173 (2023) 64–71

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Neural networks with divisive normalization for image segmentation

Pablo Hernández-Cámara *, Jorge Vila-Tomás, Valero Laparra, Jesús Malo

Image Processing Lab., Universitat de València, 46980 Paterna, Spain

ARTICLE INFO

Editor: Song Wang

Dataset link: https://www.cityscapes-dataset.com/, http://people.ee.ethz.ch/~csakarid/SFSU_synthetic/, Semantic Foggy Scene Understanding with Synthetic Data (Reference data), Cityscapes dataset (Reference data)

Keywords:
Adaptation
Manifold alignment
Nonlinear interactions
Divisive normalization
Image segmentation
Cityscapes dataset

ABSTRACT

One of the key problems in computer vision is adaptation: models are too rigid to follow the variability of the inputs. The canonical computation that explains adaptation in sensory neuroscience is *divisive normalization*, and it has appealing effects on image manifolds. In this work we show that including *divisive normalization* in current deep networks makes them more invariant to non-informative changes in the images. In particular, we illustrate this concept in U-Net architectures for image segmentation. Experiments show that the inclusion of *divisive normalization* in the U-Net architecture leads to better segmentation results with respect to the conventional U-Net. The gain increases steadily when dealing with images acquired in bad weather conditions (from 3% of IoU increase in regular weather up to 20% on high fog). In addition to the positive results on the Cityscapes and Foggy Cityscapes datasets, we explain these advantages through the visualization of the responses: the equalization induced by the *divisive normalization* leads to more invariant features to local changes in contrast and illumination.

1. Introduction

A fundamental problem in image analysis is the variability of the image manifold depending on acquisition conditions [1]. Usually sources are not stationary: for instance, the visual texture and color of an object can be very different at different locations in the image. Examples are known from long ago: Leonardo da Vinci was the first to use the term *aerial* or *atmospheric perspective* as a method of creating the illusion of depth in a painting by modulating color and contrast to simulate changes effected by the atmosphere on the things seen at a distance [2]. These effects were later explained quantitatively by the physics of light-matter interaction [3], but they have obvious negative impact on image analysis such as image segmentation. Fig. 1 shows a specific example of the problems posed by uncontrolled image acquisition. In the patches highlighted in red in Fig. 1, the visual texture (spatial frequency and contrast) and the color of the object locally change due to the illumination, shadows, atmospheric scattering, depth and perspective. See the zoomed patches in the center panel. The associated problems for statistical learning are shown in the scatter plots at the right. The top-right plot of Fig. 1 shows changes in the spatial texture defined by the luminance of neighbor pixels. In this plot each point corresponds to a three-pixel image where each dimension is the luminance value of each pixel and samples are colored according to their region in the original image. One can see that the contrast (distance from the diagonal) is smaller for lighter regions (away from the origin), and the energy in the different directions (spatial frequency) locally changes for the different clusters along the image. Moreover, atmospheric scattering also implies local changes in the hue, see the shift towards blue (towards lower dominant wavelengths) in the CIExy color diagram at bottom-right plot of Fig. 1.

Adaptation to (and compensation of) these non-informative factors of image variation is key for optimal segmentation. Successful image representations for segmentation should merge the separate clusters in the scatter plots of Fig. 1.

Current deep-learning models for image segmentation (e.g. the popular U-Nets [4]) should display this behavior when trained for a variety of acquisition conditions. However, it is not clear how these deep models deal with this problem and, more important, how this adaptive behavior could be enforced and controlled.

Our contributions in this work are the following:

- We **propose** the use of the canonical computation that accounts for adaptation in biological neurons, namely the *divisive normalization* [5]. This improves adaptation of artificial networks in an explainable way (Sections 2 and 3).

* Corresponding author.
E-mail addresses: pablo.hernandez-camara@uv.es (P. Hernández-Cámara), jorge.vila-tomas@uv.es (J. Vila-Tomás), valero.laparra@uv.es (V. Laparra), jesus.malo@uv.es (J. Malo).

https://doi.org/10.1016/j.patrec.2023.07.017
Received 18 November 2022; Received in revised form 21 July 2023; Accepted 30 July 2023
Available online 9 August 2023

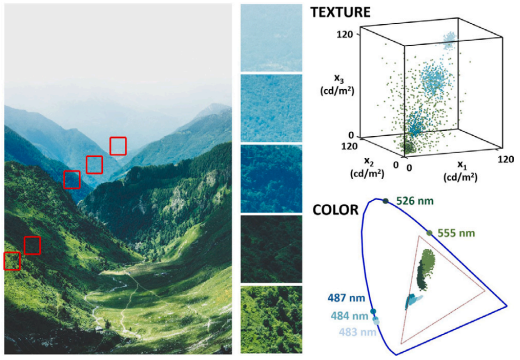


Fig. 1. The problem: non-informative variations imply that the same object has different visual features along the image, complicating the analysis. For instance, patches of the same class (e.g. forest) in different acquisition condition (depth, atmospheric scattering, illumination, fog) lead to separate clusters (different luminance, contrast and hue). See text for details.

- We **experimentally show** that the adaptation ability of network augments when including divisive normalization in classical architectures like U-Nets [4]. In particular, in our experiments we test a standard model with and without divisive normalization on an image segmentation problem under different atmospheric conditions. Results show that divisive normalization helps the model to adapt to non-informative variations for which it had not been trained (Sections 4 and 5).
- We **analyze** the qualitative reasons for this success by performing an ablation study, inspecting the feature maps before and after normalization, and by analyzing the nonlinear response induced by different divisive normalization layers (Section 6).

2. Why using divisive normalization?

The *linear+nonlinear* structure of conventional artificial neurons [6] comes from the seminal computation proposed for sensory neurons [7]:2lm

$$\mathbf{x} \xrightarrow{\mathcal{L}} \mathbf{z} \xrightarrow{\mathcal{N}} \mathbf{y} \tag{1}$$

where the *linear* response, $\mathbf{z} = \mathcal{L} \cdot \mathbf{x}$, is given by the matrix $\mathcal{L}$, that contains the so called (linear) receptive fields, and the *nonlinearities* $\mathcal{N}(\cdot)$ were originally simple thresholds or point-wise saturating functions [8] eventually rectified. These were the inspiration for current sigmoids/ReLU in deep-learning [6].

In sensory neuroscience, the *divisive normalization* model of $\mathcal{N}(\cdot)$ was a way to account for the inhibitory effect of neighbor neurons within a layer [5,9]:

$$y_k = \frac{z_k}{\left(\beta_k + \sum_s \gamma_{k,s} * |z_s|^{\alpha_s}\right)^{\epsilon_k}} \tag{2}$$

where the linear response of a neuron is inhibited (normalized) by a pool of the activity of neighbor neurons. Note that the division/is a point-wise operation, and the sum in the denominator is a convolution over the omitted spatial indices (see pseudocode in Appendix). The constant β_k determines the level in which the pool generates effective inhibition, and the exponents control the norm of the pool. While the interaction kernel in the denominator, γ, can have whatever dense structure [9], in visual neuroscience the spatial interaction is usually assumed to be convolutional in the fovea [10–12].

However, *why using this biological transform to improve artificial networks devoted to image segmentation?*

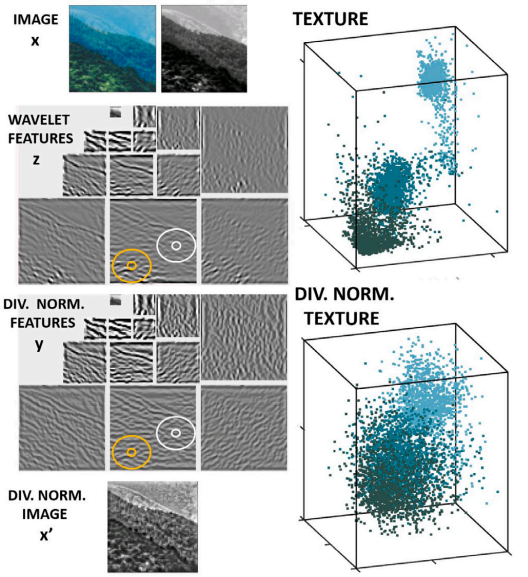


Fig. 2. The idea for the solution: manifold equalization though Divisive Normalization. At the top we have the luminance image $\mathbf{x}$. That is the input to the class of linear+nonlinear models of the V1 cortex based on wavelets and divisive normalization [10–13]. The specific parameters of this example are purely psychophysical, taken from [12], not trained for any image processing task. Below the input we have the wavelet-like linear features, $\mathbf{z} = \mathcal{L} \cdot \mathbf{x}$, and the corresponding divisive normalized features, $\mathbf{y} = \mathcal{N}(\mathbf{z})$. White and orange circles highlight low and high contrast regions (center and surrounds) in the wavelet domain that have been equalized in the divisive normalized features. At the bottom we show the "equalized" image which has been computed directly from the divisive normalized features by inverting the linear wavelet, i.e. $\mathbf{x}' = \mathcal{L}^{-1} \cdot \mathbf{y}$. The scatter plots were obtained as in Fig. 1: they display the distributions of 3×1 image patches taken from the original image, $\mathbf{x}$, and from the "equalized" image, $\mathbf{x}'$.

As a motivation, we give an explicit example on how a biologically plausible *divisive normalization* may improve image segmentation by alleviating the problem described in Fig. 1. The example in Fig. 2 shows a general property of this transform: the *divisive normalization* equalizes the image manifold in such a way that the different clusters corresponding to the same object may merge into a single cluster, thus simplifying class identification.

Fig. 2 considers an illustrative patch from Fig. 1 where the object under consideration (the forest) displays space varying features due to illumination and depth/atmospheric effects. In this example we do not consider color because the spatial texture problem is enough to stress the role of local normalization.[1] In this achromatic context, a reasonable physiological model of texture perception includes a linear layer of wavelet-like filters (in this illustration $\mathcal{L}$ is a steerable wavelet transform as in [11–13]), and a *divisive normalization* (in this case with the psychophysically tuned parameters in [12]). It is important to stress that these biologically sensible parameters were not retrained for this specific example: Fig. 2 just shows the effect of the psychophysical transform in the image manifold.

The neighborhood in γ that defines the pooling region (illustrated in Fig. 2 by the circles in the feature vectors or wavelet bands) has a

[1] Note that, despite this example is focused on luminance/texture for simplicity, local color compensation could also be done using modern divisive normalization formulations of classical Von Kries ideas [14,15].

Fig. 3. U-Net for segmentation including 4 DN layers (in green). Numbers above the layers indicate the number of features and black arrows represent the skip unions. The model without DN layers did not have any of the green layers. The consideration of the DN layers only increases the number of parameters by an 1.8% with regard the No DN model.

qualitatively relevant effect: the division by the local pool moderates the response in regions where contrast is high, in orange, while boosts the response in regions of low contrast, in white. See the difference between the features before and after the local normalization, and note the equalization of the amplitude of the edges along the field of view. The local normalization compensates for the contrast variation along the image due to atmospheric conditions. As a result, while the distribution of the samples from the original luminance channel display separate clusters corresponding to spatial variations of the texture, the different samples have been compacted due to the contrast equalization effect.

The result in Fig. 2 shows how psychophysical mechanisms may reduce the intra class differences by removing the non-informative factors in the signal. This is the reason to try the *divisive normalization* as an extra layer in artificial networks. In the analysis of the results we will explicitly check if U-nets equipped with divisive normalization trained for segmentation also develop this contrast enhancement behavior and the equalization effect shown here for the biological neurons in [12].

2.1. Divisive normalization in other applications

The effect of the *divisive normalization* on the PDF illustrated above is at the core of its use in other image processing applications different from segmentation. Its original use in compression to improve JPEG and MPEG standards [16,17] has been updated in current neural architectures [18,19]. Similarly, better image enhancement and restoration in traditional (fixed) representations using the proposed nonlinearity [20] have been extended to adaptive representations in deep-learning contexts [21]. The field of subjective image quality had benefited from divisive normalization too [22,23], but this use has gained attention again in current neural nets [24–26], and extended to audio signals [27]. Divisive normalization is also important in classification [28–31], where it helps to get higher accuracy than the usual CNNs with layer/batch normalization. Finally, it has been studied for its appealing statistical properties [13,29,32–37], showing that responses after divisive normalization are more independent. The classification models that implement it are more robust to noise and to adversarial perturbations and align better with humans, while models trained for self-supervised learning and super-resolution show contrast invariant tuning properties [38].

However, to the best of our knowledge, this is the first report on the use of Divisive Normalization for image segmentation.

3. Segmentation with and without divisive normalization

In order to confirm the intuition obtained from the 1-layer, purely biological, non-optimized model presented in the previous section just for motivation, the actual experiments implemented the generic *divisive normalization* layer in Eq. (2) in an automatic differentiation context. Therefore, one may include our layer at any place of any architecture and optimize the network for the desired task.[2]

[2] https://github.com/pablohc97/TFM/blob/main/GDN.py.

Table 1
Performance in IoU (mean and standard deviation) in test for each weather condition over ten runs. Gains in IoU percentage due to DN layers with regard to using no DN is also reported. The tick (✓) and the cross (×) indicate if the p-values from the T-test that compares the use of DN layers with regard no using DN for each weather condition are significant ($p < 0.05$) or not ($p > 0.05$) respectively.

Dataset	No DN	1–4 DN	Gain	p-val
Original	0.75 ± 0.02	0.77 ± 0.02	2.7%	✓
Low fog	0.65 ± 0.02	0.70 ± 0.03	7.7%	✓
Middle fog	0.54 ± 0.03	0.62 ± 0.04	14.8%	✓
High fog	0.40 ± 0.05	0.48 ± 0.03	20.0%	✓

Table 2
Reductions of performance due to fog in IoU percentage (test).

Dataset	No DN	1–4 DN
Original - Low fog	−13.3%	−9.1%
Original - Middle fog	−28.0%	−19.5%
Original - High fog	−46.7%	−37.7%

Qualitatively critical points are (a) the nature of the interaction kernel, γ, and (b) the specific locations to include the normalization in the U-Nets for segmentation.

Regarding the nature of γ, the above example points out that the consideration of spatial neighborhoods is convenient to get the local contrast equalization along the visual field required to overcome the local adaptation problem. The recent literature that trains divisive normalization through automatic differentiation uses a range of kernel structures: some do not consider spatial interactions (either in a dense [18,24,35] or convolutional [31] combinations of features); while others do with some restrictions (either uniform weights [39], a ring of locations [29,30], or special symmetries in the space [40]). Following biology [10–13] and the intuition pointed out in the previous section, our experiments use Eq. (2) which includes a general spatial surround, as in [39], but allowing variable weights over space. Additionally a general (dense) interaction over channels was considered. Besides the γ parameter, the β parameter is also trained, but we set α and ϵ to 1 which we found it makes the training more stable.

Regarding where to include the proposed normalization in U-Nets, we recall the *color* and *texture* problems pointed out in Fig. 1: following color appearance models [14,15], constancy maybe addressed by a single normalization of the photoreceptors (right before the first convolution), while texture equalization over space may require normalization at deeper stages where filters tuned to specific patterns have emerged [41]. This intuition lead us to the specific scheme in Fig. 3 where *green* layers stand for the proposed *divisive normalization* (DN) that we included in the four encoding blocks (model 1–4 DN). Note that we propose modifications only in the *encoding/compressive* part, where the features that will lead to the segmentation are computed. This model will be complemented with an ablation study in Section 6.1 where different number of DN layers are located at different depths.

4. Experiments

In the experiments we used the Cityscapes dataset [42], that includes real scenes with annotated segmentation ground truth, and the

Table 3
Performance in IoU (mean and standard deviation) in test for each weather condition over ten runs. Gains in IoU percentage due to DN layers with regard to using no DN is also reported. The tick (✓) and the cross (×) indicate if the p-values from the T-test of the use of DN layers with regard to using no DN are significant ($p < 0.05$) or not ($p > 0.05$). For an easier comparison, results of mean IoU for No DN and 1-4 DN from Table 1 are included.

Dataset	No DN	1 DN	1–2 DN	1–3 DN	2–4 DN	1–4 DN
Original	0.75	0.75 ± 0.02, 0.0%, ×	0.76 ± 0.02, 1.3%, ✓	0.77 ± 0.02, 2.7%, ✓	0.77 ± 0.02, 2.7%, ✓	0.77
Low fog	0.65	0.65 ± 0.04, 0.0%, ×	0.68 ± 0.02, 4.6%, ✓	0.69 ± 0.02, 6.2%, ✓	0.70 ± 0.03, 7.7%, ✓	0.70
Middle fog	0.54	0.53 ± 0.04, −1.9%, ×	0.57 ± 0.03, 5.6%, ×	0.60 ± 0.03, 11.1%, ✓	0.61 ± 0.03, 13.0%, ✓	0.62
High fog	0.40	0.38 ± 0.04, −5.0%, ×	0.41 ± 0.04, 2.5%, ×	0.46 ± 0.04, 15.0%, ✓	0.49 ± 0.03, 22.5%, ✓	0.48

Foggy Cityscapes dataset [43], that includes degraded versions of the scenes simulating poor weather conditions of controlled (low, medium, high) severity. In foggy images, we expect that the spatial varying degradation will be a major problem for models trained with good weather condition images if they are not able to adapt their behavior.

The bottom pixels of each image were cropped from 1024×2048 to 768×2048 in order to remove the front of the car. After that, the images are normalized to the range $[0, 1]$ and resized to 96×256 to reduce the computational demand for the models but maintaining the aspect ratio. We used 2675, 300 and 500 images for training, validation and testing respectively. We used *MAE* loss function (which we found more stable and get better results than the classical cross-entropy), a batch-size of 64 and Adam optimizer with a learning rate of 0.001 during 500 epochs with the original scenes (in regular weather conditions). We performed 10 runs for each model with different seeds. We kept the models with higher Intersection over Union (IoU) in validation (in regular weather) and used these for the test in the four weather conditions. Testing in more general datasets where visibility is strongly reduced will confirm or refute the intuition in Fig. 2.

5. Segmentation results

Table 1 shows the results in test for each model in the four weather conditions for models trained only using the original (good weather conditions) images. It also shows the mean improvements of the use of 4 DN layers with regard to not using DN layers for each weather condition and the significance of the *p*-value of the T-test performed between the ten IoU values obtained by the models that use 4 DN layers with regard the models without DN layers. Table 2 shows how the mean test IoU changes in poor conditions with regard to regular weather for different number of DN.

The first observation is that using 4 DN layers gives better results in IoU than not using them in *all* cases. Second, as expected, for progressively heavier fog, IoU gets reduced in all cases. However the conventional architecture is more sensitive to the decrease in visibility (bigger reductions in performance) than the architecture with 4 DN layers. And third, the gains due to the DN layers get progressively bigger for more challenging acquisition conditions.

Fig. 4 shows two illustrative examples of the predictions, which are consistent with average IoU's in the tables. There is a degradation of the performance when the amount of fog is increased (as expected). However, this effect is less severe for the model that include 4 DN. In particular, the 4 DN model is able to get the border in the shadow and preserve the detection of the car in heavy fog. It is important to stress that non of these models have seen a foggy image during the training, and therefore the 4 DN model is able to better adapt to bad weather conditions.

For a better quantification of the results in Fig. 4 we calculated the confusion matrices for the extreme scenarios (No fog and high fog) for the top image. Fig. 5 shows these matrices, where we calculated the Cohen kappa coefficient [44] for each matrix, which gives an estimation of how diagonal they are. As expected, 4 DN models achieve better results with the original images but specially in the high fog weather, where the use of the divisive normalization becomes more relevant. Note how both models detect the class car in the original images but in the high fog condition the no DN model predicts the car mainly as building or sky while the 4 DN model still detects it.

Table 4
Reductions of performance due to fog in IoU percentage (test).

Dataset	1 DN	1–2 DN	1–3 DN	2–4 DN
Ori. - Low	−13.3%	−10.5%	−10.4%	−9.1%
Ori. - Middle	−29.3%	−25.0%	−22.1%	−20.8%
Ori. - High	−49.3%	−46.1%	−40.3%	−36.4%

6. Analysis and discussion

6.1. Ablation study

We performed an ablation study of DN layers with two complementary objectives. First, to analyze the effect of DN use in deeper and deeper layers and, second, to analyze which layer has the strongest DN effect.

Usually in deep networks, the first layers deal with more low-level characteristics as color changes, while deeper layers are tuned to more complex/texture features [41]. We compare the results of the 1–4 DN model with different (simpler) models. One model just performs a DN at the first layer (1 DN), doing so is similar to an equalization of the colors and very low-level features. Another model performs DN normalization in the first two layers (1–2 DN), a third model has 3 DN layers located in layers 1, 2, and 3 (1–3 DN), and one last model where the DN is included in the 3 last three layers, from 2 to 4, leaving the first one out (2–4 DN).

Table 3 shows the mean results and standard deviations in the test for each model in the four weather conditions. As in Section 5, all models have been trained only using the original (good weather conditions) images. The table also shows the gain due to the DN layers with regard to not using DNs for each weather condition, and the statistical significance of the gain according to the p-values of the T-test. Table 4 shows how the mean test IoU changes in poor conditions with regard to regular weather for different configurations of DN.

First, the main observation is that models with more DN layers get better results than the No DN model, especially as the fog increases. Second, the use of more DN layers leads to higher improvements with regard to the No DN model. Third, when comparing the models with 3 DN layers that leave the first and the last out, we observed almost no difference except in the high fog scenario, when the 2–4 DN model improves the result of the 1–3 DN model. The main conclusion is that the effect of DN is more important in the last layers.

6.2. Feature equalization through DN trained for segmentation

In order to understand why the DN leads to the advantages for segmentation reported above, here we explored its effect in illustrative feature maps (in the 4 DN model). To do this, we analyze the response of some inner channels for an image *before* and *after* the DN. However, before looking at the channels in Fig. 7, consider the original scene(s) and the segmentation(s) in Fig. 6. Again, the 4 DN model gives a better segmentation, specially for the high fog image. More interestingly, Fig. 7 illustrates why a model with DN is able to adapt to the texture degradation due to the fog.

Fig. 7 shows the effect of the second DN layer on two features that we found to be tuned to horizontal and vertical edges (key for object

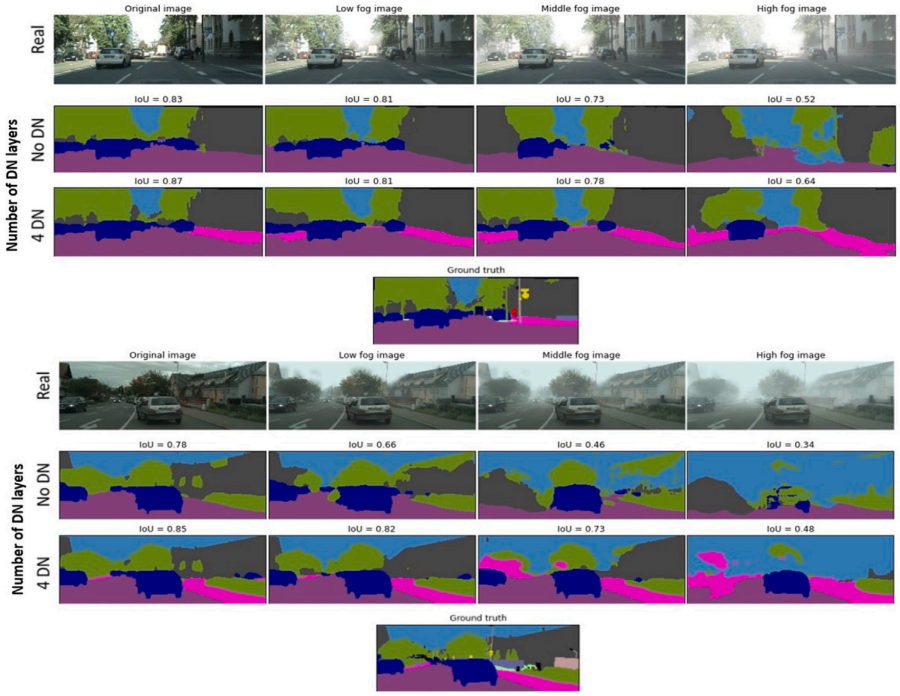


Fig. 4. Two examples of segmentation results on different fog levels (original, low, middle and high fog) for the No DN and 4 DN models.

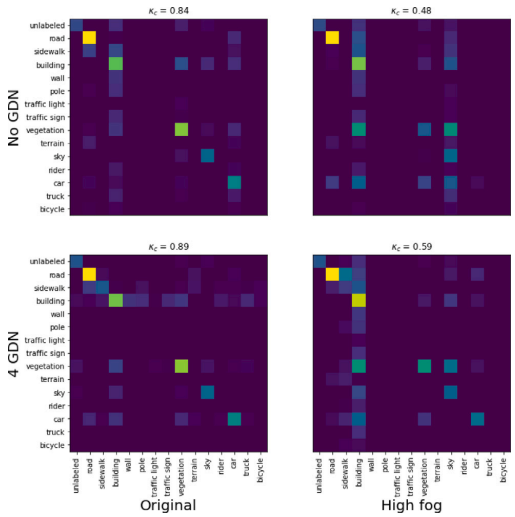


Fig. 5. Confusion matrices from extreme scenarios (No DN and 4 DN for original and high fog images) calculated from Fig. 4 top.

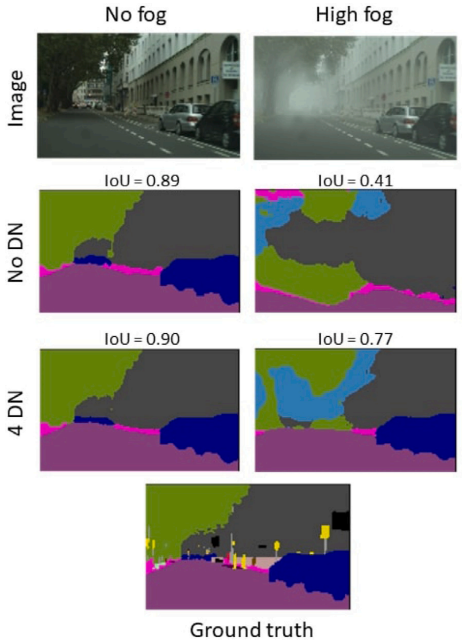


Fig. 6. Scenes and segmentations for the feature maps shown in Fig. 7.

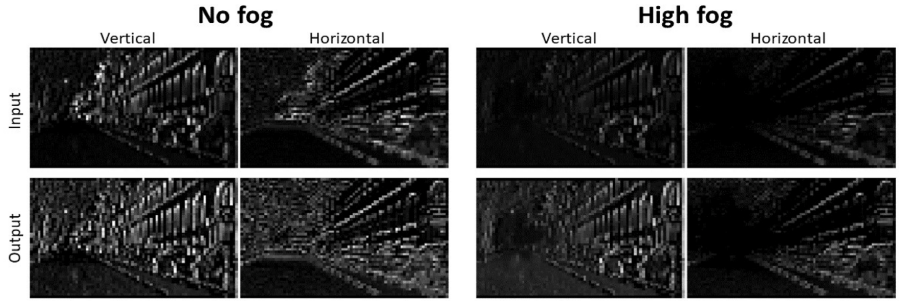


Fig. 7. Effect of the 2nd DN layer of the 4 DN model in two channels (or feature maps), one tuned to vertical edges and one tuned to horizontal edges. The plots show the activity before (input) and after (output) the divisive normalization layer. Results are shown for two different fog situations shown in Fig. 6.

recognition). At the *Input*, we see that in fog, responses to edges are weak and virtually zero at the central part of the scene. That poor information (the *Input*) is what the regular U-Net with no DN has to use. However, edge responses is substantially stronger after the DN (*Output*). The application of the DN recovers most of the edge details lost by the effect of the fog and this enhanced information can be used downstream to improve the quality of the segmentation. The effective result of DN is enhancing low amplitude responses. This is always important in distant regions where there is more scattering due to distance, but it is particularly relevant when there is additional fog which masks the signal. Fig. 7 explains the result in Fig. 6: the convolutional layers extract the edge information. When there is no fog the DN reveals extra details in the distant vegetation but (given the good visibility) both models (No DN and 4 DN) reach similar results because the linear responses are good enough. However, in the high fog scenario, the loss of detail of the convolutional layers without divisive normalization makes the No DN model perform much worse than the 4 DN model, which still detects the parked cars and better identifies distant vegetation despite the fog.

6.3. Nonlinearities in DN trained for segmentation

Fig. 8-top shows *Input* and *Output* of different DNs as nonlinear (saturating) transduction functions. This illustrates the equalization behavior shown above for the feature maps: enhancement of low input values versus moderation of high inputs values. And this happens both for pixel intensities (1st DN layer), as well as for the responses of the filters tuned to edges (2nd DN layer). Moreover, the consideration of these nonlinearities that emerge in segmentation U-Nets with DNs (Fig. 8-top) together with known behaviors in biological vision (Fig. 8-bottom) leads to two interesting suggestions: (1) the behavior of early and late nonlinearities (here 1st DN and 2nd DN) is in line with the intuition on color/texture separation mentioned in building the models, and (2) the responses to RGB channels *qualitatively seem* to saturate and adapt as human luminance/brightness receptors [45], and filters tuned to edges *qualitatively seem* to saturate and adapt as human sensors of edge/contrast [34]. This comparison relies on the fact that input pixel intensities are correlated to the input luminance in biological systems, and feature intensities in the 2nd layer are correlated with contrast of band-pass stimuli used to probe biological mechanisms.

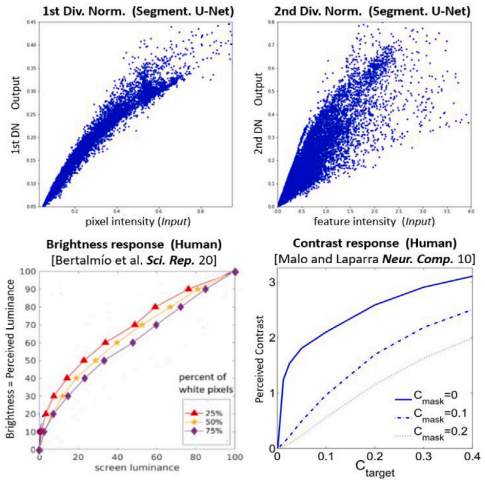


Fig. 8. *Top:* Nonlinear input/output transduction in the 1st (left) and 2nd (right) DN layers of the 4 DN model. The transductions are adaptive: the same input may have a range of outputs (depending on a context not considered in the abscisas). *Bottom:* responses of qualitatively similar sensors in human vision: Weber law for luminances [45] (left) and contrast nonlinearities of edge sensors [34] (right) both adaptive with background.

7. Conclusions

Here we have analyzed the ability of a psychophysically inspired non-linear layer (the divisive normalization) to help artificial neural networks in adaptation tasks. In particular, we presented results in image segmentation when including divisive normalization in the U-Net architecture. We found that introducing divisive normalization layers in all the encoding blocks leads to consistent improvements in IoU in the Cityscapes dataset. More important is the effect of Divisive Normalization in generalization: a model trained only with images acquired in good weather conditions obtains a substantial increase in IoU over conventional U-Net when tested with images in bad weather conditions. In fact, the higher the level of fog introduced, the higher the advantage.

We presented an ablation study to analyze the advantages of using DN at different depths of the network. The main conclusion is that the effect of DN is more relevant in the deeper layers. The presented visualizations of feature maps and nonlinearities suggest explanations

for the benefits of Divisive Normalization: this transform enhances the responses of filters tuned to edges and this compensates the effects of scattering and perspective, and this is particularly important when the acquisition conditions are poor. Finally, we report an interesting qualitative similarity between the nonlinearities that emerge in the DN layers and equivalent sensors in biological vision. In conclusion, consistently with the results in other image processing applications, this transform helps the segmentation models to generalize over the non-informative variations introduced by contrast and luminance changes.

Future work should extend the experiments to other datasets and image acquisition conditions. Quantitative analysis of the changes in the feature maps (beyond the qualitative visualization shown here) could reveal the limits of the achievable gains. In this regard, open data simulators to generate images in controlled ways will be crucial to understand the benefits of this transform. Moreover, the qualitative similarity between the behavior of the neurons trained for segmentation and biological mechanisms of brightness and contrast perception suggested here should be further studied.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

In this section we provide the data on which the models have been trained and evaluated. To train them we have used the Cityscapes Dataset (https://www.cityscapes-dataset.com/) and to test them we have also used the Foggy Cityscapes Dataset (http://people.ee.ethz.ch/~csakarid/SFSU_synthetic/).

Semantic Foggy Scene Understanding with Synthetic Data (Reference data) (http://people.ee.ethz.ch/~csakarid/SFSU_synthetic/)
Cityscapes dataset (Reference data) (https://www.cityscapes-dataset.com/)

Acknowledgments

This work was supported in part by MICIIN/FEDER/UE, Spain under Grant PID2020-118071GB-I00 and PDC2021-121522-C21, and in part by Generalitat Valenciana, Spain under Projects GV/2021/074, CIPROM/2021/056 and CIAPOT/2021/9. Some computer resources were provided by Artemisa, funded by the European Union ERDF and Comunitat Valenciana as well as the technical support provided by the Instituto de Física Corpuscular, IFIC (CSIC-UV).

Appendix. Pseudocode for the divisive normalization

Input, z, **output**, y: $z, y \in \mathbb{R}^{H\times W\times C}$, **and parameters:** $\gamma \in \mathbb{R}^{w_1\times w_2\times C\times C}$; $D \in \mathbb{R}^{H\times W\times C}$; $A \in \mathbb{R}^{H\times W\times C\times C}$ and $\alpha, \beta, \epsilon \in \mathbb{R}^{C}$.

where w_1 and w_2 are the height and the width of the convolutional kernel, H and W are the height and the width of the image, and C the number of channels.

Algorithm 1 Pseudocode of equation (2)

for k in C **do**
  for s in C **do**
    $A[:,:,k,s] = \gamma[:,:,k,s] * |z[:,:,s]|^{\alpha[s]}$
  end for
  $D[:,:,k] = \left(\beta[k] + \sum_s A[:,:,k,s]\right)^{\epsilon[k]}$
  $y[:,:,k] = z[:,:,k]./D[:,:,k]$
end for

Algorithm 2 Convolution 2D: $A[:,:,k,s] = \gamma[:,:,k,s] * |z[:,:,s]|^{\alpha[s]}$

$m = 0$
for i' in w_2 **do**
  for j' in w_1 **do**
    $m = m + \gamma[i', j', k, s] \cdot |z[i - i' + \lfloor w_2/2 \rfloor, j - j' + \lfloor w_1/2 \rfloor, s]|^{\alpha[s]}$
  end for
end for
$A[i, j, k, s] = m$

Algorithm 3 Point-wise division: $y[:,:,k] = z[:,:,k]./D[:,:,k]$

for i in H **do**
  for j in W **do**
    $y[i, j, k] = \frac{z[i,j,k]}{D[i,j,k]}$
  end for
end for

References

[1] Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (1) (2009) 1–127.
[2] L. da Vinci, Trattato Della Pittura, Ed. G. Langlois, 1651, Paris, France https://archive.org/details/in.ernet.dli.2015.82680.
[3] L. Rayleigh, On the transmission of light through an atmosphere containing small particles in suspension, and on the origin of the blue of the sky, Lond. Edin. Dubl. Phil. Mag. Sci. 47 (287) (1899) 375–384.
[4] O. Ronneberger, et al., U-net: Convolutional networks for biomedical image segmentation, in: 2015 MICCAI, 2015, pp. 234–241.
[5] M. Carandini, D.J. Heeger, Normalization as a canonical neural computation, Nat. Rev. Neurosci. 13 (1) (2012) 51–62.
[6] S. Haykin, Neural Networks and Learning Machines, third ed., Pearson Education, Upper Saddle River, NJ, 2009.
[7] D.H. Hubel, T.N. Wiesel, Receptive fields of single neurons in the cat's striate cortex, J. Physiol. 148 (3) (1959) 574–591.
[8] K.-I. Naka, W.A. Rushton, S-potentials from luminosity units in the retina of fish (cyprinidae), J. Physiol. 185 (3) (1966) 587–599.
[9] M. Carandini, D.J. Heeger, Summation and division by neurons in primate visual cortex, Science 264 (5163) (1994) 1333–1336.
[10] A.B. Watson, J.A. Solomon, Model of visual contrast gain control and pattern masking, J. Opt. Soc. Amer. A 14 (9) (1997) 2379–2391.
[11] M. Martinez-Garcia, et al., Derivatives and inverse of cascaded linear+ nonlinear neural models, PLoS One 13 (10) (2018) e0201326.
[12] M. Martinez-Garcia, et al., In praise of artifice reloaded: Caution with natural image databases in modeling vision, Front. Neurosci. 13 (2019).
[13] O. Schwartz, E.P. Simoncelli, Natural signal statistics and sensory gain control, Nature Neurosci. 4 (8) (2001) 819–825.
[14] J.M. Hillis, D.H. Brainard, Do common mechanisms of adaptation mediate color discrimination and appearance? Uniform backgrounds, J. Opt. Soc. Amer. A 22 (10) (2005) 2090–2106.
[15] M. Fairchild, Color Appearance Models, in: IS and T Wiley Series, Wiley, Sussex, UK, 2013.
[16] I. Epifanio, et al., Linear transform for simultaneous diagonalization of covariance and perceptual metric matrix in image coding, Pattern Recognit. 36 (08) (2003) 1799–1811.
[17] J. Malo, et al., Nonlinear image representation for efficient perceptual coding, Trans. Image Process. 15 (1) (2006) 68–80.
[18] J. Ballé, et al., End-to-end optimized image compression, in: ICLR, 2017.
[19] K. Islam, et al., Image compression with recurrent neural network and generalized divisive normalization, in: CVPR Workshops, 2021, pp. 1875–1879.
[20] J. Gutierrez, et al., Regularization operators for natural images based on nonlinear perception models, Trans. Image Process. 15 (1) (2005) 189–200.
[21] V. Laparra, et al., Perceptually optimized image rendering, J. Opt. Soc. Amer. A 34 (9) (2017) 1511–1525.
[22] A. Pons, et al., Image quality metric based on multidimensional contrast perception models, Displays 20 (2) (1999) 93–110.
[23] V. Laparra, et al., Divisive normalization image quality metric revisited, J. Opt. Soc. Amer. A 27 (4) (2010) 852–864.
[24] A. Hepburn, et al., Perceptnet: A human visual system inspired neural network for estimating perceptual distance, in: ICIP, 2020, pp. 121–125.
[25] K. Ma, et al., End-to-end blind image quality assessment using deep neural networks, Trans. Image Process. 27 (3) (2018) 1202–1213.
[26] E.F.W. Bowen, et al., Visual stream connectivity predicts assessments of image quality, J. Vis. 22 (11) (2022) 4.

[27] T. Namgyal, A. Hepburn, R. Santos-Rodriguez, V. Laparra, J. Malo, What you hear is what you see: Audio quality metrics from image quality metrics, 2023, arXiv:2305.11582.
[28] R. Coen-Cagli, O. Schwartz, The impact on midlevel vision of statistically optimal divisive normalization in V1, J. Vis. 13 (8) (2013) 13.
[29] L.G.S. Giraldo, O. Schwartz, Integrating flexible normalization into midlevel representations of deep convolutional neural networks, Neural Comput. 31 (11) (2019) 2138–2176.
[30] X. Pan, et al., Brain-inspired weighted normalization for CNN image classification, in: ICLR Workshop: How Can Findings About the Brain Improve AI Systems, 2021.
[31] M. Miller, et al., Divisive feature normalization improves image recognition performance in AlexNet, in: ICLR, 2022.
[32] M. Cekic, et al., Neuro-inspired deep neural networks with sparse strong activations, ICIP (2022) 3843–3847.
[33] J. Malo, J. Gutiérrez, V1 non-linear properties emerge from local-to-global non-linear ICA, Netw.: Comput. Neural Syst. 17 (1) (2006) 85–102.
[34] J. Malo, V. Laparra, Psychophysically tuned divisive normalization approximately factorizes the PDF of natural images, Neural Comput. 22 (12) (2010) 3179–3206.
[35] J. Ballé, et al., Density modeling of images using a generalized normalization transformation, in: ICLR, 2016.
[36] J. Malo, Spatio-chromatic information available from different neural layers via Gaussianization, J. Math. Neurosci. 10 (1) (2020) 1–40.
[37] A. Cirincione, R. Verrier, A. Bic, S. Olaiya, J.J. DiCarlo, L. Udeigwe, T. Marques, Implementing divisive normalization in CNNs improves robustness to common image corruptions, in: NeurIPS Workshop, 2022.
[38] V. Veerabadran, R. Raina, V.R. de Sa, Bio-inspired learnable divisive normalization for ANNs, in: SVRHM 2021 Workshop @ NeurIPS, 2021.
[39] M. Ren, et al., Normalizing the normalizers: Comparing and extending network normalization schemes, in: ICLR, 2017.
[40] M.F. Burg, et al., Learning divisive normalization in primary visual cortex, PLoS Comput. Biol. 17 (6) (2021) e1009028.
[41] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: ECCV, 2014, pp. 818–833.
[42] M. Cordts, et al., The cityscapes dataset for semantic urban scene understanding, in: CVPR, 2016.
[43] C. Sakaridis, et al., Semantic foggy scene understanding with synthetic data, Int. J. Comput. Vis. 126 (9) (2018) 973–992.
[44] J. Cohen, A coefficient of agreement for nominal scales, Educat. Psychol. Meas. 20 (1) (1960) 37–46.
[45] M. Bertalmío, et al., Evidence for the intrinsically nonlinear nature of receptive fields in vision, Sci. Rep. 10 (2020) 16277.

Neurocomputing 649 (2025) 130569

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Why Divisive Normalization works in image segmentation?

Pablo Hernández-Cámara [a,*], Jorge Vila-Tomás [a], Paula Dauden-Oliver [a], Nuria Alabau-Bosque [b,a], Valero Laparra [a], Jesús Malo [a]

[a] Image Processing Lab, Universitat de València, Paterna, 46980, Spain
[b] ValgrAI: Valencian Grad. School Research Network of AI, València, 46022, Spain

ARTICLE INFO

Communicated by M. Bianchini

Dataset link: https://huggingface.co/datasets/isp-uv-es/IPL-CARLA-dataset, https://huggingface.co/datasets/isp-uv-es/IPL-CityscapesLuminanceContrasts, https://huggingface.co/datasets/isp-uv-es/IPL-Cityscapes-Illuminants

Keywords:
Divisive Normalization
Segmentation
U-net
Invariance
Generalization
Adaptation
Autonomous driving

ABSTRACT

Scene variability and data diversity are major challenges for image segmentation. Previous work suggested that a biologically motivated computation, the so-called Divisive Normalization, could be useful to deal with image variability, but its effects have not been systematically studied over different data sources and environmental factors. Explanation of the reasons why Divisive Normalization layers lead to enhanced capabilities in segmentation networks is still an open issue. In this work, we respond to the questions of when Divisive Normalization layers are useful for image segmentation, and why. First, we propose systematic domain adaptation experiments based on dissecting the problem according to physically meaningful dimensions to delineate the limits of the gain. Then, we provide quantitative explanations of the special capabilities of Divisive Normalization layers. Results show that neural networks augmented with Divisive Normalization get better segmentation results in a wide range of shifted scenarios and their performance remains more stable with regard to the considered environmental factors and the nature of the image sources. This behavior is understood in two ways: (1) by quantifying the invariance of the responses that incorporate Divisive Normalization, and (2) by illustrating the adaptive nonlinearity of the different layers that depend on the local activity.

1. Introduction

Local response normalization in neural networks is known to be a key factor for image classification [1], and batch normalization is a standard element in deep learning [2]. Divisive Normalization (DN) is a computation known to happen in the visual brain [3,4] that has been extensively used in image coding [5,6] and in modeling subjective distances between images [7,8]. More recently, this biologically motivated form of local normalization has been proposed for semantic image segmentation, showing promising results [9,10].

In particular, these two previous works [9,10] demonstrated that the use of Divisive Normalization can be beneficial in segmentation. However, these studies did not analyze why the Divisive Normalization helps or under what conditions it offers advantages. Therefore, important questions remain open:

- In which regions of the input signal variability does Divisive Normalization improve segmentation?
- Does it enhance robustness to changes in illumination, texture, weather, or data source?
- And crucially, what is the mechanism behind these improvements?

Answering these questions is essential—especially in safety-critical tasks—because aggregate performance metrics often obscure failure modes. Understanding **when** and especially **why** a system fails or succeeds is key for trustworthy deployment [11,12]. In this regard, explainability given by the analytical nature of Divisive Normalization is a good complement to pure maximization of the performance measure [13]. And quantitative descriptions of the reasons for improvement (e.g. invariance of the representation) are highly desirable too.

The above scientific questions are still open issues as they were not addressed in [9,10], just focused on the proposal of using Divisive normalization in segmentation. This work is devoted to answer those questions. Particularly, this work builds directly upon [10] and addresses its open questions by making the following key contributions:

- **Systematic environmental control:** We propose a structured evaluation framework that manipulates visual properties such as luminance, achromatic and chromatic contrast, and spectral illumination. This setup allows us to test the performance of DN under diverse real and synthetic conditions and identify when it is most beneficial.

* Corresponding author.
E-mail address: pablo.hernandez-camara@uv.es (P. Hernández-Cámara).

https://doi.org/10.1016/j.neucom.2025.130569
Received 15 November 2024; Received in revised form 6 May 2025; Accepted 21 May 2025
Available online 23 June 2025

- **Quantitative analysis of invariance:** We introduce metrics to measure how invariant the network's predictions are to environmental changes, and we demonstrate that DN leads to more stable, generalizable representations.
- **Layer-wise nonlinearity analysis:** We investigate the adaptive behavior of DN layers across network depth and show how their data-dependent responses contribute to improved robustness.

Together, these contributions shift the focus from demonstrating that DN works to **understanding how and why it works**, offering both theoretical insights and practical guidance for deploying DN in real-world segmentation tasks.

1.1. A challenging scenario to illustrate the diversity problem: autonomous driving

Both works that propose the use of Divisive Normalization in image segmentation use autonomous driving as a useful case study [9,10]. Note that autonomous driving includes scenes at *any* time of the day and night, and may include wild weather conditions, shadows, and illumination varying at different time scales. These variations lead to high dynamic range scenes both in luminance and color: imagine bright streetlights in the night [14] and saturated colors in sunny scenes as colorfulness increases with luminance [15]. Environmental changes modify average luminance and contrast (e.g. in fog or snow) [16] and induce non-additive noise (e.g. rain) [17]. Moreover, radiance level and spectral illumination change along the day [18], and induce nontrivial contrast changes in otherwise continuous objects (e.g. shadows [15,19], or interreflections [20,21]). This uncontrolled scenario implies that similar reflectances and textures lead to highly variant measurements in the camera. The possible fatal consequences of mis-segmentation in autonomous driving [22] imply that models must have an excellent performance despite all the above variabilities, i.e. they have to be invariant under these changes. Moreover, the complexity in these scenes is not easy to simulate in synthetic scenarios, so if computer-generated images were used to train (data-hungry) algorithms [23,24], they could generalize poorly in real life.

In this work, we follow the domain choice of previous works [9,10], but with a **different objective**: rather than focusing solely on performance gains, we use the autonomous driving domain as a controlled, high-variability setting to investigate the functional benefits of Divisive Normalization. Its combination of visual diversity, simulation capabilities, and safety-critical demands makes it a uniquely powerful testbed for our analysis.

1.2. Simple example of the diversity problem

While the above diversity in the images depends on complex interactions in the scenes and on its natural or synthetic origin, a simple univariate analysis of relevant visual descriptors can illustrate the problem. Fig. 1 includes seven natural and synthetic datasets of driving scenes in different environments. We analyze these scenes according to their *luminance*, and to the *achromatic* and *chromatic contrasts* of their textures. Histograms show how diverse these descriptors are. Fig. 1 also includes a segmentation example from each dataset and the negative effect of this diversity when using a segmentation algorithm trained only with real daytime images. Note how the segmentation result becomes worse as the fog level increases, the mis-detection of the car in the night scenario and the mis-detection of a (clearly synthetic) sky in the CARLA scenario.

Description of the scene conditions according to these visual features (defined below in Section 4.1.2), is of course an oversimplification. However, note the clear patterns of change when changing the conditions. For instance, fog increases mean luminance: see how the peak in the first column (rows 1-to-4) moves to the right with the fog level. Fog reduces the achromatic contrast (see the central histogram column, rows 1-to-4) because darker objects become lighter due to the reflected light via scattering. This reduces the energy of spatial modulations of luminance thus impeding texture-based segmentation. Fog also reduces color saturation (and hence chromatic contrast in the right column) because back-scattered light is white (or broadband). This reduces the energy of spatial modulations of color thus impeding segmentation based on chromatic textures. Night scenes (5th row) are obviously darker (left), but interestingly, contrasts (both achromatic and chromatic) are substantially higher (bright streetlights, saturated neon adds...). Finally, completely synthetic (video-game-like) data shows distributions that deviate from natural scenes, as illustrated by the last two rows of Fig. 1. On the one hand, note that the luminance peak in CARLA [29] is darker than real night images in [27], while both contrasts are wider than in natural databases [26,27]. On the other hand, GTA-V [28] displays a larger luminance range than the natural scenes in [26] (day), and [27] (night). Note that the luminance peak in GTA-V is wider than the natural distributions combined.

1.3. Approaches to cope with diversity

Proper consideration of the mentioned variability for successful segmentation is a matter of active research. For instance, [30,31] propose frameworks to restore clean images by removing rain and snow from corrupted images. Also, authors from [27] train series of segmentation models for different periods of time from daylight to nighttime through the sunset in order to cope with illumination variability. In particular, they use the model trained at certain time as starting point for the model valid for the next period of time, so that the resulting model could be invariant under illumination changes. In fact, the rationale of adding synthetic weather conditions to natural scenes [16,17] is taking into account such variability through data augmentation. However, to train models over the widest range of conditions one may need completely synthetic data, as in [28]. That may be a problem because, as illustrated in Fig. 1, synthetic data may have different statistics and its use may lead to the conventional out-of-distribution problem. Then, specific ad-hoc solutions for this change in statistics have to be developed, as for instance [32] who improve synthetic-to-real adaptation using perturbations in the Fourier domain so the model relies only on low-frequency textures. Other approaches change the loss function to take into account variations in the texture statistics [33]. Other forms of data augmentation include the use of web-crawled images during training [34], and the variation of the samples in the feature domain, in some cases using GANs [35], and in other cases using language-vision models [36–38] to modify the image descriptions while preserving the semantic content.

As opposed to (computationally demanding and conceptually trivial) data augmentation, an alternative is designing a computation that embraces the large variability of the input and turns it into an inner representation with reduced variability. A recent bio-inspired layer for image segmentation has been proposed to (intrinsically) cope with this kind of data variability [9,10], the so-called Divisive Normalization of visual neuroscience [3,4]. Divisive Normalization may be good for invariant segmentation because the response of each neuron is normalized by the responses of other neurons tuned to neighbor locations and features, so each response is adapted to the local activity. The works [9,10] were the first to incorporate DN into segmentation models. The authors of [9] focused on the difference with regular batch and layer normalization. In particular, they analyzed the effect of the size and shape of the neighborhoods in the normalization. However, while they report gains in a specific segmentation task, they do not check the invariance to environmental nor data diversity. On the contrary, the proposal in [10] moved closer to this goal by evaluating DN under fog conditions and showing improved performance under reduced contrast. However, its scope was limited: only fog was considered, there was no controlled manipulation of the visual conditions, and the role of data quality (real vs. synthetic) was not explored.

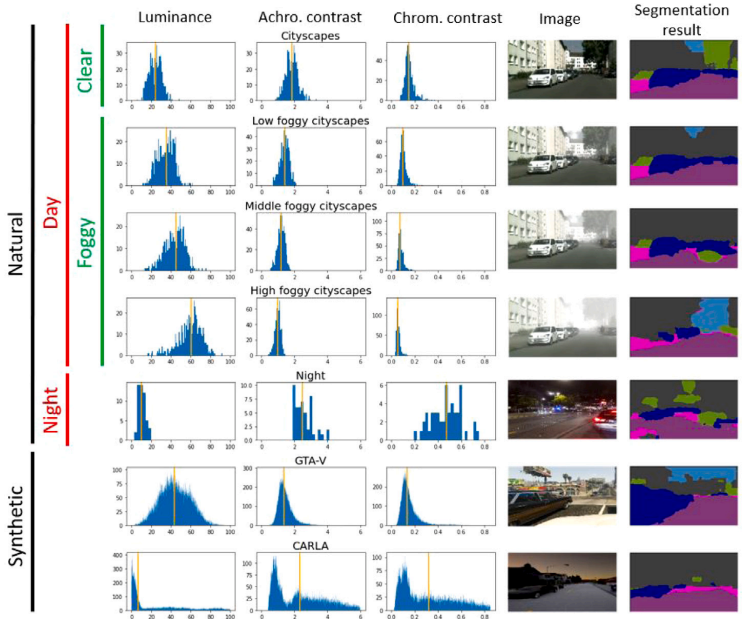


Fig. 1. Motivation: colors and the energy of visual textures change with the environment and data nature and highly affect the segmentation results. Histograms of luminances (left histogram column), achromatic contrast (middle histogram column) and chromatic contrasts (right histogram columns) for 7 different datasets (rows). Vertical orange lines in the histograms represent their median values. The definition of contrasts (energy of spatial modulation of luminance and color) is described in Section 4.1.2. Last two columns show an example of each dataset and the segmentation result of a U-Net model [25] trained with natural daytime and clean images. The first row shows the distributions corresponding to the natural, daytime scenes from Cityscapes [26]. 2nd to 4th rows correspond to the same scenes modified to include different fog levels [16]. 5th row corresponds to real urban night images [27]. 6th row corresponds to scenes from the famous video game GTA-V [28] and last row corresponds to computer-generated scenes using the virtual-reality framework CARLA [29].

In contrast, the present work directly addresses the open questions left by these prior studies: ***When*** *does DN help in segmentation under variable conditions?* ***Why*** *does it help? And* ***how*** *does DN influence internal representations and prediction stability?*

The rest of the paper is organized as follows: First, in Section 2 we review the formulation of Divisive Normalization. Second, in Section 3 we recall the different models (neural networks) and introduce the data (scenes to segment) that we use in our study. More importantly, we introduce the proposed methodology for systematic environmental control following the intuition in Fig. 1. Next, in Section 4 we expose the results that we obtain over different data diversity factors and over the different controlled changes we introduce. Then, in Section 5 we analyze where the advantage of the Divisive Normalization comes from and how it is achieved and finally we conclude in Section 6 with an example of use showing the better segmentation of the models with Divisive Normalization.

2. Background on Divisive Normalization

Divisive Normalization [3] is the canonical computation that accounts for adaptation in biological neurons [4]. It is a local normalization in which the response of a sensor is normalized not only by its own value but also by taking into account the values of its local surroundings:

$$y_k = \frac{z_k}{(\beta_k + \sum \gamma_{k,s} * |z_s|^{\alpha_s})^{\epsilon}_k} \tag{1}$$

where the linear response of a sensor, z_k, is inhibited (normalized) by a pool of the activity of neighbor sensors. In the context of computer vision models that work with images as input, it implies that one pixel is modulated not only by its own value but also by the values of close pixels. Note that the division is a point-wise operation, and the sum in the denominator represents the interaction between the considered pixel and its neighborhood. The exponents ϵ and α control the norm of the pool. The ratio between the constant β and the convolution kernel γ determines the level of non-linearity. The interaction kernel in the denominator, γ, can have any arbitrary structure to take into account the surroundings but here we build it to have the form of a usual spatial convolution with dense connections between the input channels. Unlike traditional nonlinearities such as ReLU, which apply a fixed point-wise transformation independently to each pixel, Divisive Normalization introduces a contextual, adaptive mechanism. The normalization depends on both the pixel's activation and its neighbors through learnable parameters, making it a powerful, input-dependent operation that adapts to the input signal.

3. Models, data, and proposed control of environmental factors

3.1. Models

Building the Divisive Normalization non-linearity in an automatic differentiation environment allows us to use it and optimize its parameters in any machine learning model.[1] In this work, we focus on the U-Net architecture [25], which has become a conventional backbone in image segmentation [39] due to its simplicity, effectiveness and widespread use. Importantly, prior work introducing DN into segmentation also used U-Net [10], allowing us to ensure comparability

[1] https://github.com/pablohc97/TFM/blob/main/GDN.py

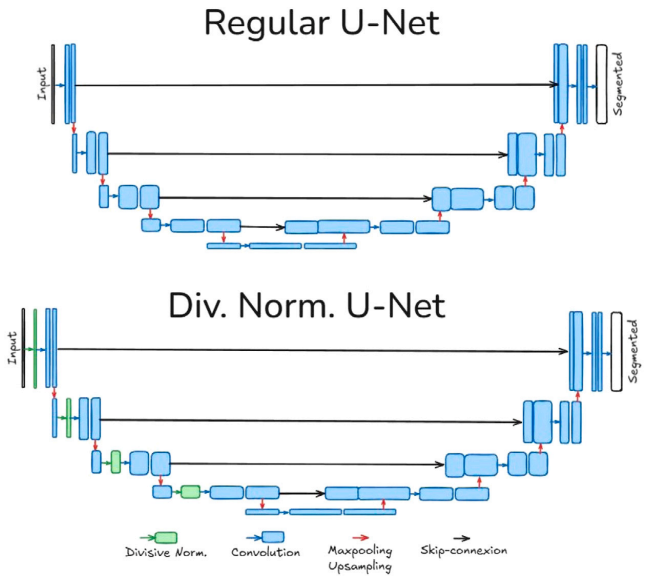


Fig. 2. U-Net for segmentation with and without Div. Norm. layers: Comparison of the baseline U-Net architecture (top) and the modified U-Net with four Divisive Normalization (DN) layers (bottom). DN layers are shown in green and are inserted at the beginning of each encoder block. Both models share the same structure in terms of convolutional layers, skip connections (black arrows), and pooling/upsampling operations (red arrows). The addition of DN layers increases the model size by only 1.8%, preserving overall capacity. The legend indicates the layer types used in the diagram. This architectural setup follows [10] and has been updated for clarity.

and isolate the specific contribution of DN. By maintaining the same architectural foundation, we avoid introducing confounding factors and can focus our analysis on how DN influences robustness and representation. Particularly, we study the behavior of Divisive Normalization by building two model variants: a baseline U-Net with no DN layers (referred to as no-DN), and a modified U-Net with four DN layers inserted at the start of each encoder block (4-DN). Fig. 2 illustrates both architectures and highlights the placement of DN layers in the modified model.

We built our models as in [10] to be able to compare with their results. Therefore, in our models, γ corresponds with a 3×3 convolution kernel so that the close neighbors around the considered pixel are taken into account. The β parameter is also trained, but we set ϵ and α to 1 which makes the training more stable.

To quantify the models' performance, we use the commonly used segmentation metric Intersection over Union (IoU). It measures the overlap between the predicted segmentation mask and the real ground truth.

3.2. Data

In our experiments, we used four different datasets to train and test the models and assess the Divisive Normalization effect.

We used Cityscapes Dataset [26], one of the most famous semantic segmentation datasets for autonomous driving. It includes scenes with 30 classes annotated segmentation ground truths. Images were taken in good weather conditions and during the day by a car on the streets of 50 cities in Germany.

In addition, we use the Nighttime Driving-test dataset [27]. It contains 50 segmented real images from Swiss cities taken at night. This dataset will allow us to test the Divisive Normalization good effect when facing low-luminance images.

The previous datasets are made of real images. However, generating real datasets is very time-consuming because we need to manually generate the ground truths and they do not cope with all the possible data variability that models can find in real life. For these reasons, synthetic datasets are rising in popularity. Following this idea, we also used two synthetic datasets. First, we used the CARLA Simulator (Car Learning to Act) [29]. It is an open simulator for urban driving, developed as an open-source layer over Unreal Engine 4. We generate a new dataset that contains 20000 images of all weather (sunny, rainy and foggy) and time conditions (morning, day, evening and night) from two different simulated cities with their corresponding segmentation ground truth [40]. We divide it into 80%, 10% and 10% to train, validate and test, which gives us 16000, 2000 and 2000 train, validation and test images.

Finally, we used the GTA-V dataset [28]. It contains almost 25000 synthetic images rendered using the open-world video game Grand Theft Auto V. Images are from the car perspective in the streets of American-style virtual cities and they also include different weather (sunny and rainy) and time (day, evening and night) conditions.

To summarize all the datasets, Fig. 3 shows an image of the four datasets and their associated segmentation ground truth, which we want to predict with our models.

3.3. Environmental factors: partition and control

To test how different environmental factors affect the models and to check the effect of Divisive Normalization in controlled experiments, we have to characterize and modify the visual appearance of the images. To do so, we follow the intuition introduced in Fig. 1.

We first consider Foggy Cityscapes [16]. It is a synthetic foggy dataset that simulates fog on the real scenes of Cityscapes. Each Cityscapes image is rendered from a clear image and a depth map to include fog of controlled (low, medium and high) severity. Each level is characterized by a constant attenuation coefficient (0.005, 0.1 and 0.2) corresponding with a visibility range of 600, 300 and 150 meters respectively. Therefore, this dataset has the same Cityscapes images but with simulated fog as shown in Fig. 4.

Fig. 3. Example images and segmentation ground truth of the datasets. From left to right an image of Cityscapes, Nighttime Driving, CARLA Simulator and GTA-V.

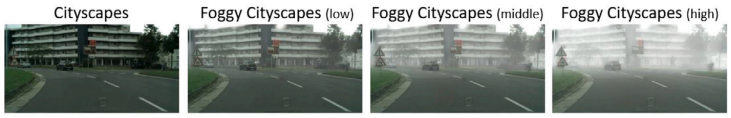


Fig. 4. Example of fog severities. An image from Cityscapes and their corresponding versions in Foggy Cityscapes with different severities.

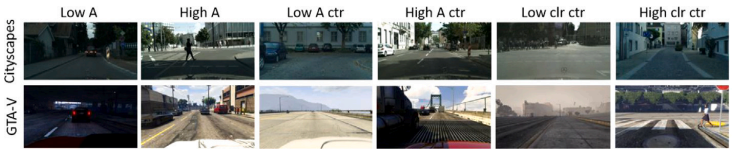


Fig. 5. Representative extreme images of the dataset partitions. From left to right it shows the images with the lowest and higher mean luminance, achromatic contrast and chromatic contrast.

In order to push the models to the limit, we test them on the more difficult or extreme scenarios, i.e. in the images of the datasets which have extreme visual appearance according to simple descriptors as those illustrated in Fig. 1. To do so, we create different partitions of the real and synthetic data (Cityscapes and GTA-V) according to the extreme values of luminance, achromatic contrast and chromatic contrast. The rationale to propose these partitions is challenging the models that may solve the segmentation problems using color or texture variations in certain luminance levels.

The methodology to identify extreme subsets is the following. We first express the real and synthetic images in a classical Achromatic Tritanopic (red-green) and Deuteranopic (yellow-blue), or ATD, color space [15]. Once in this color space, we can easily define the mean luminance, μ_A, and the achromatic and chromatic contrast, C_a and C_{chro} as:

$$C_a = \sqrt{2} \cdot \frac{\sigma_A}{\mu_A} \tag{2}$$

$$C_{chro} = \sqrt{2} \cdot \frac{\sqrt{\sigma_T^2 + \sigma_D^2}}{\mu_A} \tag{3}$$

where μ_A is the spatial mean of the achromatic channel A, and σ_i correspond to the standard deviation of channel i over the spatial extent of the image. The above descriptors are applicable to natural images but are consistent with the standard definition of Michaelson contrast for sinusoidal gratings of luminance [41], and with the definition of perceptual colorfulness [15]. These contrasts represent the amplitude (or energy) of the achromatic and chromatic textures in the image. Once we have (μ_A, C_a, C_{chro}) for each image, we sort them accordingly and we identify the subsets which are in the 15th/20th and 85th/80th percentile of each visual feature for the GTA-V and Cityscapes datasets. This determines six new partitions: low mean luminance (called Low A), high mean luminance (High A), low achromatic contrast (Low A ctr), high achromatic contrast (High A ctr), low chromatic contrast (Low chrom ctr) and high chromatic contrast (High chrom ctr).

Fig. 5 shows the extreme images of these six partitions for both datasets. Note that low mean luminance corresponds with very dark images while high mean luminance images are really bright. On the achromatic contrast partitions, the lowest ones are images that are almost flat, texture-less scenes, while the high achromatic contrast images have both really dark and bright sections. In the low chromatic contrast images, we find images without almost any color but in the high chromatic contrast images there are more colorful images over different backgrounds.

Although, as seen in the results section, these dataset partitions and the foggy scenes help us to test the model in extreme conditions and they are useful to get initial ideas of where the critical situations are, these subsets are limited in different ways. First, the values of the descriptors for these subsets cannot be controlled in a smooth way, and second, the range of the visual descriptors is limited by the set of images already available in the original datasets.

In order to overcome these limitations and extend the range of our exploration in a systematically controlled way, we artificially modify the original images in five relevant visual dimensions: the already mentioned (1) mean luminance, (2) achromatic contrast, (3) chromatic contrast, and the spectral illumination, which amounts to (4) hue angle, and (5) saturation.

First, we address the variation of the three descriptors introduced in Fig. 1. We select 100 images from the Cityscapes test dataset and manually increase and reduce their mean luminances, achromatic contrasts and chromatic contrasts in turns in the ATD space. In this way, we generated a tensor of 3 dimensions (μ_A, C_a, C_{chro}) where in each point we have 100 Cityscapes images with fixed luminance, achromatic contrast and chromatic contrast. We modified the original values of the descriptors by factors from 0.5 to 1.4 and the set is available here [42]. Fig. 6 illustrates three different slices from the tensor, fixing the constant dimension to its original value.

Second, we can introduce variations in spectral illumination in order to shift the colors in the scene in a systematic way. To do that, we use the following *approximated but convenient* approach. We generate

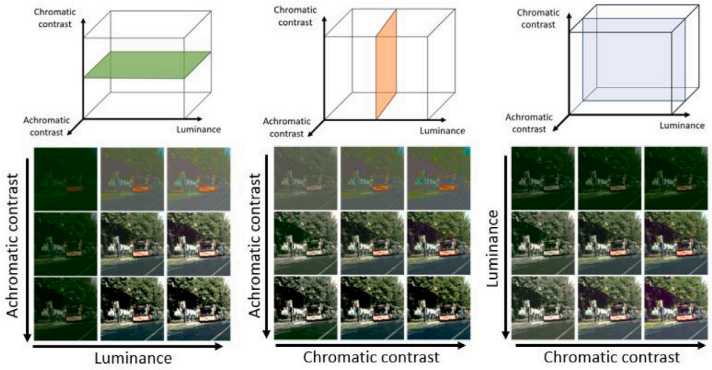


Fig. 6. Controlled modification of luminance and contrasts. Slices and examples from the 3D tensor (luminance, achromatic contrast and chromatic contrast). Note that the central image represents the original image. Figure shows 3 values/dimension but the actual intervention took 10 linearly spaced values/dimension.

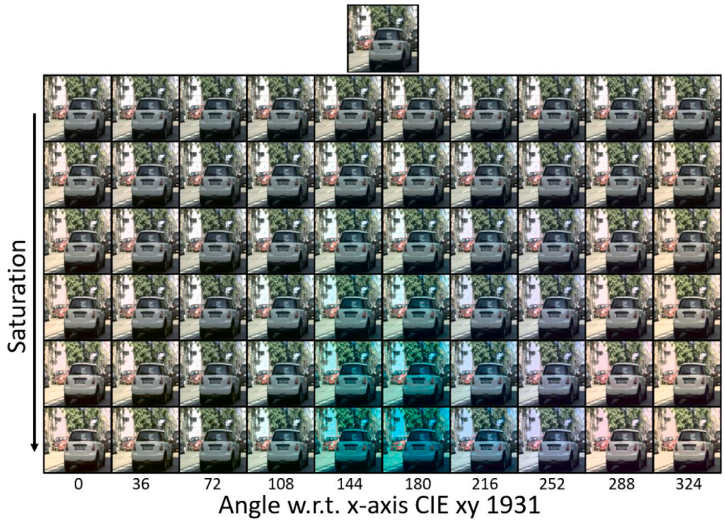


Fig. 7. Controlled change of spectral illumination. The top image is one original image and the bottom matrix shows the obtained images from different illuminant hues (orientation angle in CIE xy diagram) and saturations (distances to the white point). This illustration shows 10 hues but the intervention considered spectra of 20 linearly spaced hue angles.

a white (equienergetic) spectrum and find the lambertian reflectance of each pixel so that when the white light is reflected, we obtain the closest tristimulus values in each pixel. Once we have the reflectance of each pixel, we can generate light spectrums of different dominant wavelengths (different hues) and saturations and use them as illuminats for these reflectant scenes. In particular we select illuminants in a regular 2D-grid in polar coordinates around the white point in the CIE xy 1931 color diagram for 20 hue angles and 6 saturation distances (radius) from the white point.

This approach takes part of the physics of color into account but obviously makes gross approximations. First, it disregards complex mutual illuminations in the unknown geometry of the scenes. We worked on models of interreflections and we know that they give rise to nontrivial distributions of tristimulus values [20,21]. And these do not appear in scenes assumed to be a flat lambertian mosaic. Second, the tristimulus-to-spectrum transform is not univocal because of the strong dimensionality reduction in the spectrum-to-tristimulus transforms, or metamerism [15]. Therefore, the hack based on looking for the best matching reflectance for each pixel in the Munsell database to minimize tristimulus error (as done in Colorlab [43]) is just one of other possible solutions to build the hyperspectral scene from the tristimulus scene.

Actual changes of spectral illumination in real scenes will lead to images which will differ from the example shown in Fig. 7. However, reasonable visual aspect of the result justifies the assumption of the approximations because of the benefits we get in extending the databases in a controlled way in order to check the color constancy of the segmentation. The final dataset of 100 images with $20 \times 6 = 120$ modified illuminants is available here [44]. This form of physically meaningful data augmentation for illumination proposed here has been used to check if color discrimination of segmentation networks is aligned with tat of humans [45].

Table 1 shows a summary of the different datasets and modifications and how many images they have that we use to train, validate and test the different models. Note that these controlled experiments are

Table 1
Summary of the number of training, validation and test images per dataset.

Dataset	Train images	Val. images	Test images
Cityscapes	2675	300	500
Nighttime Driving	–	–	50
GTA-V	–	–	25 000
CARLA	16 000	2000	2000
Extreme parts. (GTA-V/City.)	–	–	3750/100
Foggy Cityscapes	8025	900	1500
Lum./contrasts Cityscapes	–	–	3300
Illuminants Cityscapes	–	–	12 000

Table 2
Mean IoU over ten trainings of models trained on real Cityscapes and synthetic CARLA and tested in: Cityscapes, CARLA and GTA-V. Improvements of the use of DN layers with regard to not using DN for each experiment in parenthesis.

Dataset	Cityscapes trained		CARLA trained	
	no-DN	4-DN	no-DN	4-DN
Cityscapes (real)	0.75	0.77 (2.7%)	0.50	0.54 (8.0%)
CARLA (synthetic)	0.51	0.54 (5.9%)	0.90	0.91 (1.1%)
GTA-V (synthetic)	0.55	0.61 (10.9%)	0.62	0.65 (4.8%)

designed to go beyond prior work, which analyzed performance only under fog conditions. Our setup enables a more precise understanding of when and why DN improves robustness to input variability.

We will use these setups to test the DN and no-DN models across a wide range of conditions, and we report both aggregate IoU metrics and performance trends within each visual partition. This allows us to identify not only whether DN helps overall, but also in which environmental regimes it offers the greatest benefits.

4. Experiments and results

Using the models and material exposed above, we perform different experiments, consisting of training and evaluating the networks presented in Section 3: the classical U-Net with no Divisive Normalization as baseline, and the modified U-Net with 4-DN layers. We take from [10] the models trained with the Cityscapes dataset and, following the same training procedure, we train more models with the synthetic CARLA dataset. We train each model ten times with the same ten different seeds as in [10]. Therefore, we have four configurations: with and without Divisive normalization trained in Cityscapes or CARLA. These models are then tested on real and synthetic datasets under varying environmental conditions.

4.1. Data diversity

First, we test and analyze the models with and without Divisive Normalization and trained with the real and synthetic images in different diverse datasets. This will allow us to analyze the effect of the Divisive Normalization trained with different data when copping with data diversity such as the nature of the data (real or synthetic), extreme conditions, day or night images or even the image resolution.

4.1.1. Real vs synthetic

First, we test how the results change between real and synthetic images. To do so, we test the models with the real images from the Cityscapes test and synthetic images from the CARLA test and GTA-V.

Table 2 shows the mean IoU results over the ten training of the four models when evaluated in the different datasets. The first thing to notice is that the Divisive Normalization produces an improvement of the IoU results in all the scenarios. Moreover, it is interesting to note that each model gets better results in the test partition of the data it has been trained with. However, this is more extreme for the model trained with CARLA. This effect is due to the high difference between CARLA images and the other datasets as shown in Fig. 1. Finally, the gain due to the Divisive Normalization is higher in the most different test data from the data the model has been trained with.

Table 3
Mean IoU across ten trainings for models trained on good-weather Cityscapes (real) and CARLA (synthetic), evaluated on both the full test sets (top row of each group) and extreme partitions of Cityscapes and GTA-V. DN-related improvements are shown in parentheses. Improvements greater than the average full-dataset gain are marked in bold.

Dataset	Cityscapes trained		CARLA trained	
	no-DN	4-DN	no-DN	4-DN
Cityscapes	0.75	0.77 (2.7%)	0.50	0.54 (8.0%)
City. (Low A)	0.76	0.78 (**3.2%**)	0.43	0.49 (**12.2%**)
City. (High A)	0.74	0.77 (**3.1%**)	0.54	0.56 (3.6%)
City. (Low A ctr)	0.72	0.75 (**3.5%**)	0.48	0.53 (**10.6%**)
City. (High A ctr)	0.78	0.80 (2.6%)	0.49	0.51 (5.0%)
City. (Low clr ctr)	0.74	0.77 (**3.2%**)	0.52	0.56 (**8.7%**)
City. (High clr ctr)	0.74	0.76 (**3.1%**)	0.46	0.52 (**11.5%**)
GTA-V	0.55	0.61 (10.9%)	0.62	0.65 (4.8%)
GTA-V (Low A)	0.46	0.52 (**13.0%**)	0.55	0.58 (**5.5%**)
GTA-V (High A)	0.61	0.67 (9.8%)	0.68	0.69 (1.5%)
GTA-V (Low A ctr)	0.57	0.63 (10.5%)	0.65	0.67 (3.1%)
GTA-V (High A ctr)	0.50	0.55 (10.0%)	0.58	0.60 (3.4%)
GTA-V (Low clr ctr)	0.50	0.57 (**14.0%**)	0.60	0.63 (**5.0%**)
GTA-V (High clr ctr)	0.54	0.59 (9.3%)	0.60	0.62 (3.3%)

Table 4
Mean IoU over ten trainings of models trained on Cityscapes good weather conditions and CARLA synthetic and tested in: Cityscapes, Nighttime Driving, CARLA and GTA-V. Improvements of the use of DN layers with regard to not using DN for each experiment in parenthesis.

Dataset	Cityscapes trained		CARLA trained	
	no-DN	4-DN	no-DN	4-DN
Cityscapes (real-day)	0.75	0.77 (2.7%)	0.50	0.54 (8.0%)
Nighttime (real-night)	0.24	0.29 (20.8%)	0.31	0.33 (6.5%)
CARLA (synthetic)	0.51	0.54 (5.9%)	0.90	0.91 (1.1%)
GTA-V (synthetic)	0.55	0.61 (10.9%)	0.62	0.65 (4.8%)

4.1.2. Regular vs extreme images

In [10] they found that the Divisive Normalization improvement increases with the fog level. It gives us the intuition that probably Divisive Normalization is even more important in extreme scenarios.

Table 3 shows the results of the models when evaluated in the entire (unpartitioned) test sets of Cityscapes and GTA-V (top row of each section), and in the six extreme subsets, based on luminance and contrast partitions. Importantly, the results show that DN improves segmentation even on the full datasets, confirming that its benefits are not limited to specific edge cases. These whole-dataset improvements establish the baseline, and the partition-based results allow us to analyze where DN has the greatest impact. Across most of the extreme partitions, especially in Cityscapes, the 4-DN models outperform the baseline by a larger margin than on the full dataset. For GTA-V, improvements are particularly notable in low luminance (dark) and low color contrast conditions. These findings highlight DN's robustness in challenging visual environments and motivate further analysis in more extreme conditions—such as real night-time scenes with minimal luminance.

4.1.3. Day vs night

To answer our previous question, we test our models in a dataset made of real nighttime images. These images have extremely low luminance and we expect a high gain when using the Divisive Normalization.

Table 4 shows the results of the models when evaluated in the Nighttime Driving dataset. As expected due to the extremely low-luminance, the worst results for all the models happen in the night images. However, in this dataset is where the models trained in real data get a higher increase due to the DN layers, showing its importance. Models trained in CARLA do not get the higher increase due to the DN layers in the night images because the CARLA dataset has some night images and therefore these models have seen low luminance images

Table 5
Mean IoU of low and full resolution models trained on Cityscapes in good weather conditions and tested in Cityscapes. Improvements of the use of DN layers with regard to not using DN for each experiment in parenthesis.

Dataset	Low resolution trained		Full resolution trained	
	no-DN	4-DN	no-DN	4-DN
Cityscapes	0.75	0.77 (2.7%)	0.73	0.76 (4.1%)

during their training. In fact, for these models, the DN higher effect happens when facing real daytime images, because as shown in Fig. 1 it has really different statistics from real daytime images.

4.1.4. High vs low resolution

To test how the image resolution affects the results, we retrain the models with the Cityscapes data but maintain the images at their full resolution, without any resizing. Therefore, we train the models exactly in the same way except for the image resolution. Due to the high computational cost of training models with this high image resolution, we only train a single model with the Cityscapes dataset.

Table 5 shows that training with low or high-resolution images does not give too much difference in the overall IoU results or in the effect of the Divisive Normalization.

4.2. Controlled changes of environment

Previous section results are really good and show that (1) the Divisive Normalization always helps the models to cope with the data variability and improve their results and (2) the use of the Divisive Normalization is even more important in extreme scenarios, such as extremely low-luminance images as night images, and also low and high chromatic contrast images. However, the images we used to understand the effect of Divisive Normalization come from different datasets with different statistics. To analyze the effect of the DN in the models in a completely controlled scenario, we need to maintain always the same base images. In that way, if we do some variation always to the same images, the only variations in the results will come from facing the introduced changes. Therefore, in this section, we select a real dataset, Cityscapes and we will perform some experiments with variations of it.

4.2.1. Fog change

First, we use the Foggy Cityscapes dataset. It consists of the same Cityscapes images with synthetic fog added to simulate three different fog severities. Using these three modified datasets we can test the effect of the Divisive Normalization when facing images with progressively more fog, i.e. progressively more luminance and less contrast as shown in Fig. 1. In addition to the two models trained on the Cityscapes and CARLA datasets, we develop a new model trained in a combination of Cityscapes and Foggy Cityspaces train images (of the three severities), so that we can compare the results between a model trained in completely synthetic images (CARLA model), real clean daytime images (Cityscapes model) and a model trained in real images but with more variability (Cityscapes + Foggy Cityscapes model).

Table 6 shows the results of the three models. [10] found that the improvements due to the Divisive Normalization increase with fog level for a model trained in Cityscapes. We found it also happens with a model trained with synthetic images. As expected, the model trained with a combination of Cityscapes and Foggy Cityscapes gets much more constant results because it has seen foggy images during its training. However, the DN layers still have a positive impact and the gain gets higher in high fog level. Interestingly, gains are more significant when dealing with very different data as the synthetic scenes.

4.2.2. Luminance and contrasts changes

Our previous results highlight that the Divisive Normalization is even more important in very dark images and as the fog level increases. Motivated by these results and to get a deeper analysis of the Divisive Normalization gains, we have to control the exact luminance and contrast of the images. We test the models with 3D (luminance, achromatic contrast and chromatic contrast) modified tensor images.

Fig. 8 shows how the IoU results of the no-DN and 4-DN models trained with the Cityscapes images depend on the luminance and achromatic contrast of the input images, with a fixed chromatic contrast equal to its original value, such as in the green slice in Fig. 6. As expected, the lowest IoU's are obtained in the low luminance and low achromatic contrast conditions. The models that implement the Divisive Normalization layers not only get higher IoU values but the region where they obtain higher values is bigger than for the no-DN models.

It is easier to visualize the benefit of the Divisive Normalization using the gain of the 4-DN models with regard to the no-DN models as shown in Fig. 9 for different values of chromatic contrast and for the models trained with the different data. The first row shows the gains for the model trained with the Cityscapes data and the second row shows the gains for the models trained with the CARLA synthetic data. We can obtain different conclusions from this figure. First, we see that using the Divisive Normalization in the models always helps to achieve better results. Second, the models trained in Cityscapes get the maximum gain at low contrast and luminances, implying that it is especially important for night and foggy conditions. However, what happens when we focus on the models trained with the CARLA synthetic images? We can see that although there is still a high gain region of the Divisive Normalization, it is not located at low luminances and contrasts as before but the greatest improvements happen at high luminances. To understand what is happening, in Fig. 10 we plot the IoU values of the no-DN and 4-DN models when fixing the chromatic contrast to its original value, as we did in Fig. 8. We get that the maximum IoU the models get is highly displaced from the 1-1 original conditions. It implies that the synthetic CARLA images have higher luminances and contrasts than real Cityscapes images and that is why the models get their highest IoU at higher luminances and contrasts than the real images. Also, it explains why the gain region is displaced from the top-left corner in the bottom panel of Fig. 9. The gain happens at lower luminances and contrasts than the original image conditions, which in the CARLA dataset are displaced to higher luminances and contrasts and so the gain region is also displaced.

4.2.3. Illuminant changes

We also test our models with the images that we modified lighting them with different illuminants of controlled dominant wavelength and saturation. Fig. 11 shows directly the gain of using 4-DN models with regard to the no-DN models when they face the different illuminant images for the models trained with the different data. We obtained that including the Divisive Normalization in the models trained with Cityscapes images (first row) improves the segmentation results in almost all the illuminant-intensity space, except for the model trained with Cityscapes data in just a small region of very high saturations, which are almost impossible images in real life.

In the CARLA training scenario (second row), we see that the gain region of the Divisive Normalization models trained with CARLA does not follow the same trend as when the models are trained with real images but again they show higher gains due to the Divisive Normalization, especially in the low saturation.

5. Analysis: invariance and nonlinearities

To understand **why** models with Divisive Normalization (DN) consistently achieve better segmentation performance, we go beyond raw accuracy and explore two key aspects of network behavior. First,

Table 6
Mean IoU over ten models trained on Cityscapes, Cityscapes + Foggy Cityscapes and CARLA and tested in Cityscapes, the different Foggy Cityscapes severities, Nighttime Driving, CARLA and GTA-V. Improvements of the use of DN layers with regard to not using DN for each experiment in parenthesis. Results marked with * are from [10].

Dataset	City. train		City. + Foggy train		CARLA train	
	no-DN	4-DN	no-DN	4-DN	no-DN	4-DN
Cityscapes	*0.75	*0.77 (2.7%)	0.78	0.79 (1.2%)	0.50	0.54 (8.0%)
Foggy (low)	*0.65	*0.70 (7.7%)	0.78	0.79 (1.2%)	0.46	0.52 (13.0%)
Foggy (middle)	*0.54	*0.62 (14.8%)	0.78	0.79 (1.2%)	0.44	0.50 (13.6%)
Foggy (high)	*0.40	*0.48 (22.5%)	0.77	0.78 (1.8%)	0.40	0.46 (15.0%)
Nighttime	0.24	0.29 (20.8%)	0.34	0.36 (5.9%)	0.31	0.33 (6.5%)
CARLA	0.51	0.54 (5.9%)	0.58	0.67 (16.7%)	0.90	0.91 (1.1%)
GTA-V	0.55	0.61 (10.9%)	0.65	0.68 (3.6%)	0.62	0.65 (4.8%)

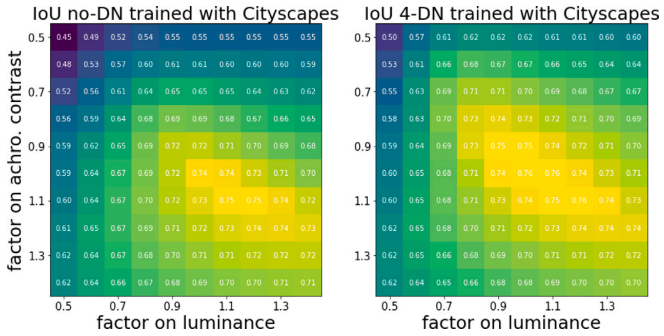


Fig. 8. IoU of the no-DN (left) and 4-DN (right) model trained with Cityscapes images depending on the luminance and achromatic contrast of the input images. In this example, the chromatic contrast of the images has not been changed.

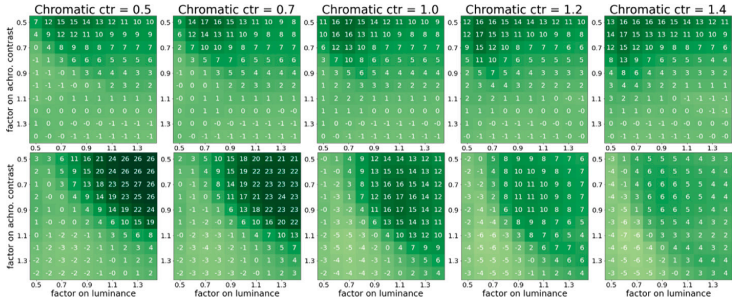


Fig. 9. Relative IoU gains (in percentage) of the 4-DN models with regard to the no-DN models trained with different data depending on the luminance, achromatic contrast and chromatic contrast of the input images. We show five different slides for five different chromatic contrasts. *Top:* models trained in Cityscapes. *Bottom:* models trained in CARLA.

the invariance of the prediction space under environmental changes. We analyze whether DN improves the stability of the model's predictions when the input domain shifts due to changes in luminance, contrast, or color statistics. Using controlled environmental transformations (described in Section 3.3), we quantify the invariance of the model's outputs. Our findings suggest that models with DN produce more invariant predictions across diverse conditions, indicating stronger generalization and a more robust internal representation.

Second, we analyze the adaptive, data-dependent nonlinearities in the intermediate representations. Particularly, we investigate the qualitative effect of DN on feature representations at different network depths. To do this, we apply targeted activations to DN layers and observe how their outputs vary as a function of neighborhood activity. This stimulation method allows us to confirm the nonlinear, adaptive behavior predicted by the DN formulation: the gain of a neuron's response decreases as the activity of surrounding neurons increases. This specific behavior is good for equalizing the response of all neurons in a feature map [46], and this is good for obtaining an invariant representation. In a model trained for segmentation, we show that this input-dependent inhibition increases for deeper layers.

5.1. Quantitative measures of the invariance

Fig. 12 describes the hypothesis we have on the effect of Divisive Normalization on the invariance of the signal representation when the input undergoes changes in the environment. Given the qualitative behavior expected from the literature on biological adaptation [4,46], confirmed in the next section for networks trained for segmentation, one expects that samples that have been put away in the input domain due to changes in the environment will remain close in the inner representation of systems augmented with Divisive Normalization. In contrast, conventional systems would keep these samples separated in

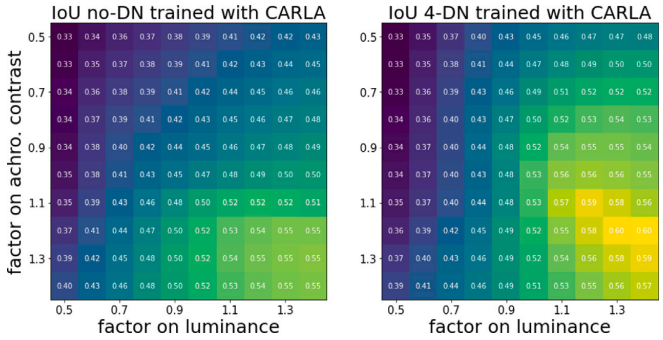


Fig. 10. IoU of the no-DN (left) and 4-DN (right) model trained with CARLA images depending on the luminance and achromatic contrast of the input images. In this example, the chromatic contrast of the images has not been changed.

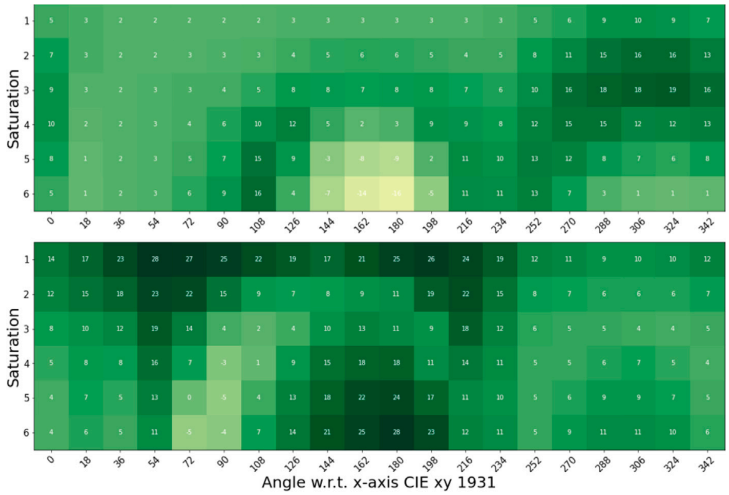


Fig. 11. Relative IoU gains (in percentage) of the 4-DN models with regard to the no-DN models trained with the different datasets depending on illuminant hue and saturation. *Top:* models trained in Cityscapes. *Bottom:* models trained in CARLA.

the inner representation. In other words, the inner representation of the nets with Divisive Normalization will be more invariant under environmental changes than the representation in conventional nets (with no DN). This invariance can be quantitatively assessed by measuring the overlap between the responses of the networks to original and distorted inputs. This is what we do in the experiment shown here.

To do so, we compute the prediction of the no-DN and 4-DN models for the original and the modified images (under fog, luminance, contrast and illuminant changes). Then, we calculate the IoU metric between the prediction of the original image and the prediction of the modified image. This IoU-overlap measure is a measure of the *length* of the black arrows in the inner domain shown in Fig. 12: bigger IoU implies more overlap and smaller distance.

Table 7 shows the IoU results between the predictions over the original images and the predictions over the modified images both for the no-DN and 4-DN models. It shows results for all the modifications we tested: fog, change of luminance, achromatic and chromatic contrasts and illuminants. We got that for all the modifications, the IoUs of the 4-DN model are always higher than the IoUs of the no-DN model. This implies that the predictions of the model that implements the Divisive Normalization change less under all the image changes

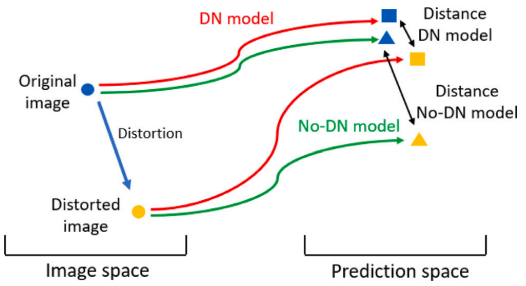


Fig. 12. Quantification of invariance. If a model presents invariance to some distortion, it means that images that are further away in their original space due to that distortion are transformed to close points in the model prediction domain. If the DN induces invariance, the distance between the predictions of the DN model should be smaller than the distance between the predictions of the no-DN model.

Table 7
Iou between the prediction on the original image and the prediction on the modified image for some of the modifications, comparing the results of the models with and without Divisive Normalization layers.

Dataset	no-DN IoU($pred_{ori}$, $pred_{mod}$)	4-DN IoU($pred_{ori}$, $pred_{mod}$)
Low fog	0.766	0.826
Middle fog	0.621	0.714
High fog	0.478	0.609
Achrom ctr = 0.6	0.649	0.786
Achrom ctr = 1.4	0.793	0.832
Luminance = 0.6	0.739	0.781
Luminance = 1.4	0.848	0.886
Chrom ctr = 0.6	0.842	0.844
Chrom ctr = 1.4	0.848	0.887
Angle 0	0.464	0.800
Angle 72	0.751	0.826
Angle 162	0.517	0.562
Angle 252	0.657	0.772
Angle 343	0.365	0.753

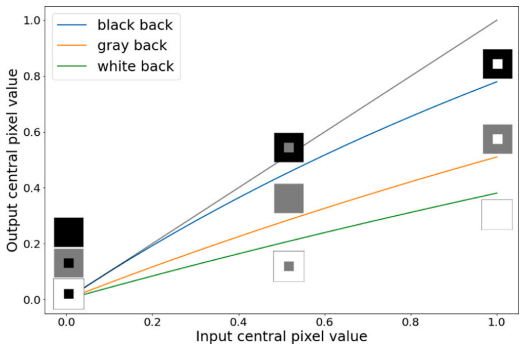


Fig. 13. Checking the nonlinearity of responses. Effect of the first Divisive Normalization model layers on some example images that have different central pixel values and different surrounding values. It shows how the central pixel normalized value depends not only on its value but also on the neighbor pixels.

than the ones of the no-DN model, and therefore that the models with Divisive Normalization are more invariant to these changes.

5.2. Where does the invariance come from? Adaptive nonlinearities

We argue that invariance comes from the adaptation of the responses of each neuron to the responses of the neighbor neurons that is enforced by the local normalization. In order to illustrate this intuition, we visualize the response of neurons of the four Divisive Normalization layers obtained in a model trained for segmentation. In particular, we build 3×3 (same size as the γ kernel in the Divisive Normalization denominator) neighborhoods where the central pixel value changes between 0 and 1 (or the minimum and maximum of the corresponding input feature). We try different surrounding pixel values such as 0, 0.5, and 1.0 (or equivalent fractions of the dynamic range of the input): implying a black, gray, or white background. We pass these images through the four Divisive Normalization layers and register how the central pixel value has been transformed depending on its own value and the value of the surround. Fig. 13 shows some of the generated images and the effect of the first divisive normalization layer depending on the central pixel and surrounding values. Fig. 14 shows the effect for the four Divisive Normalization layers.

We obtain that inputs where the background is very active are strongly inhibited with regard to inputs with inactive backgrounds: the central pixel is more inhibited as the background changes from completely black (surrounding pixel values of 0 and then only the central value is considered) to white (surrounding pixel values of 1, which are the higher values they can have). Looking at the different Divisive Normalization layers we observe this trend in the four layers. Still, it is interesting to note that the non-linearity (departure from the reference black line) increases its effect with the layer depth because the ratio between β over γ value (which controls the linear/non-linear behavior) found in our training is reduced with the layer depth.

The two properties of these curves tend to *equalize* the responses: (1) the saturating shape of the blue curves amplifies low values and moderates high values, and (2) the presence of high values (in the neighbors) moderates the responses (attenuated orange and green curves), while the presence of low values amplifies the responses. The first property transforms the input responses (with eventually disparate range) into output responses within a more compact and stable range performing a sort of univariate histogram equalization. And the second property makes the range of the outputs more equal along (spatial/feature) regions where the inputs had different ranges. The combined effect of these properties makes the range of the output responses kind of invariant to changes in the range of the input, probably leading to the invariance measured in the previous section.

These nonlinearities had been described before in vision science to explain adaptation to luminance, color [47] and contrast [48], and their effect on multivariate equalization has been extensively illustrated too [46,49,50]. The interesting novelty here is that this behavior, convenient for invariance, easily emerged in networks trained for segmentation because we used the layer with the appropriate analytical expression (or the appropriate capacity).

6. Conclusion

We have shown that introducing the Divisive Normalization layer in segmentation models helps them to achieve better results in all the tested scenarios, not only in a range of natural and synthetic databases, but also when we systematically extended the changes over five extra visual dimensions: luminance, achromatic contrast, chromatic contrast, spectral illumination (i.e. hue, and saturation). These improvements are not only reflected in higher overall IoU scores, but also in a deeper understanding of when and why DN helps. Our results show that the use of Divisive Normalization becomes more important in extreme scenarios, such as low luminances (night images) or low contrast (high fog). Crucially, we found that these gains are not just quantitative but also qualitative: DN improves segmentation by increasing the invariance of the model's output under environmental changes. That is, networks with DN layers produce more stable predictions across input variations, leading to better generalization in real-world conditions. These effects are visually summarized in Fig. 15 which shows how the model with Divisive Normalization gets similar segmentation results when environment changes are applied to the image, while the segmentation produced by the classical U-Net is clearly affected by these changes.

More detailed, panel 15(a) shows predictions when changing the illuminant by increasing the saturation. When the saturation is high, the car (segmented in blue) disappears from the no-DN prediction. Panel 15(b) shows results for different levels of fog, this variation also affects the no-DN model by stopping to detect a car even for the middle fog level. Panels 15(c) and 15(d) show how the prediction changes when the luminance and achromatic contrast are increased or reduced. In agreement with results obtained in Section 4.2.2, there are higher gains due to the Divisive Normalization in the low luminances and low contrast. When the luminance decreases the no-DN model does not detect the truck while the 4-DN model still detects part of it. The same happens when the contrast gets reduced, the no-DN model completely mismatches the car, labeling it as vegetation.

Despite the valuable insights provided, this study has some limitations that should be considered. First, regarding the modification of the scenes to control illumination, a more physically accurate simulation could be achieved using virtual reality tools. However, these methods

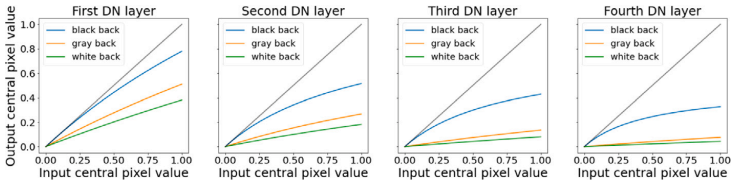
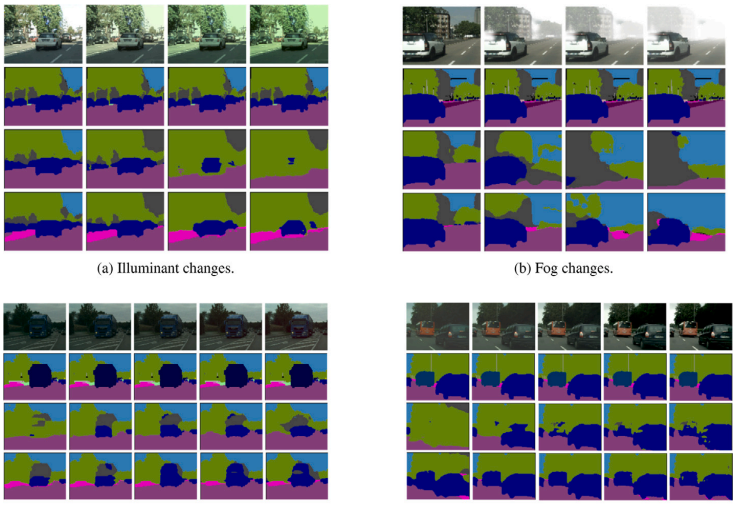


Fig. 14. Nonlinearities at different depths. Effect of the four Divisive Normalization model layers depending on the pixel value and its surround, showing the adaptation of the normalization.



(a) Illuminant changes.

(b) Fog changes.

(c) Luminance changes.

(d) Achromatic contrast changes.

Fig. 15. Summary: dealing with diversity. Example of the models' predictions under the different controlled changes. In each panel, the first row shows the input images; the second row shows the segmentation ground truth; and the third and fourth rows show the predictions of the no-DN and 4-DN models respectively. Panel 15(a) and 15(b) show the predictions when the saturation of an illuminant and the fog level increases (left to right). Panel 15(c) and 15(d) show the predictions when the image mean luminance and achromatic contrast are reduced (from the central image to the left) or increased (from the central image to the right).

are extremely computationally expensive and would still rely on non-photorealistic approximations. In our case, assuming that the original images originate from certain spectral reflectances under a given spectral illumination is sufficient to generate visually acceptable and useful images for evaluating segmentation robustness. Second, while the models used in our experiments are not extremely large (they contain only a few million parameters), they are adequate to demonstrate the effectiveness of Divisive Normalization (DN) in improving generalization under environmental variability. Additionally, our experiments focus on the U-Net architecture. This was a deliberate choice to isolate the effect of DN in a well-understood, widely-used model. Nevertheless, testing DN within a broader set of architectures will further support our results. Similarly, our evaluation was in the autonomous driving domain, which, while rich in environmental diversity, does not cover all types of segmentation tasks.

Regarding future work, we plan to extend our evaluation to include other segmentation architectures (e.g., transformer-based or multi-scale models) to further assess the generality of DN. We also intend to apply DN in larger models to test its scalability and effectiveness in more complex learning setups. In terms of datasets, we aim to explore scene types beyond autonomous driving, including indoor and natural environments, to evaluate generalization across domains. Furthermore, while this study focused on illumination, contrast, and color variability, we are interested in testing DN's benefits under other natural distortions, such as snow, rain, and motion blur. Finally, we plan to develop a parametric version of the Divisive Normalization layer, enabling the use of larger fixed-form kernels with fewer parameters and greater flexibility.

To conclude, our results show that including the Divisive Normalization in the segmentation algorithm makes them invariant under many image changes, which is helpful for example for autonomous driving.

CRediT authorship contribution statement

Pablo Hernández-Cámara: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jorge Vila-Tomás:** Writing – review & editing, Writing – original draft, Software, Data curation. **Paula Dauden-Oliver:** Writing – review & editing, Writing – original draft, Visualization, Software, Data curation. **Nuria Alabau-Bosque:** Writing – review & editing, Writing – original draft, Software, Methodology, Data curation. **Valero Laparra:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Jesús Malo:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by MICIIN/FEDER/UE under Grants PID2020-118071GB-I00, PDC2021-121522-C21 (funded by MCIN/AEI/10.13039/501100011033 and the EU NextGenerationEU/PRTR) and Grant PID2023-152133NB-I00; in part by Spanish MIU under Grant FPU21/02256; and in part by Generalitat Valenciana, Spain under Projects GV/2021/074, CIPROM/2021/056 and CIAPOT/2021/9. The authors gratefully acknowledge the computer resources at Artemisa and the technical support provided by the European Union through the 2014–2020 ERDF Operative Programme of Comunitat Valenciana, project IDIFEDER/2018/048.

Data availability

The datasets used during the current study are available from the following links:

(1) IPL-CARLA-dataset: https://huggingface.co/datasets/isp-uv-es/IPL-CARLA-dataset.

(2) IPL-Cityscapes-LuminanceContrasts : https://huggingface.co/datasets/isp-uv-es/IPL-Cityscapes-LuminanceContrasts.

(3) IPL-Cityscapes-Illuminants: https://huggingface.co/datasets/isp-uv-es/IPL-Cityscapes-Illuminants.

References

[1] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Adv. Neur. Inf. Proc. Syst., vol. 25, 2012.
[2] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, http://www.deeplearningbook.org.
[3] D.J. Heeger, Normalization of cell responses in cat striate cortex, Visual Neurosci. 9 (2) (1992) 181–197.
[4] M. Carandini, D.J. Heeger, Normalization as a canonical neural computation, Nat. Rev. Neurosci. 13 (1) (2012) 51–62.
[5] J. Malo, I. Epifanio, R. Navarro, E. Simoncelli, Nonlinear image representation for efficient perceptual coding, IEEE Trans. Images Proc. 15 (1) (2006) 68–80.
[6] J. Ballé, V. Laparra, E.P. Simoncelli, End-to-end optimized image compression, in: Int'L Conf on Learning Representations, ICLR, 2017.
[7] V. Laparra, J. Muñoz-Marí, J. Malo, Divisive normalization image quality metric revisited, J. Opt. Soc. Amer. A 27 (4) (2010) 852–864.
[8] A. Hepburn, et al., Perceptnet: A human visual system inspired neural network for estimating perceptual distance, in: ICIP, 2020, pp. 121–125.
[9] A. Ortiz, et al., Local context normalization: Revisiting local normalization, in: CVPR, 2020.
[10] P. Hernández-Cámara, J. Vila-Tomás, V. Laparra, J. Malo, Neural networks with divisive normalization for image segmentation, Pattern Recognit. Lett. 173 (2023) 64–71.
[11] R. Burnell, et al., Rethink reporting of evaluation results in AI, Science 380 (6641) (2023) 136–138.
[12] L. Zhou, et al., Predictable artificial intelligence, 2023, arXiv:2310.06167.
[13] M. Martinez, M. Bertalmío, J. Malo, In praise of artifice reloaded: Caution with natural image databases in modeling vision, Front. Neurosci. 13 (2019).
[14] M.D. Fairchild, The HDR photographic survey, in: Color and Imaging Conference, vol. 15, 2007, pp. 233–238.
[15] M. Fairchild, Color Appearance Models, John Wiley & Sons, 2013.
[16] C. Sakaridis, et al., Semantic foggy scene understanding with synthetic data, Int. J. Comput. Vis. 126 (9) (2018) 973–992.
[17] X. Hu, C.-W. Fu, L. Zhu, P.-A. Heng, Depth-attentional features for single-image rain removal, in: 2019 CVPR, 2019, pp. 8014–8023.
[18] S. Jiménez, J. Malo, The role of spatial information in disentangling the irradiance–reflectance–transmittance ambiguity, IEEE Trans. Geosci. Remote Sens. 52 (8) (2013) 4881–4894.
[19] R. Casati, P. Cavanagh, The Visual World of Shadows, MIT Press, 2023.
[20] V. Laparra, S. Jiménez, et al., Nonlinearities and adaptation of color vision from sequential principal curves analysis, Neural Comp. 24 (10) (2012) 2751–2788.
[21] R. Deeb, et al., Interreflections in computer vision: a survey and an introduction to spectral infinite-bounce model, J. Math. Imag. Vis. 60 (2018) 661–680.
[22] A. Hawkins, Tesla's autopilot and full self-driving linked to hundreds of crashes, dozens of deaths, 2024, https://www.theverge.com/2024/4/26/24141361/tesla-autopilot-fsd-nhtsa-investigation-report-crash-death. (Accessed 08 July 2024).
[23] F. Saleh, M. Aliakbarian, et al., Effective use of synthetic data for urban scene semantic segmentation, in: 2018 ECCV, 2018, pp. 84–100.
[24] G. Ros, L. Sellart, et al., The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes, in: 2016 CVPR, 2016, pp. 3234–3243.
[25] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: MICCAI, 2015, pp. 234–241.
[26] M. Cordts, et al., The cityscapes dataset for semantic urban scene understanding, in: 2016 CVPR, 2016.
[27] D. Dai, L. Van Gool, Dark model adaptation: Semantic image segmentation from daytime to nighttime, in: IEEE International Conference on Intelligent Transportation Systems, 2018.
[28] S.R. Richter, V. Vineet, S. Roth, V. Koltun, Playing for data: Ground truth from computer games, in: Eur. Conf. Comp. Vis., 2016, pp. 102–118.
[29] A. Dosovitskiy, G. Ros, et al., CARLA: An open urban driving simulator, in: Proceedings of the 1st Annual Conference on Robot Learning, 2017, pp. 1–16.
[30] P. Wang, P. Wang, et al., Mask-DerainGAN: Learning to remove rain streaks by learning to generate rainy images, Pattern Recognit. 156 (2024) 110840.
[31] Y. Wen, T. Gao, et al., Restoring vision in rain-by-snow weather with simple attention-based sampling cross-hierarchy transformer, Pattern Recognit. 156 (2024) 110743.
[32] P. Chattopadhyay, K. Sarangmath, et al., PASTA: Proportional amplitude spectrum training augmentation for syn-to-real domain generalization, in: International Conference on Computer Vision, ICCV, 2023, pp. 19288–19300.
[33] S. Kim, D.-h. Kim, H. Kim, Texture learning domain randomization for domain generalized segmentation, in: International Conference on Computer Vision, ICCV, 2023, pp. 677–687.
[34] N. Kim, T. Son, et al., WEDGE: Web-image assisted domain generalization for semantic segmentation, in: International Conference on Robotics and Automation, ICRA, 2023, pp. 9281–9288.
[35] Q. Sun, P. Melnyk, et al., Augment features beyond color for domain generalized segmentation, 2023, arXiv preprint arXiv:2307.01703.
[36] C. Hümmer, M. Schwonberg, et al., VLTSeg: Simple transfer of CLIP-based vision-language representations for domain generalized semantic segmentation, 2023, arXiv preprint arXiv:2312.02021.
[37] M. Fahes, T.-H. Vu, et al., PØDA: Prompt-driven zero-shot domain adaptation, in: International Conference on Computer Vision, ICCV, 2023.
[38] M. Fahes, T.-H. Vu, et al., A simple recipe for language-guided domain generalized segmentation, in: Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 23428–23437.
[39] R. Azad, et al., Medical image segmentation review: The success of U-Net, IEEE Trans. Patt. Anal. Mach. Intell. 46 (12) (2024) 10076–10095, http://dx.doi.org/10.1109/TPAMI.2024.3435571.
[40] IPL-CARLA-dataset, 2024, https://huggingface.co/datasets/isp-uv-es/IPL-CARLA-dataset. (Accessed 10 July 2024).
[41] E. Peli, Contrast in complex images, J. Opt. Soc. Amer. A 7 (10) (1990) 2032–2040.
[42] IPL-cityscapes-LuminanceContrasts, 2024, https://huggingface.co/datasets/isp-uv-es/IPL-Cityscapes-LuminanceContrasts. (Accessed 10 July 2024).
[43] J. Malo, M. Luque, ColorLab: A matlab toolbox for color science and calibrated color image processing, Univ. Val. (2002) http://isp.uv.es/code/visioncolor/colorlab.html.
[44] IPL-cityscapes-illuminants, 2024, https://huggingface.co/datasets/isp-uv-es/IPL-Cityscapes-Illuminants. (Accessed 10 July 2024).
[45] P. Hernández-Cámara, P. Daudén-Oliver, V. Laparra, J. Malo, Alignment of color discrimination in humans and image segmentation networks, Front. Psychol. 15 (2024).
[46] J. Malo, V. Laparra, Psychophysically tuned divisive normalization approximately factorizes the PDF of natural images, Neural Comput. 22 (12) (2010) 3179–3206.
[47] A.B. Abrams, J.M. Hillis, D.H. Brainard, The relation between color discrimination and color constancy: When is optimal adaptation task dependent? Neural Comput. 19 (10) (2007) 2610–2637.
[48] A.B. Watson, J.A. Solomon, Model of visual contrast gain control and pattern masking, J. Opt. Soc. Amer. A 14 (9) (1997) 2379–2391.
[49] M. Martinez-Garcia, P. Cyriac, T. Batard, M. Bertalmío, J. Malo, Derivatives and inverse of cascaded linear+nonlinear neural models, PLoS One 13 (2018) 1–49.
[50] J. Malo, Spatio-chromatic information available from different neural layers via Gaussianization, J. Math. Neurosci. 10 (1) (2020) 1–40.

Pablo Hernández-Cámara received the B.Sc. degree in Physics from University of Valencia, Spain, in 2020 and the M.Sc. degree in Data Science from University of Valencia, Spain, in 2022.He is currently pursuing the Ph.D. degree in Computational Neuroscience and Machine Learning with the University of Valencia, Spain. His research interests include bio-inspired models, deep learning and computational neuroscience.

Jorge Vila-Tomás obtained his degree in Physics in the University of Valencia. He then went to pursue a M.Sc. in Artificial Intelligence and another one in Data Science. His current interests comprise computer vision, image statistics and deep neural networks' training dynamics, and he is currently doing a Ph.D. in the intersection between computational neuroscience and computer vision in the Image Processing Laboratory in Valencia.

Paula Daudén-Oliver is a researcher in Image Processing Laboratory, University of València. She studied a degree and a M.Sc. in Optics. She is interested in comparing machine learning vision models with human vision doing human experimentation and analysis of models in classic psychophysics.

Nuria Alabau-Bosque has a degree in Computer Science from the University of Valencia. She then studied an MSC in Computer Engineering and Mathematics at the Rovira i Virgili University. She is currently doing her Ph.D. at the Image Processing Laboratory of the University of Valencia. His research interests include computer vision, image processing and machine learning. His current work focuses on the inclusion of invariance in image quality models and classification models.

Valero Laparra was born in València, Spain, in 1983. He received the B.Sc. degree in telecommunications engineering and the B.Sc. degree in electronics engineering from Universitat de València, València, in 2005 and 2007, respectively, the B.Sc. degree in mathematics from Universidad Nacional de Educación a Distancia, Madrid, Spain, in 2010, and the Ph.D. degree in computer science and mathematics from Universitat de València in 2011. He is currently an Assistant Professor with Escola Tècnica Superior d'Enginyeria, Universitat de València, where he is also a Researcher with the Image Processing Laboratory.

Jesús Malo received the M.Sc. and Ph.D. in Physics in 1995 and 1999 both from the Universitat de València. He received the Vistakon European Research Award in 1994 for his work on physiological optics. He worked on computational neuroscience during his postdoc stays at NASA Ames and NYU (2000–01) and as a visiting professor at Stanford and NYU (2013). He served as Editor of the IEEE Trans. Im. Proc., PLoS ONE, and Front. Neurosci, and has been on the committees of NeurIPS, ICLR, and the CIE. He was the first male member of the Asociación de Mujeres Investigadoras y Tecnólogas (AMIT). Currently, he is a Professor of Vision Science and a member of the Image and Signal Processing Group at the Universitat de València. His interests include (but are not limited to) models of low-level human vision, their relations with information theory, their applications to image processing, vision science experimentation, and beauty in general.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 23 October 2024
DOI 10.3389/fpsyg.2024.1415958

OPEN ACCESS

EDITED BY
Koen V. Haak,
Tilburg University, Netherlands

REVIEWED BY
Andrew Coia,
Science Applications International Corporation, United States
Eric Postma,
Tilburg University, Netherlands

*CORRESPONDENCE
Pablo Hernández-Cámara
pablo.hernandez-camara@uv.es

RECEIVED 11 April 2024
ACCEPTED 08 October 2024
PUBLISHED 23 October 2024

CITATION
Hernández-Cámara P, Daudén-Oliver P, Laparra V and Malo J (2024) Alignment of color discrimination in humans and image segmentation networks.
Front. Psychol. 15:1415958.
doi: 10.3389/fpsyg.2024.1415958

Alignment of color discrimination in humans and image segmentation networks

Pablo Hernández-Cámara*, Paula Daudén-Oliver, Valero Laparra and Jesús Malo

Image Processing Lab, Parc Científic, Universitat de València, València, Spain

The experiments allowed by current machine learning models imply a revival of the debate on the causes of specific trends of human visual psychophysics. Machine learning facilitates the exploration of the effect of specific visual goals (such as image segmentation) by different neural architectures in different statistical environments in an unprecedented manner. In this way, (1) the principles behind psychophysical facts such as the non-Euclidean nature of human color discrimination and (2) the emergence of human-like behaviour in artificial systems can be explored under a new light. In this work, we show for the first time that the *tolerance* or *invariance* of image segmentation networks for natural images under changes of illuminant in the color space (a sort of insensitivity region around the *white*) is an *ellipsoid* oriented similarly to a (human) MacAdam ellipse. This striking similarity between an artificial system and human vision motivates a set of experiments checking the relevance of the statistical environment on the emergence of such insensitivity regions. Results suggest, that in this case, the statistics of the environment may be more relevant than the architecture selected to perform the image segmentation.

KEYWORDS

vision models, color discrimination, image segmentation, artificial neural networks, U-Nets, image statistics, chromatic adaptation, Divisive Normalization

1 Introduction

Natural images and principled explanations in vision science. A long-standing hypothesis in vision science assumes that sensory behaviour derives from an evolutionary adaptation to the regularities of the environment (Barlow, 1959, 2001). This hypothesis is statistical in spirit because it assumes that certain architecture (network of sensors and neurons) is progressively updated to become optimal according to certain task for the inputs faced by the system (Richards et al., 2019). As a result, the concept of *natural images* has become central in this kind of principled explanation (Field, 1987; Simoncelli and Olshausen, 2001; Torralba and Oliva, 2003; Hyvärinen et al., 2009), because it refers to stimuli (e.g., photographic images) which are representative of certain visual environments and constitute the training set for the system.

Linear statistical models in color vision. The link between color vision and the statistics of the natural environment has a long and fruitful history. Classical approaches often employ linear models to explain different aspects of color vision. For instance, one seminal study derived opponent color channels from the statistics of color samples (Buchsbaum and Gottschalk, 1983): authors assumed that the goal of the color sensors is to decorrelate the neural responses after the photoreceptors so they computed the linear Principal Component Analysis (PCA) of color samples in natural images. PCA

transforms data into a set of linearly uncorrelated components (Jolliffe, 2002), identifying the directions (principal components) in which the data varies the most. It turns out that the best directions to encode natural colors are the *luminance*, the *red-green*, and the *yellow-blue* directions. This statistical explanation of the physiological achromatic and opponent channels (Shapley and Hawken, 2011) is more conclusive than classical hue cancellation experiments (Vila-Tomás et al., 2023). This is because the opponent spectral sensitivities obtained from PCA are more similar to the final sensitivities in multi-stage models such as (DeValois and DeValois, 1993), while hue cancellation results are mainly determined by the experimental choice of the cancellation stimuli (Vila-Tomás et al., 2023). Another notable linear approach involved the derivation of chromatic Contrast Sensitivity Functions (CSFs) through linear filters designed to maximize information transmission (Atick and Redlich, 1992; Atick et al., 1992). In this case the authors also assumed a decorrelation goal but in the presence of retinal noise. The optimal filters amplify certain spatial frequencies to whiten the responses (to make their spectrum flat) while attenuating the spatial frequencies where the noise is bigger than the typical signal. Additionally, explanations of chromatic adaptation, a process by which the visual system adjusts to changes in the lighting conditions, have been based on linear shifts in the average and covariance of color samples (Webster and Mollon, 1997; Clifford et al., 2007). The average represents the mean color value, while the covariance indicates how color values vary together, providing insights into the overall color distribution in the visual scene. Adaptation is understood as a transform to an invariant inner representation that compensates for the color shifts induced by changes in the environment (illumination, shadows, etc.). In summary, linear statistical models have identified opponent chromatic channels, the frequency bandwidth for achromatic and opponent chromatic patterns, and adaptation mechanisms based on the mean and the covariance of the chromatic signals.

Nonlinear statistical models in color vision. More recently, nonlinear descriptions of color statistics have been used to reproduce the nonuniform resolution and adaptation of the response of opponent mechanisms. In particular von der Twer and MacLeod (2001); MacLeod and von der Twer (2003) suggested that the nonlinear behaviour of opponent channels could be explained by using univariate Cumulative Density Functions (CDFs) of color samples. The CDF transforms the input probability into a uniform probability. This means that if the sensor responses are related to the CDF, simple uniform resolution in the response domain minimizes the error introduced in the representation of the signal (Lloyd, 1982). This philosophy was further extended to other optimization principles and higher-dimensional scenarios using Sequential Principal Curves Analysis (SPCA), a statistical method that generalizes PCA by fitting smooth curves through the data allowing for the representation of nonlinear structures. The different nonlinearities that can be accommodated in SPCA (Laparra and Malo, 2016) extend the cumulative density approach from optimal error minimization (Lloyd, 1982) to optimal information maximization (Laughlin, 1983). In this way, new explanations of color adaptation, color constancy and color illusions were proposed (Laparra et al., 2012; Laparra and Malo, 2015).

Signal statistics and model architecture. By definition, nonlinear models are more accurate and general than linear models. However, the above nonlinear descriptions of color phenomena were more focused on the statistics of the color signals rather than on the architecture, i.e., they oversaw the specific network required for the implementation of the computations. In general, the interactions between the statistical goal and architecture are not trivial (Poggio, 2021; Hernández-Cámara et al., 2023a; Hernández-Cámara et al., 2024). For example, different deep-learning architectures trained according to the same statistical goal may lead to critically different behaviours. This has been the case in studying color illusions (Gomez-Villa et al., 2020), or chromatic contrast sensitivity, either from low-level (Li et al., 2022), or higher-level principles (Akbarinia et al., 2023). In these studies authors show that *for the same functional goal* deeper networks may get better performance in the goal, but they display less-human behaviour than shallow networks (in terms of bandwidth or visual illusions).

Open issues in statistical explanations of color discrimination. The metric of the tristimulus space is not Euclidean, for instance, the discrimination region around the white has a specific asymmetry and orientation (MacAdam, 1942). Current statistical explanations of that fact are based on very low-level principles: error-minimization or information-maximization using SPCA (Laparra et al., 2012) or Gaussianization techniques (Jiménez et al., 2013), or the techniques based on Fisher information (da Fonseca and Samengo, 2016, 2018) which is another form of information maximization. Neither of these explanations take the architecture of the system into account (they only describe the properties of color distributions), and the principles are so low-level that are not directly connected to actual visual tasks.

Questions addressed in this work:

- *Is it possible to derive basic properties of human color discrimination ellipses from visual tasks of higher-level than error-minimization or information-maximization?* Particularly [as opposed to the cited low-level literature (Laparra et al., 2012; Jiménez et al., 2013; da Fonseca and Samengo, 2016, 2018)] by explicitly optimizing a neural architecture with certain resemblances to the retina-cortex pathway.
- In solving the considered higher-level visual task, *what is the relative relevance of the color statistics of the environment versus the consideration of reasonable variants in the network architecture?*

In this work, we address these questions using networks trained to perform image semantic segmentation (Guo et al., 2018), which is a mid-level vision task that consists of identifying the objects in the input images by classifying each pixel into one semantic category. We implement this task using variants of the successful U-net architecture (Ronneberger et al., 2015). The encoding part of this architecture is a cascade of linear-nonlinear stages which displays certain resemblances (in connectivity and function) with early vision (Jacob et al., 2021). Moreover, we augment the conventional U-net by including biologically-inspired layers, the so-called *Divisive Normalziation* (DN) (Hernández-Cámara et al., 2023b). This DN layer is a canonical non-linearity in sensory

neuroscience (Carandini and Heeger, 2012) that takes into account the inhibitory effect of neighbour neurons and explains chromatic adaptation too (Abrams et al., 2007; Hillis and Brainard, 2005). Finally, to check the relevance of the statistics of the environment, we conduct the training and testing of the networks with different kinds of images with distinctly different color statistics.

The idea is to check if human-like tolerance regions to color changes emerge in these networks tuned to solve semantic image segmentation. And, if they sometimes do, does it depend more on the statistics of the environment or on the variants introduced in the architecture?

In this work, we report the following finding: the region of invariance to changes in illumination in image segmentation networks trained with naturally illuminated images is similar to the region of insensitivity (or invariance) to color changes in humans (the MacAdam ellipse around the white). Therefore, this mid-level task may be an alternative to previous lower-level explanations. However, we find that the statistics of the colors in the environment are more relevant to explain color discrimination than the considered variants in the architecture in the segmentation network.

2 Materials and methods

Here we introduce the six methodological elements required for our experiments: (1) various distinct chromatic environments for the segmentation goal: a naturally illuminated scenario (regular photographic scenes with daylight illumination), and then two counter-examples selected to have quite different color statistics (submarine images and achromatic images respectively). Then, (2) we outline the methodology we follow to compare the tolerance to color changes in artificial networks and in humans. As this general methodology implies generating consistent color shifts in scenes annotated for segmentation, (3) we select one of the possible approximated ways to introduce such color shifts, namely the variation of a simulated spectral illumination. Then, (4) we illustrate the shape of the tolerance of humans to color shifts around the white color (or the anisotropy of that MacAdam ellipse), (5) we present the scenes with shifted colors to check the tolerance of the networks, and finally, (6) we present details of the neural architectures of the considered image segmentation networks.

2.1 Environments of different statistics

The analysis of color discrimination of different image segmentation networks requires training these artificial systems in visual environments with substantially different color statistics. The idea is checking if differences in color statistics induce consistent changes in color discrimination. To this end, we considered three datasets with known segmentation ground truth, but distinct scene statistics: *Cityscapes* (Cordts et al., 2016),[1] *SUIM* (Islam et al., 2020),[2] and *Oxford-IIIT Pets* (Parkhi et al., 2012).[3] While *Cityscapes* consists of a range of urban photographic scenes under natural illumination (see Figure 1), the other two environments are counter-examples specifically selected to have distinct color statistics. On the one hand, *SUIM* is shifted to blue because it consists of underwater pictures. On the other hand, while *Pets* also has natural daylight illumination, we intentionally removed all chromatic information by changing the images to gray-scale so that the segmentation has to be based on alternative (non-chromatic) visual features such as shape or texture. The term "natural scenes" is applicable to the "urban scenes" in CityScapes, because "natural" refers more to the low-level statistical features of the images (smoothness, edge consistency and continuity, or day-light illumination) rather than to the presence of natural versus man-made objects. In fact, similarly to the achromatic literature (Field, 1987; Olshausen and Field, 1996), when dealing with spatio-chromatic scenes Gabor-like sensors in chromatically-opponent channels emerge both in forest-landscape scenes and in urban scenes (Doi et al., 2003; Gutmann et al., 2014).

The color statistics of these environments are illustrated by the scatter plots of color samples in the 1931 CIE xy chromatic diagrams in Figure 1. In these diagrams, the spectral locus and the triangle defined by the red, green and blue primaries of regular displays have been plotted for useful reference. For each environment, we took the 1000 nearest neighbours to the average chromaticity and computed the local Principal Component Analysis (local PCA) as in Laparra et al. (2012) and Laparra and Malo (2015). The local principal components (in green and blue) and ellipses (in orange) associated with the local covariance matrices from the local PCA highlight the difference in color statistics. Therefore, systems trained for information maximization or error minimization in these environments should have different metrics when considering color differences. Of course, nothing can be said for systems trained in (artificially) achromatic environments.

A technical note on the color of the databases. The images in the considered databases are expressed in digital values. This device-dependent color representation is transformed into standard 1931 CIE XYZ tristimulus vectors assuming a standard display calibration (Hunt, 2005; Malo and Luque, 2002). To simplify the implementation of the experiments involving changes of illumination in the following sections, we reduced by a factor 0.75 the excitation purity of all the colors in *Cityscapes* and *SUIM*. This can be easily seen in the sharp edge in the cyan colors of the underwater environment. Incidentally, this sharp edge suggests that camera recordings in this region are already saturated in the blue channel. This bias does not represent a problem for our study because this is just a counter-example with substantially non-natural statistics. We applied this small reduction because changes in spectral illumination imply movements of the color manifold towards the limits of the color gamut that can be properly represented in digital systems (the triangle in dotted style). This reduction in the original saturation allows stronger changes in the illumination. Nevertheless, it is important to note that this does not change the relative shape of the color distributions (does not change

1 https://www.cityscapes-dataset.com/

2 https://irvlab.cs.umn.edu/resources/suim-dataset

3 https://www.robots.ox.ac.uk/~vgg/data/pets/

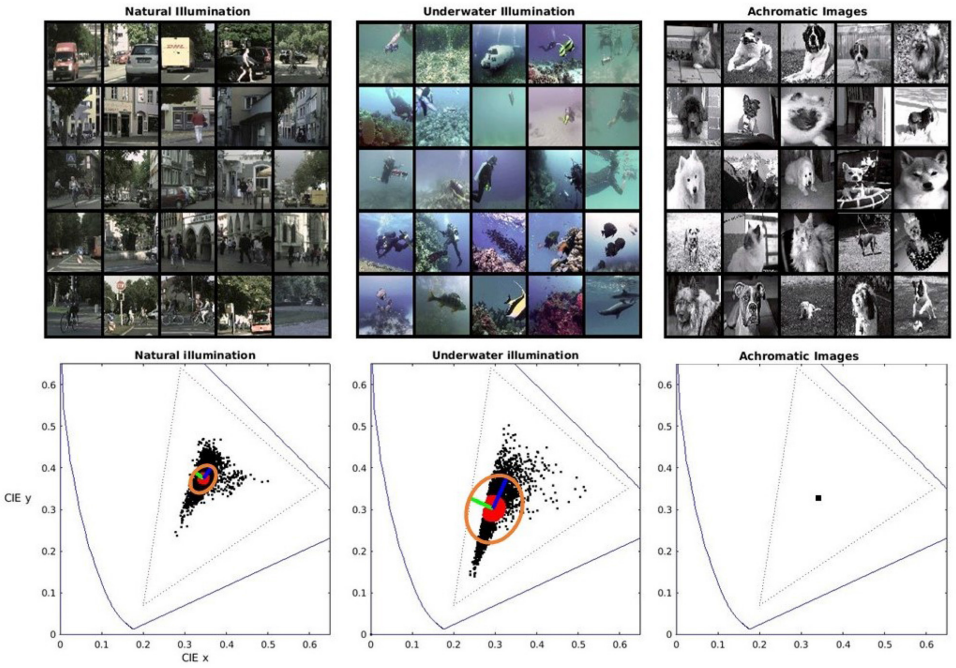


FIGURE 1
Different environments **(top)** and associated color statistics **(bottom)**. *Left:* daylight natural illumination. *Center:* daylight+underwater filtering and scattering. *Right:* artificial achromatic scenes (flat spectral reflectances and equienergetic illuminant). The corresponding 1931 CIE xy diagrams show representative color samples from the scenes (in black) and the closest neighbours to the average chromaticity (in red). It also shows local principal components (in green and blue) and the associated ellipsoid (in orange) computed from the local Principal Component Analysis (local PCA) of the nearest neighbours to the average chromaticity.

the orientation of the covariance matrices nor its relative size) and then, it does not modify the generality of the results.

2.2 Comparing tolerance to color shifts in humans and in machines

Color discrimination in humans has been defined in different ways depending on the stimuli and experimental task done by the observers. For instance, the classical MacAdam results are based on the variability of color matching experiments with flat patches of light sources (MacAdam, 1942; Wyszecki and Stiles, 2000). The covariance of this variability leads to the well-known ellipses in the 1931 CIE xy diagram. However, detection thresholds of deviations in different chromatic directions using randomly textured stimuli (Barbur, 2004) leads to ellipses with the same shape and orientation but larger size, about a $\times 5$ factor in size (Jennings and Barbur, 2010). Similar detection thresholds measured with natural images under controlled changes in illumination (Alabau-Bosque et al., 2024) are compatible with the results by Jennings and Barbur (2010).

All these descriptions are qualitatively equivalent: the relevant facts (Stockman and Brainard, 2010; Wyszecki and Stiles, 2000) that should be reproduced by the models is that the higher sensitivity (lower threshold and equivalently lower variability of the color matches) is observed in a specific *red-green* direction while the lower sensitivity (bigger threshold and bigger variability) is in an almost orthogonal *yellow-blue* direction. In Section 2.5 we visually illustrate that these are really robust trends.

In the case of artificial networks we will use the concept of *tolerance region*. Note that the performance of the neural net in the visual task (in this case segmentation) has a certain value given that the images are illuminated as in the training conditions (and hence the test images have the same texture and color statistics). However, if the images are consistently color-shifted (for instance by changes in the spectrum of the light source) the performance will drop. If the network is able to cope with the color-shift with a negligible drop in performance one can say that the network is insensitive (or tolerant) to that color-shift. Setting an arbitrary threshold on the network performance one may define a *tolerance region* in the color space so that performance drops less than this value. This tolerance region is a description of the insensitivity of the network

to color shifts, similarly to the MacAdam ellipses for humans. Obviously, tolerance regions in humans (as classically defined) and in machines (as defined here) are not identical concepts, but a convenient analogy to compare their behaviours.

Will these (artificial) tolerance regions have something in common with the human insensitivity (MacAdam ellipse) region around the white? Alignment between these two concepts would suggest a common explanation of both behaviours.

In order to check the above in different training environments one must: (1) train the considered networks for the task in the different environments and (2) test the tolerance of those networks in scenes where color has changed in a consistent form (that can be systematically represented in the chromatic diagram).

In the next subsections we discuss how to introduce systematic color changes in photographic images via simulated changes in spectral illumination and how this can be used to illustrate human color discrimination around the white.

2.3 Systematic color-shifts via changes in spectral illumination

To test the tolerance of the segmentation networks to color shifts in a meaningful way, one should use convenient ways to generate systematic, chromatically-controlled and consistent changes in the images of the different environments so that the networks face new (equivalent and controlled) situations not considered in the training.

The required color shifts in the test sets can be introduced in different ways. In the context of color constancy, different approaches have been used to model color changes in the images. These different approaches represent different degrees of approximation to the physics of image generation. The approximations differ on how well the geometry of rendering and the spectrum-to-tristimulus transforms are taken into account. Approaches to include consistent color-shifts which are progressively closer to the physics of image generation include:

1. Following simple models of illumination compensation (Finlayson et al., 1993; Chong et al., 2007), one should express the color of the images in certain tristimulus space and introduce independent linear variation in the tristimulus values. This is clearly better than naive operation in RGB digital counts, but the diagonal linear transform is still rather restrictive: the authors recommended this when the intrinsic dimensionality of spectral reflectance of surfaces and spectral radiance of the illuminant is as low as two or three.
2. Following general linear models of illumination compensation (Webster and Mollon, 1997; Clifford et al., 2007), one could apply a rotation and a scaling matrix to the tristimulus values. This transform is more general than the previous method based on diagonal matrices but still disregards the huge dimensionality reduction process that happens in the spectrum-to-tristimulus transform.
3. Virtual environments [such as CARLA simulator (Dosovitskiy et al., 2017)] are appealing to change the chromaticity of the illumination because they consider the 3-dimensional scene in the rendering. However, conventional programs usually make gross approximations from the colorimetry point of view: spectral distributions are not controllable and they usually operate in RGB digital counts. As a result, it is not obvious how to control the changes in the illumination to systematically sample chromatic directions to check discrimination in the 1931 CIE xy diagram.
4. A convenient alternative is assuming that the original images come from certain spectral reflectances under a given spectral illumination and recreating new images by applying the tristimulus equation assuming Lambertian surfaces with no mutual illumination. As opposed to methods that operate on tristimulus values, this method does take into account the huge dimensionality reduction in the spectrum-to-tristimulus transform, so illumination change is richer than a rotation+scaling in the tristimulus space. However, this method has also been criticized because it disregards the nonlinearities that come from mutual illumination (Laparra et al., 2012; Deeb et al., 2018).
5. Create annotated scenes for segmentation using the unconventional virtual reality tools that take into account both the geometry and the spectral content of light and reflectance of surfaces, as for instance (Heasly et al., 2014).
6. Take real scenes where the spectral illumination can be physically modified and measure (take pictures) using colorimetrically calibrated cameras (Laparra et al., 2012; Gutmann et al., 2014), or spectro-radiometrically calibrated cameras (Foster et al., 2016; Nascimento et al., 2016).

Of course, the best methods (5th and 6th) are not straight forward. The 6th case implies building a database from scratch (in case of having the expensive measurement equipment). Moreover the mentioned databases that include physical changes in the spectra are not good for our purposes because the spectral change is uncontrolled or does not properly sample the chromatic diagram. Moreover, they are not annotated for segmentation. In the 5th case, one would have to build virtual scenes from scratch and then use the internal (non-standard) code for the objects to derive the image segmentation maps. Therefore, methods 5 and 6 are too complicated for the illustrative test sets that we want to generate to check the invariance/tolerance of the segmentation networks. Then, between the next two methods (3rd and 4th, each with advantages and shortcomings) we chose the 4th method for its balance between complexity and colorimetric realism.

2.4 Human color discrimination illustrated via changes in the spectra

After the previous discussion about the different ways to introduce the color shift, here we describe in more detail the chosen option. Particularly, here we describe the change of tristimulus values of a surface of known spectral reflectance when we change the spectral illumination, and then we explicitly illustrate how uniform changes in hue and saturation over the chromatic diagram are not perceived uniformly. This

anisotropic tolerance to color shifts,[4] known as the MacAdam ellipses (Wyszecki and Stiles, 2000; MacAdam, 1942), is the human behaviour that we want to compare with the invariance region of the models.

Given an object of spectral reflectance, $\rho_\lambda \in [0, 1]$, illuminated by an illuminant with spectral radiance s_λ in $W/m^2 str$, its tristimulus values in certain color representation, T_i with $i = 1, 2, 3$, are given by Wyszecki and Stiles (2000):

$$T_i = k_m \int_{380}^{770} \rho_\lambda \, s_\lambda \, \bar{T}_i(\lambda) \, d\lambda \tag{1}$$

where $\bar{T}_i(\lambda)$ are the *color matching functions*, or the sensitivity of the color sensors in that representation, and $k_m = 683\, lm/W$, is the *luminous efficacy* constant. This implies that the *chromatic coordinates*, $t_i = T_i / \sum_i T_i$, also change with the illuminant.

Figure 2 shows the variation of the color appearance of a flat reflectance, $\rho_\lambda = 1 \; \forall\lambda$, when it is illuminated by a set of sources with spectral radiances, $s^\star_\lambda$, taken so that the color of the sample has the desired tristimulus vectors, $T^\star$, with chromatic coordinates represented in the 1931 CIE xy diagram at the left and a constant luminance of 35 cd/m^2. The spectral sources were computed via:

$$s^\star_\lambda = \arg\min_{s_\lambda} \left| T^\star - T\,(s_\lambda) \right|_2 \tag{2}$$

where $T\,(s_\lambda)$ was computed as in Equation 1. Metamerism means that Equation 2 is ill-posed (Wyszecki and Stiles, 2000). The algorithm we use[5] breaks the multiplicity of solutions by looking for the illuminant that minimizes the error in tristimulus values using an exhaustive search in a structured dataset of 20,000 spectral radiances/reflectances. The structure of this dataset (the way the spectral shapes are ordered) is based on the Munsell book of color. This guarantees that the considered spectra represent a perceptually uniform sampling of the color space. In this example the considered illuminants are organized as a function of *hue* and *saturation*, i.e., angle with respect to the x axis, and distance with respect to the central white point respectively.

The uniform distribution of color variations in a polar representation along the 1931 CIE xy diagram in Figure 2A illustrates the fact that that human color discrimination is not isotropic around the white, i.e., it is non-uniform. Note that when linearly increasing the saturation of the color along the different hue directions (going down along each column of the colored panel), the perception of *colorfulness* (Fairchild, 2013) is not uniform. See that, qualitatively and just for illustrative visualization, the circles in the colored panel define a boundary between clearly chromatic patches (below the curve) and mainly achromatic patches (above the curve). This human region of tolerance or invariance around the white can be plotted in the chromatic diagram (ellipse represented by the orange dots in Figure 2B) by using the corresponding cartesian to polar transform. As an example of this transformation, see for instance the position of the solid circle located in the colored panel Figure 2A (fourth hue and fifth saturation index) and its corresponding location in the chromatic diagram.

Figure 2 is just a compelling visual illustration of the anisotropy of human tolerance to color shifts: the tolerance is maximal in the *yellow-blue* direction and minimal in the *red-green* direction. Of course, this visualization is not an accurate measurement of the color discrimination ellipse (MacAdam, 1942; Jennings and Barbur, 2010; Alabau-Bosque et al., 2024). Interestingly, even though this visualization has all the limitations of color reproduction in displays (Hunt, 2005), the anisotropy of human tolerance to color shifts is so robust that the characteristic two-minima-shape of the achromatic-to-chromatic boundary is clearly visible. Note that the orientation of this qualitatively drawn boundary-and-ellipse is consistent with the classical experimental ellipses (MacAdam, 1942; Wyszecki and Stiles, 2000) depicted in Figure 2C.

We will be back to this two-minima shape in the hue-saturation plane and the associated ellipsoid when we present the results of the tolerance regions of the segmentation networks in Section 3.

2.5 Spectral illumination changes in the environments

Once the networks are trained in the considered environments, the scenes are modified to introduce changes in spectral illumination using the sources, $s^\star_\lambda$, shown in Figure 2. To do so, spectral reflectances have to be associated to each region of the original scenes. This association is done by assuming that the tristimulus vectors, T, in the original scenes (e.g., the chromaticities in Figure 1 with their corresponding luminances) come from the illumination of certain reflectances, $\rho^\star_\lambda$, with an equienergetic illuminant:

$$\rho^\star_\lambda = \arg\min_{\rho_\lambda} \left| \, T - k_m \int_{380}^{770} \rho_\lambda \, \bar{T}_i(\lambda) \, d\lambda \, \right|_2 \tag{3}$$

Again, the solution of the ill-posed Equation 3 was obtained through the function `tri2spec.m` of Colorlab (Malo and Luque, 2002) because the spectra in the Munsell database do a thorough sampling of the color space. Once each pixel has an associated reflectance, $\rho^\star_\lambda$, its new color, T', under the new illumination is computed with Equation 1 using $s^\star_\lambda$. Finally, the new 1931 CIE XYZ colors are transformed into digital values assuming a standard display calibration (Hunt, 2005; Malo and Luque, 2002). Figure 3 shows an example of the result of this procedure applied to one image of each of the three different training environments considered.

The saturation of the considered spectral sources was limited by the fact that we did not want the manifold of modified colors to lie outside the triangle of primary colors in a regular display.

These modified scenes can be used to test each of the image segmentation networks which were trained on the three different original scenes. The performance is expected to be similar for illuminants with small spectral contrast: the segmentation results for the scenes of the first row in Figure 3 will be similar to the performance on the original scenes. However, it is expected to change for illuminants of bigger saturation and different hues (down along the different columns).

4 Here color-shifts modelled as changes in the spectral reflectance or spectral radiance.

5 The function `tri2spec.m` of Colorlab (Malo and Luque, 2002).

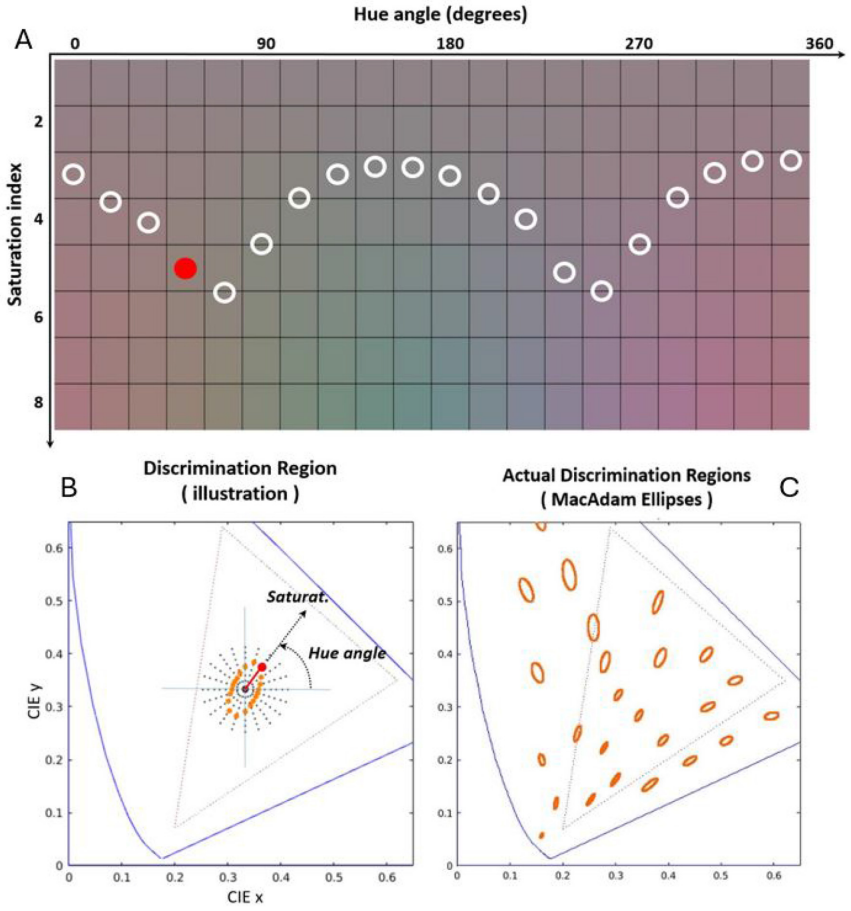


FIGURE 2
Illustration of human color discrimination: tolerance to saturation for different hues. **(A)** Shows patches of flat spectral reflectance illuminated by sources with spectral radiances selected to cover the 1931 CIE xy diagram as seen on a standard CRT display (Malo and Luque, 2002). Black dots in the CIE xy chromatic diagram of **(B)** show the polar distribution of the chromaticity of the considered illuminants. The illuminants are organized as a function of hue and saturation, i.e. angle with respect to the *x* axis, and distance with respect to the white point respectively. For each hue [each column in the colored panel of **(A)**] the Euclidean distance in the chromatic diagram required to induce certain perceptual departure from the white color of the same luminance is different. That is why the insensitivity region around the white [determined by the circles in **(A)**] is an ellipse with certain orientation [orange dots in **(B)**]. The diagram in **(C)** displays the insensitivity regions for humans measured by MacAdam (1942) at a number of color locations over the chromatic diagram.

2.6 Networks for image segmentation

In this work, we used U-Nets networks to perform image semantic segmentation following the state-of-the-art for this visual task (Ronneberger et al., 2015). See Figure 4 for an illustration of this architecture. In these networks, the input images (in digital values) go through a set of layers with progressively lower spatial resolution, i.e., the image dimensions decrease as the image passes through each block. Also, each block has a progressively higher number of features, i.e., different attributes detected by the network, such as different patterns, to capture more complex information up to the network bottleneck. From this inner representation, the signal is spatially expanded again up to the original resolution ending with a layer (in white in the figure) with a number of features equal to the number of distinct classes to be identified. Part of the high-resolution information is passed from the early

10.3389/fpsyg.2024.1415958

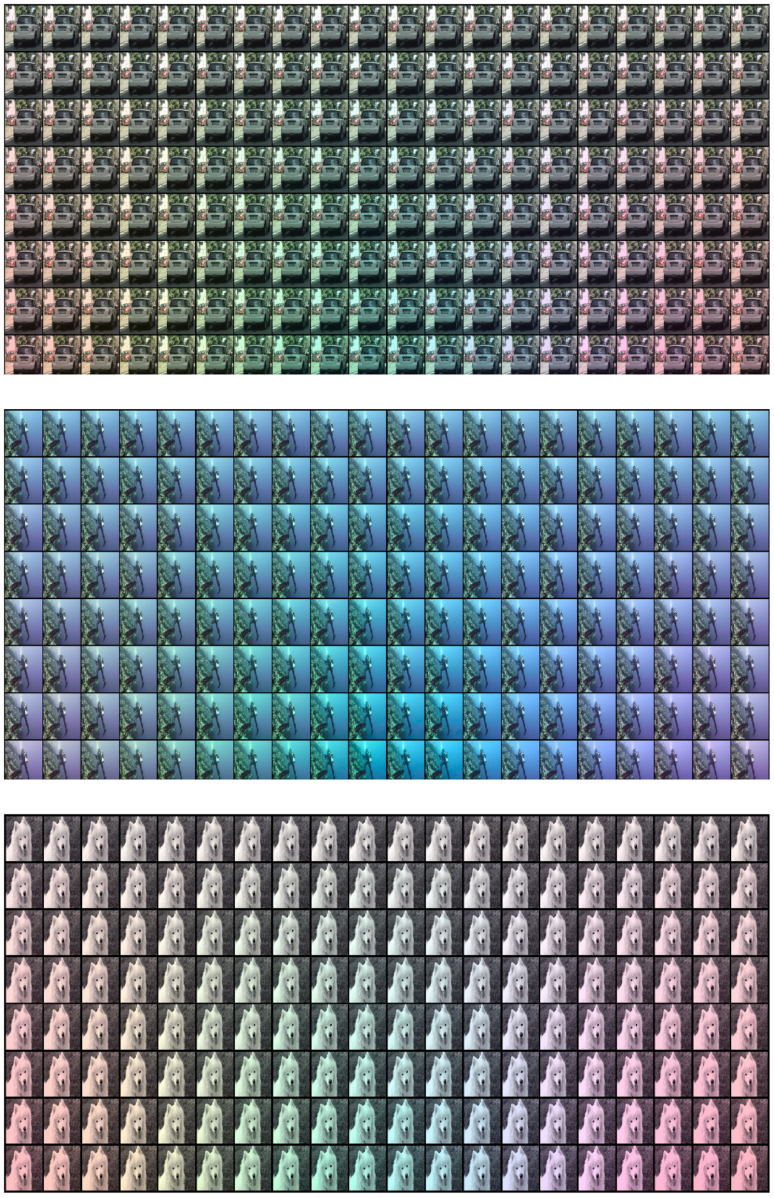


FIGURE 3
Scenes with modified illumination starting from a different original image: natural **(top)**, underwater **(center)**, and flat-reflectance, i.e., achromatic **(bottom)**.

layers to the late layers after the bottleneck through the so-called skip connections. The final layer performs the classification of each pixel to one of the possible classes in the dataset, i.e., assigned to the class that achieves higher response in this final layer. Note that this implies that this layer depends on the number of possible classes of the considered dataset and therefore a model trained in one dataset can not be applied to a different dataset with a different number of classes.

Apart from the standard U-nets, we considered the biologically inspired modification proposed in Hernández-Cámara et al. (2023b). This modified architecture considers *Divisive Normalization* (DN) layers in the encoding part of the U-Net (layers depicted in green in Figure 4). This nonlinear computation, $y = \mathcal{N}(x)$, is relevant because the response of each unit, x_i, is normalized by a pool of the responses of the neurons tuned to neighbour features:

$$y_i = sign(x_i)\,\frac{|x_i|}{b_i + \sum_j H_{ij}|x_j|} \tag{4}$$

and this normalization has proven to be important to explain both chromatic adaptation (Abrams et al., 2007; Hillis and Brainard, 2005; Fairchild, 2013) and contrast and texture adaptation (Watson and Solomon, 1997; Martinez et al., 2019). This previous literature on the benefits of Divisive Normalization for adaptation suggests that U-Nets with Divisive Normalization may be more tolerant to changes in illumination, and their insensitivity regions may be more similar to those of humans.

3 Experiments and results

In this section, the considered networks are first trained and evaluated for the image semantic segmentation task on the different environments. Then, the models are tested in the scenes under the new spectral illuminations covering a range of hue and saturation values to see the shape of the tolerance region to changes in illumination. We show that in the naturally illuminated environment human-like tolerance regions emerge, but they do not in the counter-example environments where the color statistics are markedly different.

3.1 Model training and segmentation performance

In both kinds of architectures (without and with Divisive Normalization) the parameters of the nets are obtained via supervised learning: the models are trained to minimize a measure of the segmentation error over a set of images from the considered original environments. In this case, the selected measure was the Mean Absolute Error (MAE), which is maximised if, for each pixel, the correct class is predicted with probability one and the other classes have probability zero, and therefore lower MAE is better. The final performance of the networks was measured using the Intersection over Union (IoU) measure (Rahman and Wang, 2016) over the validation data, a subset of the 20% of the training images that are not used in the training process. IoU takes into account the predicted area and the real area for each class and how much they intersect and therefore higher is better, with IoU $\in [0, 1]$. We train each network during 200 epochs (each complete pass of the whole training data) using Adam as the optimizer (Kingma and Ba, 2014) and a batch size of 16 images. We keep the model parameters that achieve higher IoU on the validation data, which we compute after each epoch.

We trained *six* artificial systems performing image semantic segmentation: 2 *architectures* $\times$ 3 *environments*. This includes *two* biologically interesting cases (both architectures trained on the naturally illuminated images under daylight source), and *four* counter-examples: the ones trained in environments with non-natural illumination (underwater) or spectrally flat reflectances (achromatic images).

Given that the encoding part of the considered networks has certain resemblances with the retina-cortex pathway (Jacob et al., 2021), and the aforementioned biological inspiration of the divisive normalization layer (Abrams et al., 2007; Hillis and Brainard, 2005), our U-nets *have the ability* to use color information to solve segmentation. However, as other features (e.g., edges, shape, and textures) may also contribute to the solution of the problem there is no guarantee that these nets develop human-like tolerance to color shifts. The counter-example case that consists of achromatic images particularly was chosen to ensure that the networks trained in this condition do not use color information at all.

We tested the performance of the considered nets (U-Net and U-Net+DN) in the three environments where they were trained (numbers in bold-face in Table 1). Moreover, we did two extra tests in order to check the relevance of color in the segmentation problem. To do so we considered the databases that originally consisted on scenes under natural daylight illumination (CityScapes and Pets). In particular, we removed the color information in CityScapes, and we recovered the original color information in Pets (numbers in light-face in Table 1).

To test each model we perform 300 realizations where we randomly select 20 test images from their corresponding test set and compute the IoU performance. Table 1 summarizes the results.

First, as expected results show that using the model with DN layers generally improves the segmentation results (compare first and second row of numbers in bold). Second, when comparing the color segmentation importance, the most important factor seems to be consistency with training, i.e., the models trained with color images get worse when removing color information, and the models trained with achromatic images get worse when facing color images (compare columns in bold and light in the non-underwater environments). However, if we compare the reductions in performance, we see higher reductions in the color-trained models tested with achromatic images (21% average reduction in IoU) than in the achromatic-trained models tested with color images (12% average reduction in IoU). This highlights that color is certainly beneficial for segmentation.

In order to check the significance of the differences between the performances seen in Table 1, we carried out a Mann-Whitney U-test (Mann and Whitney, 1947). In this non-parametric test the null hypothesis is that the distribution of the set of samples of a variable is the same as the distribution of the samples of another

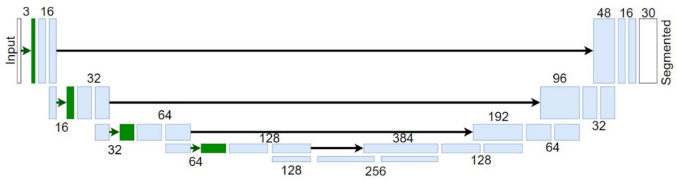


FIGURE 4
Illustrative U-Net architecture for image segmentation. The blocks in *blue* represent regular convolutional layers, and the blocks in *green* represent bio-inspired *Divisive Normalization* layers. Numbers by the layers indicate the number of features and black arrows represent the skip unions.

TABLE 1 Segmentation performance: test IoU results (mean ± standard deviation) of the models trained in the different environments when performing 300 evaluations over subsets of the test images.

Training env.	Color Natural Illum.		Underwater Illum.	Achrom. images	
Test env.	Color	Achrom.	Color	Achrom.	Color
U-Net	**0.77** ± 0.02	0.54 ± 0.03	**0.66** ± 0.05	**0.77** ± 0.02	0.72 ± 0.04
U-Net + DN	**0.78** ± 0.02	0.68 ± 0.02	**0.70** ± 0.04	**0.80** ± 0.02	0.66 ± 0.04

variable. Therefore, rejection of the null hypothesis implies that the compared variables are significantly different, i.e. one is larger than the other (Howell, 2013; Corder and Foreman, 2009). Table 2 shows the U-statistic over the number of samples (the effect size, or the proportion of pairs that support that items from group 2 are larger than items from group 1) and the corresponding p-values for all the different comparisons. We compared the IoU performance of the no-DN vs the DN models within each training environment (Table 2 top). We also compared the performance in the chromatic vs the achromatic version of the datasets (Table 2-bottom). In all the cases the null hypothesis was rejected (all p-values < 0.001), meaning that all the differences are significant. It is important to note that comparisons can be made only within each training environment because, as stated in the model definition, each dataset has a different number of classes and intrinsic difficulty.

3.2 Tolerance to illuminant change in segmentation networks

To test the tolerance to illuminant changes we evaluate the segmentation performance of the different networks (trained with the three different types of images and the two types of architecture) with the color-shifted scenes. We do the evaluation 300 times with subsets of the test images, following the same procedure we did to obtain the results in Table 1. We compare the performance of the models with images with spectral changes along the hue-saturation plane with regard to their results on their training set from Table 1. Then, we can define the tolerance/invariance region of a model as the hue and saturation combinations where the results change less than a certain threshold.

Figures 5, 6 show the variation of the segmentation performance as a function of the change of *saturation* and *hue* of the illuminant for the different architectures and the different environments considered. It also shows the corresponding *tolerance regions* for 3%, 5%, and 10% changes in performance with regard to the training situation. Finally, Figures 5, 6 also show the corresponding regions in the 1931 CIE xy diagram.

The gray level in the first row (zero saturation) of the *saturation-hue* planes represents the IoU performance of the segmentation network in the original scenes. The values of Table 1 are taken as reference in each case. Then, darker or lighter values for other illuminants correspond to lower or higher performance in the image segmentation task with regard to its reference. In particular, the curves in *purple*, *orange*, and *green*, represent variations of the performance of 3%, 5%, and 10%, respectively, with regard to their reference. Therefore, these curves also represent *tolerance regions* in the chromatic diagram where the performance departs from the original reference less than a certain threshold.

The specific size of the tolerance regions of course depends on the (arbitrarily) selected threshold for the departure with respect to the reference value. However, the fact that, given a threshold, the scale of the region is fundamentally different for the different environments is certainly relevant. Moreover, the (non-circular) shape and orientation of the regions indicate that the segmentation function learnt in a certain environment may imply anisotropies of the robustness of the (artificial) visual system under changes of illumination.

Chromatic diagrams in Figures 5, 6 display the root mean square error (RMSE) distance (in chromatic coordinates) between the mean 3% tolerance regions over the 300 iterations in the artificial systems and the corresponding color discrimination MacAdam ellipse in humans. For this comparison, the tolerance region in humans for that specific chromatic location was obtained by interpolating the parameters of the three closer ellipses out of the 25 regions measured in MacAdam (1942). In the corresponding (human and artificial) regions we took 20 points at uniformly distributed angles and we computed the average distance between

TABLE 2 Significance of the differences in segmentation performance: Mann-Whitney U-test statistic and *p*-values for the different comparisons: models (top), and achrom/color tests (bottom).

Training env.	Color natural illum.		Underwater illum.	Achrom. images	
Test env.	Color	Achro	Color	Achro	Color
MW-stat (no-DN vs. DN)	0.70	0.9998	0.73	0.76	0.82
p-val (no-DN vs. DN)	$1.1 \cdot 10^{-16}$	$1.2 \cdot 10^{-99}$	$1.7 \cdot 10^{-22}$	$4.4 \cdot 10^{-29}$	$5.9 \cdot 10^{-43}$

Training env.	Natural Illum.		Achro illum.	
Model	U-Net	U-Net+DN	U-Net	U-Net+DN
MW-stat (Achrom. vs. Color)	1.0	0.9998	0.90	0.99
p-val (Achrom. vs. Color)	$1.1 \cdot 10^{-99}$	$1.3 \cdot 10^{-99}$	$1.1 \cdot 10^{-65}$	$2.9 \cdot 10^{-99}$

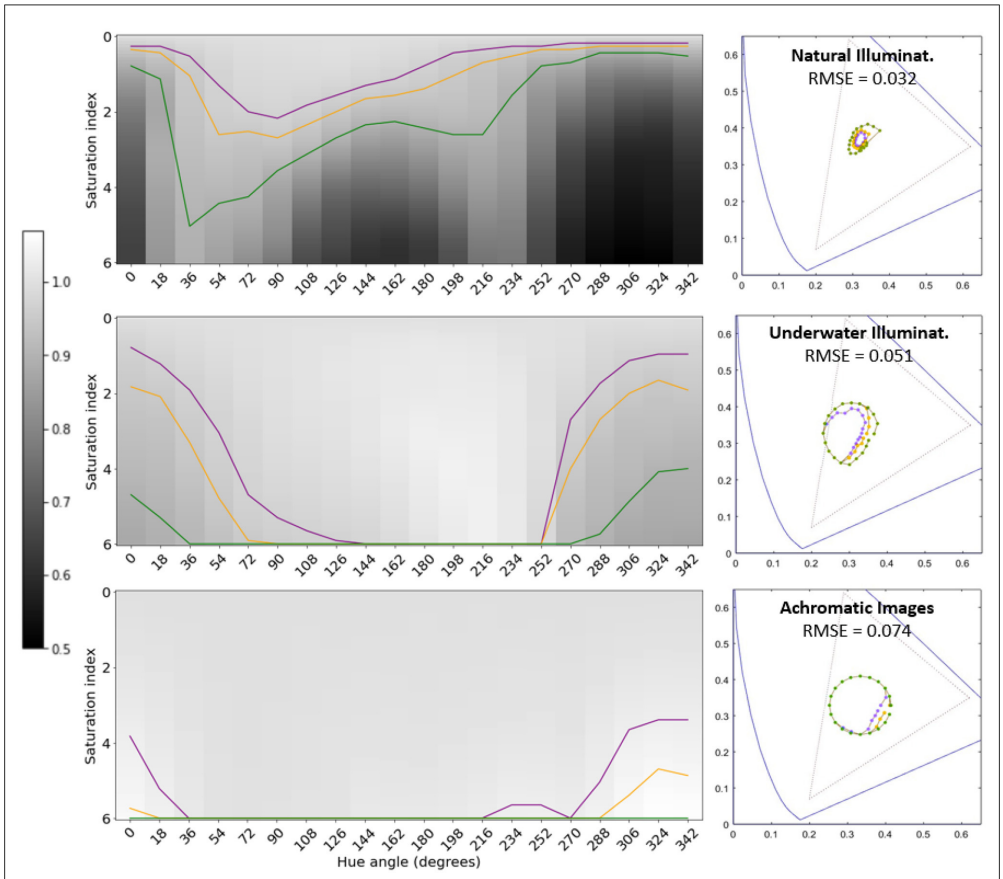


FIGURE 5
Tolerance of segmentation performance to illuminant change for different environments (regular U-Nets). The results of the *natural*, *underwater*, and *achromatic* environments are represented in the **top**, **middle**, and **bottom** rows respectively. Gray level represents the segmentation performance under different illuminations with regard to the reference performance obtained for the original scenes. Darker values represent lower performance. The curves in *purple*, *orange* and *green*, represent variations of the performance of 3%, 5%, and 10%, respectively. These curves define *tolerance regions* for performance in the chromatic diagram. The RMSE values represent the distance between the average of these tolerance regions in the artificial system and the corresponding tolerance ellipse in humans.

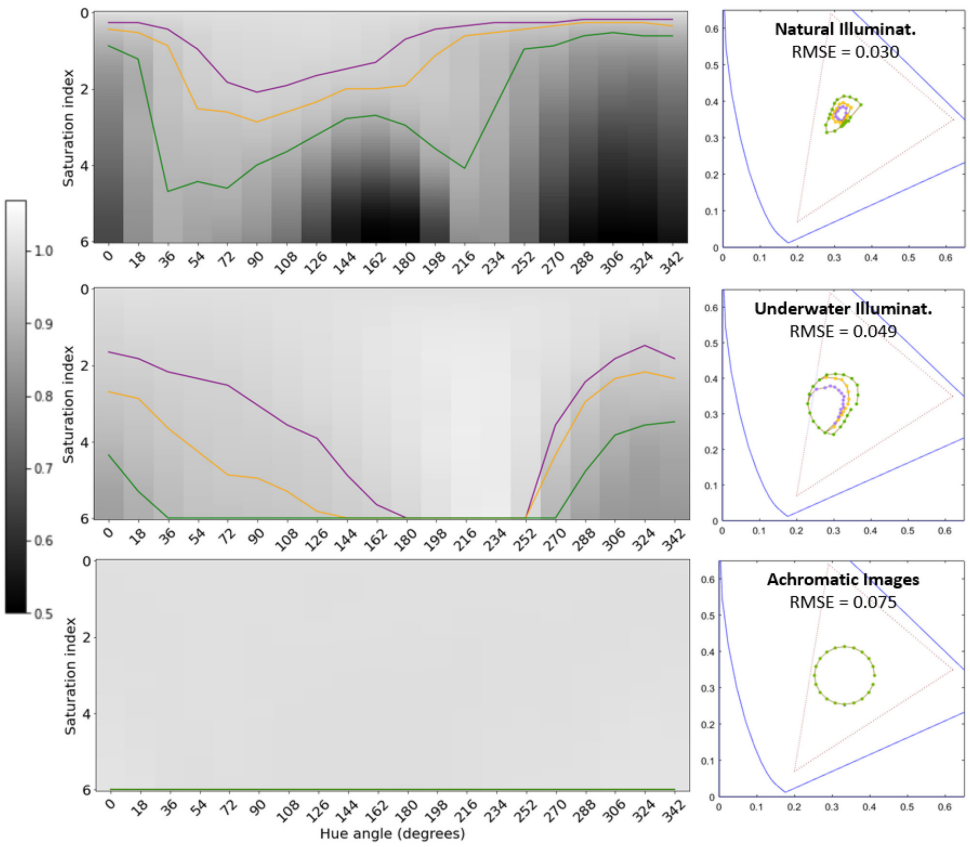


FIGURE 6
Tolerance of segmentation performance to illuminant change for different environments (U-Nets with Divisive Normalization). Same results as in Figure 5, but for the architecture with Divisive Normalization.

the corresponding points at those angles, leading to the reported RMSE value (in chromatic coordinates).

Results show an interesting alignment of the anisotropy of artificial systems with human anisotropy but only for natural scenes under daylight illumination. The counter-examples with unusual color statistics lead to non-human tolerance regions and anisotropies. In both counter-example environments, the performance is more insensitive to the changes in illumination, and this is particularly true for the architectures trained on images with flat spectral reflectance (achromatic images). As a result, the tolerance regions are substantially bigger for the same thresholds, and the insensitivity is more isotropic.

There may be two causes for this effect. On the one hand, the underwater scenes seem to have a wider color gamut with a smaller peak in the probability of colors around the mean (see scatter plots in the diagrams of Figure 1 and the corresponding ellipses representing the covariance matrices). This wider spread of colors (wider than in scenes in daylight illumination) would explain the bigger tolerance to color change of the systems trained in this unusual environment. On the other hand, the segmentation systems trained on images of flat spectral radiance may be insensitive to color just because (by construction of the training set) their ability for segmentation has to be based on non-chromatic features. Therefore, substantial changes of color should not affect much their performance, leading to big (and isotropic) tolerance regions.

The effect of the considered architectures in the size and orientation of the sensitivity regions is secondary: although the absolute performance of the networks equipped with Divisive Normalization is better (see Table 1 and slightly bigger areas of

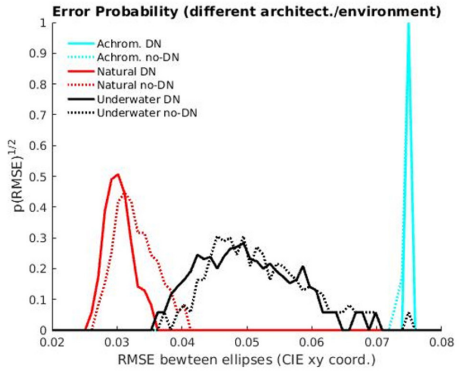


FIGURE 7
Distances to MacAdam ellipses: Histograms of the RMSE errors comparing the human MacAdam ellipses and the tolerance region of the models for 300 realizations with test subsets.

the insensitivity regions), this has low impact on the anisotropy depending on saturation and hue. The differences in the shape of the tolerance regions depend more strongly on the different image statistics rather than on the considered architectures.

To confirm the statistical significance of the results mentioned above we display the distributions of errors with the human discrimination ellipse and we perform non-parametric Kolmogorov-Smirnov tests to check if these samples of errors come from the same distribution or not. Figure 7 shows the histograms of the RMSE between the tolerance region of the models and the human MacAdam ellipse for the 300 realizations performed in the evaluation.

The distances between the histograms of errors confirm that the environment is the major factor in getting human-like ellipses, and it is way more important than the explored variants of the architecture. Of course, the 2-sample non-parametric Kolmogorov-Smirnov tests also confirm that these big differences (basically non-overlapping histograms for the different environments) are significant, with $p < 0.001$, (see the test statistics and the p-vaues in Table 3). The results of different architectures only introduce slight shifts in the histograms, so this is clearly a secondary (less relevant factor). In fact, the KS-tests reveal that differences in architecture are not significant in the non-natural cases (underwater images and achromatic images, in black and cyan, $p > 0.001$), but they are for the natural images (in red, $p < 0.001$). The histograms reveal that the significance of the difference between no-DN and DN networks according to the KS-test for natural images does not modify the fact that the environment is way more important than the architecture to get human-like results. Interestingly, the significance of the difference between the errors in the DN vs. no-DN case for natural images means that in the ecologically sensible situation, DN is important to increase alignment with humans, as expected from the rationale suggested in Hernández-Cámara et al. (2023b, 2024).

4 Discussion and conclusions

4.1 Summary of results

Artificial networks trained for image segmentation develop human-like tolerance to changes in illumination (around the white) when they are trained on natural images under daylight illumination. Similarly to humans, these networks are more tolerant to variations in the yellow-blue direction rather than in the red-green direction: see the similarity between two-minima the curve in the colored saturation-hue panel of Figure 2 and the shape of the performance surfaces in the top panels of Figures 5, 6. This anisotropy occurs both for regular U-Net architectures, Figure 5-*top row*, and with architectures augmented with the biologically-inspired Divisive Normalization, Figure 6-*top row*.

However, alternative environments with markedly different image statistics (e.g., underwater scenes and achromatic scenes) lead to systems in which the tolerance to color changes is not aligned with human color discrimination (substantially bigger insensitivity with lower anisotropy), Figures 5, 6-*middle and bottom rows*.

4.2 Function, architecture, or just image statistics?

The reported emergence of a human-like anisotropy in the tolerance to color changes in artificial systems trained in a natural environment with natural illumination means that image segmentation (which is, at least partially, based on color) *could be* the principle behind the development of the anisotropy observed in humans for color discrimination.

However, not all the explanations can be attributed to the specific *segmentation* function. First, lower-level functions that involve local equalization of the color manifold, such as *error minimization* and *information maximization* (Laparra et al., 2012; Laparra and Malo, 2015; da Fonseca and Samengo, 2016, 2018), also lead to this kind of asymmetry. See that the ellipses from the local-PCA in Figure 1-*left and middle* (good for local equalization) have qualitatively similar properties as the tolerance regions that emerge in the segmentation networks. Second, more than the function, it is the data distribution that may lead to the observed asymmetry in the behaviour of the networks. In fact, statistical analysis (Figure 7 and associated Kolmogorov-Smirnov tests) shows that a natural color distribution is the major factor in getting human-like ellipses. The following example connects a strong physical constraint with a major asymmetry in the color data that may explain differences in performance and discrimination. If the spectrum of the sunlight at different times of the day can be approximated by a black-body radiator (Malo and Jiménez, 2011; Jiménez and Malo, 2014) the manifold of natural colors will be elongated along the Planckian locus in the 1931 CIE xy diagram. This locus, for the white (Wyszecki and Stiles, 2000), approximately has the orientation of the ellipse in Figure 1-*left*, and the regions of Figures 5, 6 *Top*. This makes sense because the natural dataset will have multiple examples of similar objects with different illuminations along that (yellow-blue) direction. As

TABLE 3 Significance of the differences of RMSE errors.

		Natural illum.		Underwater		Achromatic	
		No DN	DN	No DN	DN	No DN	DN
Natural illum.	No DN	-	0.79	0.9998	0.9992	1.0	1.0
	DN		-	1.0	1.0	1.0	1.0
Underwater	No DN			-	0.56	0.998	0.998
	DN				-	1.0	1.0
Achromatic	No DN					-	0.52
	DN						-
		Natural illum.		Underwater		Achromatic	
		No DN	DN	No DN	DN	No DN	DN
Natural Illum.	No DN	-	$4.2 \cdot 10^{-35}$	$1.3 \cdot 10^{-99}$	$2.2 \cdot 10^{-99}$	$1.9 \cdot 10^{-111}$	$1.1 \cdot 10^{-113}$
	DN		-	$1.1 \cdot 10^{-99}$	$1.1 \cdot 10^{-99}$	$1.9 \cdot 10^{-111}$	$1.9 \cdot 10^{-113}$
Underwater	No DN			-	0.01	$8.2 \cdot 10^{-111}$	$4.4 \cdot 10^{-113}$
	DN				-	$1.9 \cdot 10^{-111}$	$1.1 \cdot 10^{-113}$
Achromatic	No DN					-	$8.9 \cdot 10^{-5}$
	DN						-

a result, in order to obtain good segmentation performance, the networks (of whatever architecture) have to be more invariant to changes of illumination in that direction. Third, the counter-examples of markedly different color statistics imply that the same functional goal leads to very different tolerance regions.

Finally, the architecture selected to perform the image segmentation does not seem to have a big impact on the alignment of the asymmetries of humans and networks (see histograms in Figure 7). In fact, the functions related to information maximization and error minimization reduce to local PCA (Laparra et al., 2012), and hence they are independent of the architecture. In Figure 1 we see that local PCA leads to regions which are similar to the tolerance regions found in the different image segmentation networks when trained in similar environments, see Figures 5, 6. However, there is a small, but statistically significant difference (histograms in red in Figure 7 with hypothesis-zero rejected by KS-test with $p < 0.001$) that suggests that the Divisive Normalization is important to improve the alignment with humans in color discrimination in the ecologically significant case. This is consistent with the suggestions done in Hernández-Cámara et al. (2023b, 2024).

4.3 Conclusions

Artificial networks for image segmentation trained in natural environments with natural illumination exhibit human-like tolerance to changes in illuminant, aligning with human color discrimination. This is the first report on the emergence of the alignment of image segmentation networks with human color discrimination. However, in environments with markedly different image statistics, the tolerance to color changes in these artificial systems deviates from human color discrimination. This suggests that the regularities of the environment are much more significant in shaping the behaviour for color discrimination than the architecture of the image segmentation network. This is in contrast with other chromatic properties, e.g., color induction (Gomez-Villa et al., 2020) or color CSFs (Li et al., 2022), where the architecture strongly modifies the human-machine similarities. In fact, in the discrimination case considered here, alternative functional principles such as error minimization or information maximization (Laparra et al., 2012; Laparra and Malo, 2015; da Fonseca and Samengo, 2016, 2018) which only depend on the data (e.g., local PCA), also lead to tolerance regions of human-like orientation if applied in the proper environment. In conclusion, the anisotropy in human color discrimination is also present in segmentation neural networks. This is probably due to the adaptation of (both natural and artificial) neural networks to the color data distribution.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/pablohc97/SegmentationModelsAlignedColorDiscrimination.

Author contributions

PH-C: Conceptualization, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. PD-O: Investigation, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. VL: Writing – original draft, Writing – review & editing. JM: Conceptualization,

Formal analysis, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by MICIIN/FEDER/UE under Grants PID2020-118071GB-I00, PDC2021-121522-C21 (funded by MCIN/AEI/10.13039/501100011033 and the EU NextGenerationEU/PRTR) and Grant PID2023-152133NB-I00; in part by Spanish MIU under Grant FPU21/02256; and in part by Generalitat Valenciana under Projects GV/2021/074, CIPROM/2021/056, and CIAPOT/2021/9. The authors gratefully acknowledge the computer resources at Artemisa and the technical support provided by the Instituto de Fisica Corpuscular, IFIC(CSIC-UV). Artemisa was co-funded by the European Union through the 2014-2020 ERDF Operative Programme of Comunitat Valenciana, project IDIFEDER/2018/048.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Abrams, A. B., Hillis, J. M., and Brainard, D. H. (2007). The relation between color discrimination and color constancy: when is optimal adaptation task dependent? *Neural Comput.* 19, 2610–2637. doi: 10.1162/neco.2007.19.10.2610

Akbarinia, A., Morgenstern, Y., and Gegenfurtner, K. R. (2023). Contrast sensitivity function in deep networks. *Neur. Netw.* 164, 228–244. doi: 10.1016/j.neunet.2023.04.032

Alabau-Bosque, N., Daudén-Oliver, P., Vila-Tomás, J., Laparra, V., and Malo, J. (2024). Invariance of deep image quality metrics to affine transformations. *arXiv preprint arXiv:2407.17927*.

Atick, J. J., Li, Z., and Redlich, A. N. (1992). Understanding retinal color coding from first principles. *Neural Comput.* 4, 559–572. doi: 10.1162/neco.1992.4.4.559

Atick, J. J., and Redlich, A. N. (1992). What does the retina know about natural scenes? *Neural Comput.* 4, 196–210. doi: 10.1162/neco.1992.4.2.196

Barbur, J. L. (2004). "Double-blindsight" revealed through the processing of color and luminance contrast defined motion signals. *Progr. Brain Res.* 144, 243–259. doi: 10.1016/S0079-6123(03)14417-2

Barlow, H. (1959). "Sensory mechanisms, the reduction of redundancy, and intelligence," in *Proceedings of the National Physics Laboratory Symposium on the Mechanization of Thought Process*, 535–539.

Barlow, H. (2001). Redundancy reduction revisited. *Network: Comp. Neur. Syst.* 12, 241–253. doi: 10.1080/net.12.3.241.253

Buchsbaum, G., and Gottschalk, A. (1983). Trichromacy, opponent colours coding and optimum colour information transmission in the retina. Proceedings of the Royal Society of London. Series B. *Biol. Sci.* 220, 113–189. doi: 10.1098/rspb.1983.0090

Carandini, M., and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* 13, 51–62. doi: 10.1038/nrn3136

Chong, H. Y., Gortler, S. J., and Zickler, T. (2007). "The von kries hypothesis and a basis for color constancy," in *International Conference on Computer Vision*, 1–8. doi: 10.1109/ICCV.2007.4409102

Clifford, C., Webster, M., Stanley, G., Stocker, A., Kohn, A., Sharpee, T., et al. (2007). Visual adaptation: neural, psychological and computational aspects. *Vision Res.* 47, 3125–3131. doi: 10.1016/j.visres.2007.08.023

Corder, G., and Foreman, D. (2009). *Comparing Two Unrelated Samples: The Mann-Whitney U-Test, chapter 4*. London: John Wiley and Sons, Ltd. 57–78. doi: 10.1002/9781118165881.ch4

Cordts, M., Omran, M., et al. (2016). "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR). doi: 10.1109/CVPR.2016.350

da Fonseca, M., and Samengo, I. (2016). Derivation of human chromatic discrimination ability from an information-theoretical notion of distance in color space. *Neural Comput.* 28, 2628–2655. doi: 10.1162/NECO_a_00903

da Fonseca, M., and Samengo, I. (2018). Novel perceptually uniform chromatic space. *Neural Comput.* 30, 1612–1623. doi: 10.1162/neco_a_01073

Deeb, R., Muselet, D., et al. (2018). Interreflections in computer vision: a survey and an introduction to spectral infinite-bounce model. *J. Math. Imaging Vis.* 60, 661–680. doi: 10.1007/s10851-017-0781-x

DeValois, R. L., and DeValois, K. K. (1993). A multi-stage color model. *Vision Res.* 33 8, 1053–1065. doi: 10.1016/0042-6989(93)90240-W

Doi, E., Inui, T., Lee, T., Wachtler, T., and Sejnowski, T. (2003). Spatiochromatic receptive field properties derived from information-theoretic analyses of cone mosaic responses to natural scenes. *Neural Comp.* 15, 397–417. doi: 10.1162/089976603762552960

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., and Koltun, V. (2017). "CARLA: an open urban driving simulator," in *Annual Conference on Robot Learning*, 1–16.

Fairchild, M. (2013). *Color Appearance Models. The Wiley-IST Series in Imaging Science and Technology*. New York: Wiley. doi: 10.1002/97811186 53128

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A* 4, 2379–2394. doi: 10.1364/JOSAA.4.002379

Finlayson, G. D., Drew, M. S., and Funt, B. V. (1993). "Diagonal transforms suffice for color constancy," in *International Conference on Computer Vision*, 164–171. doi: 10.1109/ICCV.1993.378223

Foster, D. H., Amano, K., and Nascimento, S. M. (2016). Time-lapse ratios of cone excitations in natural scenes. *Vision Res.* 120, 45–60. doi: 10.1016/j.visres.2015.03.012

Gomez-Villa, A., Martin, A., Vazquez, J., Bertalmío, M., and Malo, J. (2020). Color illusions also deceive CNNs for low-level vision tasks: analysis and implications. *Vision Res.* 176, 156–174. doi: 10.1016/j.visres.2020.07.010

Guo, Y., Liu, Y., Georgiou, T., and Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *Int. J. Multim. Inf. Retr.* 7, 87–93. doi: 10.1007/s13735-017-0141-z

Gutmann, M., Laparra, V., Hyvärinen, A., and Malo, J. (2014). Spatio-chromatic adaptation via higher-order canonical correlation analysis of natural images. *PLoS ONE* 9:e86481. doi: 10.1371/journal.pone.0086481

Heasly, B. S., Cottaris, N. P., Lichtman, D. P., Xiao, B., and Brainard, D. H. (2014). Rendertoolbox3: Matlab tools that facilitate physically based stimulus rendering for vision research. *J. Vision* 14, 6–6. doi: 10.1167/14.2.6

Hernández-Cámara, P., Vila-Tomás, J., Dauden-Oliver, P., Alabau-Bosque, N., Laparra, V., and Malo, J. (2024). Image segmentation via divisive normalization: dealing with environmental diversity. *arXiv preprint arXiv:2407.17829*.

Hernández-Cámara, P., Vila-Tomás, J., Laparra, V., and Malo, J. (2023a). *Dissecting the Effectiveness of Deep Features as a Perceptual Metric*. Available at: https://ssrn.com/abstract=4609207

Hernández-Cámara, P., Vila-Tomás, J., Laparra, V., and Malo, J. (2023b). Neural networks with divisive normalization for image segmentation. *Pattern Recognit. Lett.* 173, 64–71. doi: 10.1016/j.patrec.2023.07.017

Hernández-Cámara, P., Vila-Tomás, J., Malo, J., and Laparra, V. (2024). "Measuring human-clip alignment at different abstraction levels," in *ICLR 2024 Workshop on Representational Alignment*.

Hillis, J., and Brainard, D. (2005). Do common mechanisms of adaptation mediate color discrimination and appearance? Uniform backgrounds. *J. Opt. Soc. Am. A* 22, 2090–2106. doi: 10.1364/JOSAA.22.002090

Howell, D. (2013). *Statistical Methods for Psychology*. New York: Thomson Wadsworth.

Hunt, R. (2005). *The Reproduction of Colour*. Sussex, UK: Wiley. doi: 10.1002/0470024275

Hyvärinen, A., Hurri, J., and Hoyer, P. O. (2009). *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.*, *volume 39*. Cham: Springer Science Business Media. doi: 10.1007/978-1-84882-491-1

Islam, M. J., Edge, C., Xiao, Y., Luo, P., Mehtaz, M., Morse, C., et al. (2020). Semantic segmentation of underwater imagery: Dataset and benchmark," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS). doi: 10.1109/IROS45743.2020.9340821

Jacob, G., Pramod, R., Katti, H., and Arun, S. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nat. Commun.* 12:1872. doi: 10.1038/s41467-021-22078-3

Jennings, B. J., and Barbur, J. L. (2010). Colour detection thresholds as a function of chromatic adaptation and light level. *Ophthal. Physiol. Opt.* 30, 560–567. doi: 10.1111/j.1475-1313.2010.00773.x

Jiménez, S., Laparra, V., and Malo, J. (2013). "Visual discrimination and adaptation using non-linear unsupervised learning," in *Human Vision and Electronic Imaging XVIII*, 395–400. doi: 10.1117/12.2019008

Jiménez, S., and Malo, J. (2014). The role of spatial information in disentangling the irradiance-reflectance-transmittance ambiguity. *IEEE Trans. Geosci. Rem. Sens.* 52, 4881–4894. doi: 10.1109/TGRS.2013.2285731

Jolliffe, I. T. (2002). *Principal Component Analysis*. New York: Springer.

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Laparra, V., Jimnez, S., Camps, G., and Malo, J. (2012). Nonlinearities and adaptation of color vision from sequential principal curves analysis. *Neural Comput.* 24, 2751–2788. doi: 10.1162/NECO_a_00342

Laparra, V., and Malo, J. (2015). Visual aftereffects and sensory nonlinearities from a single statistical framework. *Front. Hum. Neurosci.* 9:557. doi: 10.3389/fnhum.2015.00557

Laparra, V., and Malo, J. (2016). Sequential principal curves analysis. *arXiv preprint arXiv:1606.00856*.

Laughlin, S. (1983). "Matching coding to scenes to enhance efficiency," in *Physical and Biological Processing of Images: Proceedings of an International Symposium*, 42–52. doi: 10.1007/978-3-642-68888-1_4

Li, Q., Gomez-Villa, A., Bertalmío, M., and Malo, J. (2022). Contrast sensitivity functions in autoencoders. *J. Vis.* 22:8. doi: 10.1167/jov.22.6.8

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Trans. Inf. Theory* 28, 129–137. doi: 10.1109/TIT.1982.1056489

MacAdam, D. L. (1942). Visual sensitivities to color differences in daylight. *J. Opt. Soc. Am.* 32, 247–274. doi: 10.1364/JOSA.32.000247

MacLeod, D., and von der Twer, T. (2003). "The pleistochrome: optimal opponent codes for natural colors," in *Color Perception: From Light to Object*, eds. D. Heyer (Oxford, UK: Oxford Univ. Press). doi: 10.1093/acprof:oso/9780198505006.003.0005

Malo, J., and Jiménez, S. (2011). *The statistics of Remote Sensing Images, chapter 2, pages 19-35*. San Rafael: Morgan Claypool Publishers.

Malo, J., and Luque, M. (2002). *ColorLab: A Matlab Toolbox for Color Science and Calibrated Color Image Processing*. Servei de Publicacions de la Universitat de Valencia. Available at: http://isp.uv.es/code/visioncolor/colorlab.html (accessed October 1, 2024).

Mann, H. B., and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Mathem. Stat.* 18, 50–60. doi: 10.1214/aoms/1177730491

Martinez, M., Bertalmío, M., and Malo, J. (2019). In praise of artifice reloaded: Caution with natural image databases in modeling vision. *Front. Neurosci.* 13:8. doi: 10.3389/fnins.2019.00008

Nascimento, S. M., Amano, K., and Foster, D. H. (2016). Spatial distributions of local illumination color in natural scenes. *Vision Res.* 120, 39–44. doi: 10.1016/j.visres.2015.07.005

Olshausen, B. A., and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609. doi: 10.1038/381607a0

Parkhi, O., Vedaldi, A., Zisserman, A., and Jawahar, C. (2012). "Cats and dogs," in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 3498–3505. doi: 10.1109/CVPR.2012.6248092

Poggio, T. (2021). *From Marr's vision to the problem of human intelligence*. Center for Brains Minds Machines CBMM Memo.

Rahman, M. A., and Wang, Y. (2016). "Optimizing intersection-over-union in deep neural networks for image segmentation," in *International Symposium on Visual Computing* (Cham: Springer International Publishing), 234–244. doi: 10.1007/978-3-319-50835-1_22

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nat. Neurosci.* 22, 1761–1770. doi: 10.1038/s41593-019-0520-2

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention-MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18* (MICCAI), 234–241. doi: 10.1007/978-3-319-24574-4_28

Shapley, R., and Hawken, M. (2011). Color in the cortex-single- and double-opponent cells. *Vision Res.* 51, 701–717. doi: 10.1016/j.visres.2011.02.012

Simoncelli, E. P., and Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24, 1193–1216. doi: 10.1146/annurev.neuro.24.1.1193

Stockman, A., and Brainard, D. (2010). *OSA Handbook of Optics (3rd. Ed.)*, *chapter. Color Vision Mechanisms*. NY: McGraw-Hill, 147–152.

Torralba, A., and Oliva, A. (2003). Statistics of natural image categories. *Network* 14:391. doi: 10.1088/0954-898X_14_3_302

Vila-Tomás, J., Hernández-Cámara, P., and Malo, J. (2023). Artificial psychophysics questions classical hue cancellation experiments. *Front. Neurosci.* 17:1208882. doi: 10.3389/fnins.2023.1208882

von der Twer, T., and MacLeod, D. I. A. (2001). Optimal nonlinear codes for the perception of natural colours. *Network* 12:395. doi: 10.1080/net.12.3.395.407

Watson, A. B., and Solomon, J. A. (1997). Model of visual contrast gain control and pattern masking. *J. Opt. Soc. Am. A* 14, 2379–2391. doi: 10.1364/JOSAA.14.002379

Webster, M. A., and Mollon, J. (1997). Adaptation and the color statistics of natural images. *Vis. Res.* 37, 3283–3298. doi: 10.1016/S0042-6989(97)00125-9

Wyszecki, G., and Stiles, W. (2000). *Color Science: Concepts and Methods, Quantitative Data and Formulae*. New Jersey: John Wiley Sons.

Neural Networks 185 (2025) 107189

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Full Length Article

Dissecting the effectiveness of deep features as metric of perceptual image quality

Pablo Hernández-Cámara *, Jorge Vila-Tomás, Valero Laparra, Jesús Malo

Image Processing Lab., Universitat de València, 46980 Paterna, Spain

ARTICLE INFO

Dataset link: KADID-10k (Reference data), TID2013 (Reference data)

Keywords:
Image quality
Neural networks
Visual neuroscience
Functional principle
Learning environment
Architecture

ABSTRACT

There is an open debate on the role of artificial networks to understand the visual brain. Internal representations of images in artificial networks develop human-like properties. In particular, evaluating distortions using differences between internal features is correlated to human perception of distortion. However, the origins of this correlation are not well understood.

Here, we dissect the different factors involved in the emergence of human-like behavior: *function*, *architecture*, and *environment*. To do so, we evaluate the aforementioned human-network correlation at different depths of 46 pre-trained model configurations that include no psycho-visual information. The results show that most of the models correlate better with human opinion than SSIM (a de-facto standard in subjective image quality). Moreover, some models are better than state-of-the-art networks specifically tuned for the application (LPIPS, DISTS). Regarding the function, supervised classification leads to nets that correlate better with humans than the explored models for self- and non-supervised tasks. However, we found that better performance in the task does not imply more human behavior. Regarding the architecture, simpler models correlate better with humans than very deep nets and generally, the highest correlation is not achieved in the last layer. Finally, regarding the environment, training with large natural datasets leads to bigger correlations than training in smaller databases with restricted content, as expected. We also found that the best classification models are not the best for predicting human distances.

In the general debate about understanding human vision, our empirical findings imply that explanations have not to be focused on a single abstraction level, but all *function*, *architecture*, and *environment* are relevant.

1. Introduction

The internal representation of artificial networks that solve vision tasks is somehow correlated to the image representation in the visual brain (Hong, Yamins, Majaj, & DiCarlo, 2016; Kheradpisheh, Ghodrati, Ganjtabesh, & Masquelier, 2016; Yamins & DiCarlo, 2016), and deep models are the state-of-the-art in predicting human opinion of visual distortion (Ding, Ma, Wang, & Simoncelli, 2020; Hepburn, Laparra, Malo, McConville, & Santos-Rodriguez, 2020; Zhang, Isola, Efros, Shechtman, & Wang, 2018). However, the similarity between the image distances computed from networks and subjective image quality is not well understood. In fact, the emergence of this human behavior in artificial networks has been qualified as *unreasonable* (Zhang et al., 2018).

More generally, there is an open debate about the role of artificial networks as a model of the visual brain (Bowers, Malhotra, et al., 2022): while part of the community considers deep-nets as the current best option to understand biological vision (Cadena et al., 2019; Kriegeskorte, 2015; Yamins, Hong, Cadieu, Solomon, Seibert, & DiCarlo, 2014), there is a growing body of evidences pointing out their limitations (Bertalmío, Gomez-Villa, Martín, Vazquez-Corral, Kane, & Malo, 2020; Geirhos, Bethge, & Wichmann, 2020; Geirhos et al., 2019; Gomez-Villa, Martín, Vazquez-Corral, Bertalmío, & Malo, 2020; Li, Gomez-Villa, Bertalmío, & Malo, 2022) and stressing the need of proper comparison of artificial and natural visual networks (Akbarinia, Morgenstern, & Gegenfurtner, 2023; Funke et al., 2021; Geirhos, Meding, & Wichmann, 2020; Martinez-Garcia, Bertalmío, & Malo, 2019; Wichmann et al., 2017).

Understanding metrics of perceptual image quality with deep-nets is a perfect case study in this broader context. On the one hand, *subjective* distance is related to visual neuroscience for obvious reasons: even if nets are naively thought of as pure regressors tuned to solve a specific problem (as in Shakhnarovich, Batra, Kulis, & Weinberger,

* Corresponding author.
E-mail addresses: pablo.hernandez-camara@uv.es (P. Hernández-Cámara), jorge.vila-tomas@uv.es (J. Vila-Tomás), valero.laparra@uv.es (V. Laparra), jesus.malo@uv.es (J. Malo).

https://doi.org/10.1016/j.neunet.2025.107189
Received 17 October 2023; Received in revised form 7 January 2025; Accepted 15 January 2025
Available online 27 January 2025

Table 1
Our method: we isolated each potentially independent cause to explain similarity with human judgment. Within the big conceptual categories, *function*, *architecture*, *environment*, each column shows the different factors that we varied in our study and each row shows the different options explored.

Function	Architecture			Environment
task	Connections	Read-out		training data
Supervised Self-supervised	AlexNet	Euclidean	No concat.	ImageNet-1K
Supervised Classif.	AlexNet VGG-16 ResNet-50 DenseNet-121 EfficientNet-B0 ConvNeXt ViT	Euclidean	No concat.	ImageNet-1K
Supervised Classif.	AlexNet	Euclidean Mean Means-Sigmas Gram	No concat. Concat.	ImageNet-1K
Supervised Classif.	AlexNet	Euclidean	No concat.	ImageNet-1K Places-365 Cifar-10

2011), somehow they model the early stages of human vision. On the other hand, following Barlow's intuitions (Barlow, 1959, 2001), recent works stress the link between (artificial) statistical learning and (human) measures of subjective image distance (Hepburn, Laparra, Santos-Rodríguez, Ballé, & Malo, 2022; Kumar, Houlsby, Kalchbrenner, & Cubuk, 2022).

In this regard, approaches focused on image statistics (Hepburn et al., 2022) are consistent with the classical conceptual separation between *function* and *architecture* proposed by Marr and Poggio (1976). However, given the recent advances in deep nets for vision, this conceptual separation has been questioned by Poggio himself because, in general, the interactions between function (task), data (environment), and architecture are not obvious (Malo & Hernandez-Camara, 2024; Poggio, 2021).

In this work, we analyze the problem of metrics for subjective image quality (or perceptual metrics) in deep-nets by isolating *function*, *environment* and *architecture*. Specifically, we study the correlation between human opinion of distortion and image distances computed in the feature space of networks pre-trained for different vision tasks (though not explicitly trained to reproduce visual psychophysics). We do this by analyzing the key factors in turn: we consider *the function* through different supervised, self-supervised, and no-supervised goals, we consider *the environment* by using training data with different statistics, we consider *the architecture* in two ways: by using both shallow and deep networks and by checking the correlation at different layers of the models, and finally, we consider different *read-out strategies*, as plain Euclidean difference in the feature space, non-Euclidean (fine-tuned) differences, or the use of statistical summaries of the responses.

The detailed empirical analysis done here includes the following novel aspects with some connections to previous work. In particular, (a) We consider 29 different tasks (not only classification) because the consideration of multiple possible functions is central in the debate about the underlying principles of neural organization. (b) We analyze both shallow and very deep architectures for perceptual distances because, for other visual behaviors (Akbarinia et al., 2023; Bertalmío et al., 2020; Gomez-Villa, Martín, et al., 2020; Li et al., 2022), simpler architectures have been consistently identified as *more human* than deeper architectures. (c) We report correlations using multiple concatenated features for different depths. It gives an upper bound on the information that can be extracted. However, here we also report the correlation *at each layer*. This (per layer) analysis is more consistent with what happens in psychophysical models of early vision in which the signal is progressively refined (or information is progressively discarded). To the best of our knowledge, this is the first work that analyzes how well all the features extracted (layer by layer) correlate with human perception. The progressive consideration of additional distinct psychophysical facts in the different layers leads to progressive improvements in the correlation with humans as reported in Gomez-Villa, Bertalmio, and Malo (2020), Malo (2020), Malo and Simoncelli (2015), Martinez, Cyriac, Batard, Bertalmío, and Malo (2018). (d) On top of the Euclidean metric, we also consider non-Euclidean metrics (fine-tuned weights for the different features in a similar way to studies on contrast sensitivity (Akbarinia et al., 2023)). (e) Besides we study the effect of different statistical summaries for the read-out. (f) We consider datasets with different amounts and sizes of images, and where the statistics of the environment are different. (g) We compute the correlation with two different image quality assessment databases that provide mean opinion scores (TID-2013 (Ponomarenko et al., 2015) and KADID-10K (Lin, Hosu, & Saupe, 2019)). And finally, (h) we avoid our own bias in the training process by using only publicly-available models pre-trained by other groups.

The rest of the article is organized as follows: first in Section 2 we describe our methodology, the tested models and how we evaluate the models. Then, in Section 3 we show the different results we obtained when dissecting the correlation with human perception. Next, in the discussion (Section 4) we describe the non-trivial relation between accuracy in the function and correlation with perception. Moreover, we analyze the effect of language in the training and discuss the use of perceptual data to fine-tune the best model found in our previous (perceptually-agnostic) analysis. Finally, we also analyze in detail how the different distortion types and images affect the network correlation with human perception. Finally in Section 5 we summarize the conclusions of our work.

2. Methods

Our objective is to dissect the different factors that could affect the emergence of human-like behavior, such as the *architecture*, the *function*, the *environment* or the *read-out* of the model. Table 1 summarizes the methodology we followed to perform this analysis. Each row describes one experiment, and each column lists the factors analyzed. We successively isolated the key factors, fixing all but one.

We show the same results in each experiment: the correlation between human evaluations (perception) and inter-image distances computed using the features of neural network models at different depths (see, for example, Fig. 1). In the following, we describe the different neural network models analyzed, the human rate databases used for evaluation, and how distances are computed from the neural networks.

2.1. Models

We restrict ourselves to already trained models by third-parties and we use the models with default parameters. We do so under the reasonable assumption that the authors did their best to optimize the performance of their models, and thus the pre-trained networks likely represent the best achievable performance in the tasks for their respective architectures. Note that meaningful comparisons can only be made between models trained with the most appropriate hyperparameters for each architecture, ensuring convergence to the optimal performance. In this way, here we do not address the effects of controlled modifications in the architectures and training procedures. These are out of the scope of this work. This limitation, to be addressed in future work, does not preclude the interest of the analysis of well-established, pre-trained models since they are widely used in both scientific research and practical applications in their current form. Following this to avoid our own bias in trainig the networks, we used a range of pre-trained models tuned on ImageNet (Deng et al., 2009), Places-365 (Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014) and Cifar-10 (Krizhevsky, Hinton, et al.,

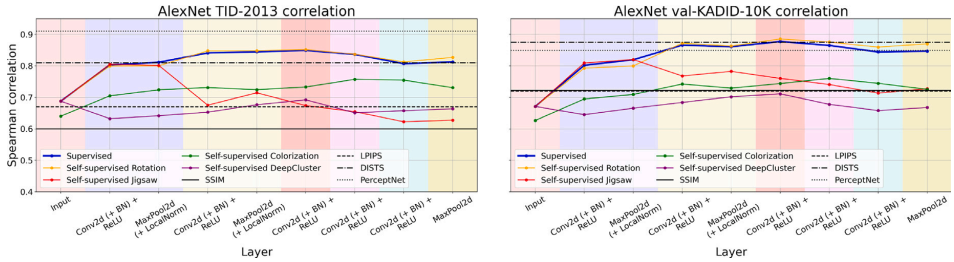


Fig. 1. Function and task 1: AlexNet. Plots display the Spearman correlation with human opinion at different layers of AlexNet trained for *different tasks*. We consider the similarity with humans in the TID-2013 database (left) and in the val-KADID-10K database (right). Background colors represent the different AlexNet blocks. The performance of some standard image quality measures in each database is shown as black lines for convenient reference.

2009) databases, in supervised, self- and no-supervised ways.

More particularly, for the supervised models, we used AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), VGG-16 (Simonyan & Zisserman, 2015), DenseNet-121 (Huang, Liu, Van Der Maaten, & Weinberger, 2017), ResNet-50 (He, Zhang, Ren, & Sun, 2016), EfficientNet-B0 (Tan & Le, 2019), ConvNeXt-tiny (Liu et al., 2022) and Vision Transformer base-patch16-224 (Dosovitskiy et al., 2021). We downloaded all of them with ImageNet-1K pre-trained weights from TorchVision (TorchVision maintainers and contributors, 2016). For other training datasets we downloaded them trained in Places-365 (Zhou et al., 2014) from MIT CSAIL Computer Vision (Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2017) and trained in Cifar-10 (Krizhevsky et al., 2009) from Raschka (2023).

For the self-supervised models, we use the architecture that obtains higher correlations (see Section 3.2) and download it from Facebook research VISSL, a library for state-of-the-art self-supervised models, Goyal et al. (2021) with ImageNet pre-trained weights. We tested different self-supervised tasks such as RotNet (Komodakis & Gidaris, 2018), Jigsaw (Noroozi & Favaro, 2016), Colorization (Zhang, Isola, & Efros, 2016) and DeepCluster (Caron, Bojanowski, Joulin, & Douze, 2018).

As an extension, we also analyze the ResNet models trained in the Taskonomy database (Zamir et al., 2018) for 24 different tasks. Appendix A shows a resume table with information of all the analyzed models and their accuracy in classification in different databases.

2.2. Distance measurement in deep features

There are several ways to measure the distance between two images. For example, let $x_0 \in \mathbb{R}^{(H,W,C)}$ and $y_0 \in \mathbb{R}^{(H,W,C)}$ denote an image and its distorted version, and $\hat{x}^l \in \mathbb{R}^{(H^l,W^l,C^l)}$ and $\hat{y}^l \in \mathbb{R}^{(H^l,W^l,C^l)}$ their corresponding feature maps at the lth layer of a network. Then, their euclidean distance at the lth layer is just:

$$d^l(x_0, y_0) = \sqrt{\sum_{H^l,W^l,C^l} (\hat{x}^l - \hat{y}^l)^2} \tag{1}$$

However, it is also possible to statistically summarize the layer output before computing the distance in a similar way to what some classical and state-of-the-art image quality assessment models do (Ding et al., 2020; Wang, Bovik, Sheikh, & Simoncelli, 2004). They compute the mean and standard deviation of different layers features when two images pass through networks to calculate the distance between them. This use of the statistics of images has been suggested as a signature in the visual cortex (*A functional and perceptual signature of the second visual area in primates*, 2013) and as an image texture descriptor (Portilla & Simoncelli, 2000). Besides style transfer techniques are based on these descriptors (Gatys, Ecker, & Bethge, 2016). The usual procedure in style transfer is computing the mean of different layer features and imposing them when a different image is passed through the network.

Here we used three different ways of summarizing the layer output. First, we can use the mean of each feature:

$$d^l_\mu(x_0, y_0) = \sqrt{\sum_{C^l} (\hat{\mu}^l_x - \hat{\mu}^l_y)^2} \tag{2}$$

where $\hat{\mu}^l_x \in \mathbb{R}^{C^l}$ and $\hat{\mu}^l_y \in \mathbb{R}^{C^l}$ are the spatial averages of the outputs of the lth layer for the image x_0 and its distorted version y_0 respectively. In this case, we compute the spatial averages to calculate the distance with them. Second, we can use not only the spatial averages but also the spatial standard deviations. In this case, we concatenate the spatial averages and standard deviation before computing the distance:

$$d^l_{\mu,\sigma}(x_0, y_0) = \sqrt{\sum_{C^l} (\hat{\mu}^l_x - \hat{\mu}^l_y)^2 + (\hat{\sigma}^l_x - \hat{\sigma}^l_y)^2} \tag{3}$$

where $\hat{\mu}^l_x \in \mathbb{R}^{C^l}$, $\hat{\mu}^l_y \in \mathbb{R}^{C^l}$ and $\hat{\sigma}^l_x \in \mathbb{R}^{C^l}$, $\hat{\sigma}^l_y \in \mathbb{R}^{C^l}$ are the spatial averages and standard deviations of the outputs of the lth layer for the image x_0 and its distorted version y_0 respectively. Finally, we can summarize the outputs through their Gram matrix:

$$d^l_G(x_0, y_0) = \sqrt{\sum_{C^l} (\hat{G}^l_x - \hat{G}^l_y)^2} \tag{4}$$

where $\hat{G}^l_x \in \mathbb{R}^{(C^l,C^l)}$ and $\hat{G}^l_y \in \mathbb{R}^{(C^l,C^l)}$ are the Gram matrices of the outputs at the lth layer for the image x_0 and its distorted version y_0.

Inspired by some of the state-of-the-art image quality models (LPIPS (Zhang et al., 2018)), one can also concatenate the outputs of different layers in order to introduce more information to calculate the distance. Besides that, it is also possible to weight (fine-tune) the output features of a model (concatenating or not) so that the correlation with a specific database is maximized.

2.3. Evaluation in perceptual databases

To test the models described previously we used two image quality databases: TID-2013 (Ponomarenko et al., 2015) and KADID-10K (Lin et al., 2019). Both consist of pairs of images and distorted versions of the same image with a mean opinion score (MOS) for each image-distorted image pair, which represents the distance between them as estimated by humans. More particularly, we used the whole TID-2013 (3000 image pairs) and 30% of KADID-10K, which we call val-KADID-10K (3038 image pairs), for testing the different models.

As a baseline, we used some image quality models. Some of them classical ones but highly used by the community such as SSIM (Wang et al., 2004), others based on neural networks which are currently state-of-the-art in image quality such us LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2020) and some that include some bio-inspired architecture such as PerceptNet (Hepburn et al., 2020). DISTS was trained to maximize the correlation using the KADID-10K database and PerceptNet was trained to maximize the correlation in TID-2008 (similar to TID-2013). Therefore they obtain high correlations at their

respective databases since they are overfitted to them. We used the implementation from PyTorch Image Quality (Kastryulin, Zakirov, & Prokopenko, 2019; Kastryulin, Zakirov, Prokopenko, & Dylov, 2022) with the default parameters.

Our procedure is to take both the reference image and its distorted version and pass both through each model. We extract the features of all the model layers and calculate the distance between the two images using one of the distance definitions we described in the previous section. Then we just calculate the Spearman correlation between that distance with the experimental MOS. With this procedure, we can obtain a correlation with human behavior for every layer of the models.

We also tested weighting the features extracted by the models (fine-tuning), following a similar strategy to LPIPS. We used TID-2008 (1700 image pairs) and the remaining 70% of KADID-10K (train-KADID-10K) (7087 image pairs) to weight each output feature in order to maximize the correlation with the MOS in these training databases.

Summing up, to obtain the results we pass each image-distorted image pair through the different models and record the features at different layers. Then, we calculate the distance between them at the different layers using one of the distance definitions from above (d^l-d^l_μ-$d^l_{\mu,\sigma}$-d^l_G; concatenating or not the outputs of different layers; weighting or not the outputs).

3. Experiments and results

The goal is to analyze how each isolated factor (function, architecture and environment) affects the perceptual behavior of the neural network models. Analyzing all the possible combinations would be unfeasible, on the one hand, because of the amount of models would be around one thousand. On the other hand, we are restricting ourselves to already trained models, and not all the combinations are available.

Therefore we followed the methodology stated in Table 1. In each experiment, we fix all the factors but one, using as baseline the best perceptual performance model, AlexNet trained on ImageNet-1K for classification, using Euclidean distance without concatenation. We represent this model with the same characteristics in all the plots (thicker solid dark blue line). We also use a colored background for the figures that highlight the different AlexNet (best perceptual performance model) blocks.

3.1. Function

First, we check how training a simple network for supervised vs. self-supervised goals modifies the correlation with human perception. Following previous reports on the human-like nature of different architectures, and also confirmed by our own empirical exploration shown below (in Section 3.2), simpler nets tend to be more human-like, both in many visual aspects (Akbarinia et al., 2023; Bertalmío et al., 2020; Gomez-Villa, Martín, et al., 2020; Li et al., 2022) and also in perceptual distances (Kumar et al., 2022; Zhang et al., 2018). Therefore, here we select AlexNet as the default simple architecture. From VISSL, a library for state-of-the-art self-supervised learning from images, we select all the models that use AlexNet architecture to analyze them. As training data, we chose ImageNet-1K, the way we perform the readout to calculate the distance is plain Euclidean, $d^l(x_0, y_0)$, so we did not weigh the features. In short, we calculate the Euclidean distance using the features of each layer for AlexNet trained for the different goals using the same data. Fig. 1 shows how AlexNet trained with different objectives correlates with human perception (MOS) at different layers for TID-2013 and val-KADID-10K.

While all training objectives have good perceptual properties, some have better properties than others. Most of the models achieve a higher correlation than SSIM. We obtain that, between the explored goals and with these perceptual databases, the supervised and the RotNet (self-supervised rotation invariance) models obtain the highest correlations. Between the other self-supervised tasks, Jigsaw obtains correlations at the level of the supervised model only in its first layers but, as depth increases, the correlation goes down. The colorization goal shows a small linear increase in correlation with layer depth but it always has a lower correlation than the supervised one. Finally, the DeepCluster model remains always almost at the same correlation level as in the RGB (input) domain, which is just the correlation with the RMSE.

3.1.1. Taskonomy

As stated in Section 2.1, in order to cover a wider range of tasks, we consider the models in *Taskonomy* (Zamir et al., 2018). We cannot compare them directly with the results from Fig. 1 because in this scenario the architecture is a modified ResNet-50 that has been trained with images from buildings so that the architecture and the training data are not the same. Each one of the 24 taskonomy goals has its own labels for the same images and has been trained using the needed decoder/classifier after the same ResNet-50 feature extractor.

Fig. 2 shows how the ResNet architecture trained with the same images but different goals correlates with human perception. Note that tasks are colored according to the different task clusters the *Taskonomy* (Zamir et al., 2018) authors discovered. In particular, they found five clusters named 3D, 2D, low dimensional geometric, semantic and denoising. Following them, we select the colors for each task according to this division: purple for semantic tasks, red for low dimensional geometric, blue for 2D, green for 3D and black for denoising tasks.

Although the training data and architecture are different from the previous results from Fig. 1, here we also found that semantic goals (such as semantic segmentation or scene classification) are the ones that obtain the highest correlation together with some of the 2D tasks. Specifically, they obtain higher correlations with human perception than the majority of the 3D and low-dimensional geometric training tasks. Also, the same trend we found in Fig. 1 can be seen here, i.e. jigsaw obtains high correlations in its first stages but then the correlation goes down with layer depth and in general classification stays at the top.

The results in our two experiments involving tasks suggest that supervised/semantic (e.g. classification or segmentation) goals lead to models with human-like perceptual properties in intermediate-last layers. This result is in concordance with other perceptual analyses such as the emergence of human-like color categories or the contrast sensitivity function, found in semantic visual tasks while 2D or low-level tasks fail to reproduce human-like color categories (Akbarinia, 2025; Akbarinia et al., 2023). However, it is important to highlight that it does not imply that the human brain relies solely on supervised goals but that the semantic ones have better perceptual correlations within the analyzed architectures, datasets and objectives. In the discussion section, we hypothesize of why we, among other works, are finding that semantic supervised models appear to have better perceptual properties.

Classical literature on image quality suggests that the human Contrast Sensitivity Function (CSF) is a major factor in predicting subjective distances (Malo, Pons, & Artigas, 1997; Watson, 1993; Watson & Malo, 2002). Moreover, classical functional explanations of the CSF have been attached to signal denoising and deblurring in autoencoder-like settings (Atick, Li, & Redlich, 1992; Atick & Redlich, 1992). Therefore, the poor correlation of denoising autoencoders in the Taskonomy experiment (Fig. 2) seems to contradict those classical results. However, note that more recent studies with deep nets have stressed the relevance of shallow architectures in the emergence of the CSFs in autoencoders (Li et al., 2022), and CSF-like bandwidths also emerge in higher abstraction tasks such as classification (Akbarinia et al., 2023). This apparent contradiction in Fig. 2 for denoising autoencoders may be just an interaction between task and architecture, due to the relatively deep nature of ResNet50, or to the fact that the CSF is important, but not the only factor behind subjective image quality (Laparra, Muñoz Marí, & Malo, 2010; Malo et al., 1997).

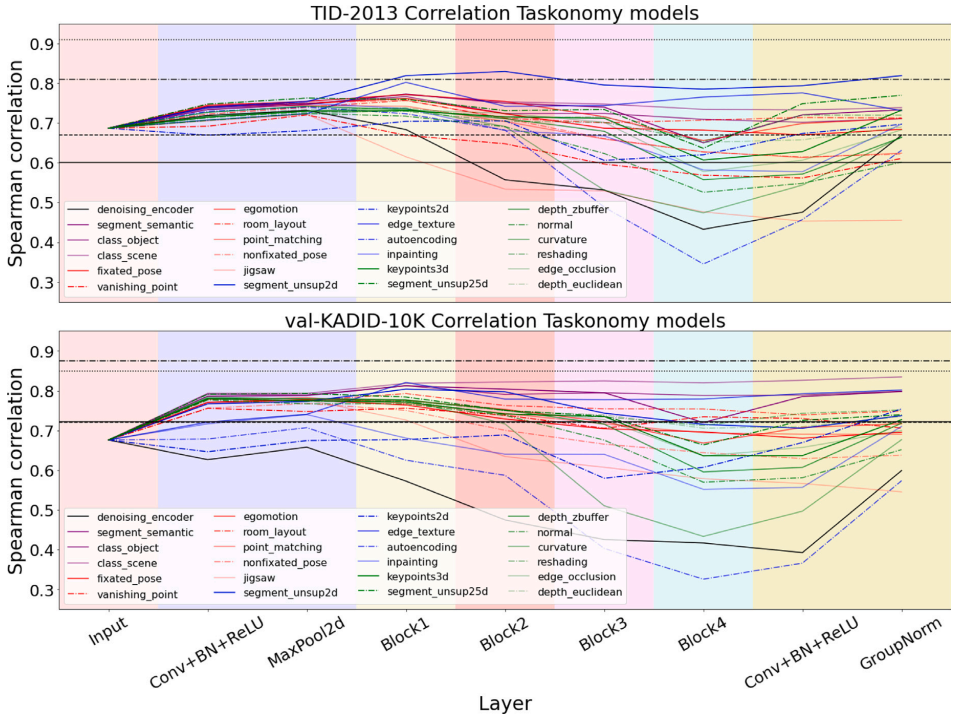


Fig. 2. Function and task 2: Taskonomy (ResNet50). Plots display the Spearman correlation with human opinion at different layers of ResNet trained for *different tasks* from Taskonomy. We consider the similarity with humans in the TID-2013 database (left) and in the val-KADID-10K database (right). Background colors represent the different ResNet blocks.

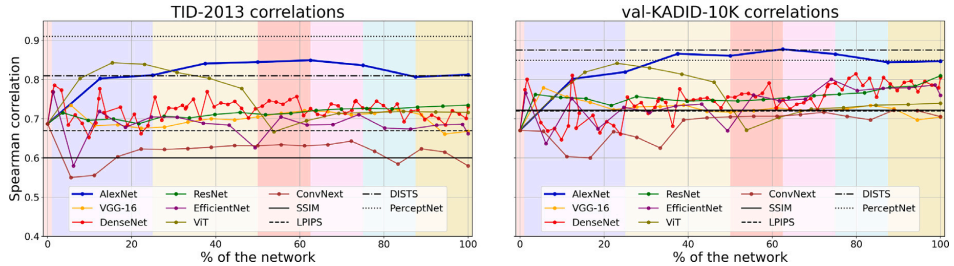


Fig. 3. Architecture 1: connections. Plots display the Spearman correlation with human opinion at different depths of networks with *different connections* trained for the same computer vision task. Note that each model has a different number of layers so that to plot altogether, the abscissa represents the percentage of the network (0% being the *retina* and 100% being the deepest layer). We consider the similarity with humans in the TID-2013 database (left) and in the val-KADID-10K database (right). Background colors represent the different areas of AlexNet. As in the previous figure, the performance of some standard image quality measures in these databases is shown in black for convenient reference.

3.2. Architecture 1: connections

Secondly, we tested how different architectures (AlexNet, VGG-16, ResNet-50, DenseNet-121, EfficientNet-B0, ConvNext and ViT) correlate with human perception. We fixed the goal of the models (ImageNet supervised classification task), the training data (ImageNet-1K), the way we perform the read-out to calculate the distance (euclidean: $d^l(x_0, y_0)$) and we did not weight the output features, so we just calculate the euclidean distance using the features of each layer of the networks. Fig. 3 shows how different architectures correlate with human perception (MOS) at different depths (different layers) for TID-2013 and val-KADID-10K.

There are several results to notice from this figure. First, most of the models perform better than classical statistical image quality models (SSIM). Also, in the TID-2013 database simpler models (AlexNet and VGG-16) perform better than modern image quality algorithms based on neural networks (DISTS and LPIPS) and only some biologically inspired image quality algorithm (PerceptNet) trained specifically for a very similar database performs better than the majority of the models. In the KADID-10K database, AlexNet and VGG-16 also perform better than modern image quality algorithms based on neural networks (LPIPS) and on pair to other state-of-the-art models stunned specifically on this database (DISTS). Second, simpler models achieve higher correlations. Specifically, AlexNet and VGG-16, get a bigger correlation

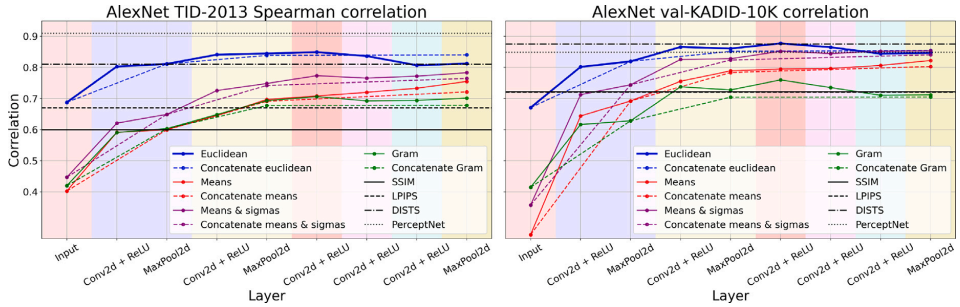


Fig. 4. Architecture 2: feature descriptors and readout strategies. Plots display the Spearman correlation with human opinion at different depths of AlexNet with *different readout strategies* trained for the same computer vision task. The different options to *read-out the distance* include are: the (plain) Euclidean distance between the original and the distorted features, using a statistical descriptor of the features and computing the difference between descriptors and/or concatenating the outputs of different layers. We consider the similarity with humans in the TID-2013 database (left) and in the val-KADID-10K database (right). Background colors represent the different areas of AlexNet. The performance of some standard image quality measures in these databases is shown in black for convenient reference.

with human perception than modern models with higher ImageNet accuracies (see Fig. 7). Although it is a completely different task and therefore is not comparable, other words found that simpler networks work better such as in style transfer where the authors from (Wang, Li, & Vasconcelos, 2021) found that using more complex models with skip connections got worse results than simpler models such as VGG. Third, in the simplest models (AlexNet and VGG-16), there is a relationship between the depth of the layer used to measure image distances and the correlation with perception, with higher correlations obtained in deeper layers. However, both models show a decrease in correlation for the last two layers. More complicated models (ResNet-50, DenseNet, EfficientNet and specially ConvNext and ViT) have much more complex correlation diagrams with depth, not showing a clear relation between deepness and correlation, except for the ViT which shows a high correlation in the started layers and the correlation goes down with the layer depth.

This result suggests that image quality algorithms based on deep learning models should use classical networks without skip connections even if they do not achieve high accuracies in ImageNet classification.

3.3. Architecture 2: feature descriptors and readout

Next, we tested how the way the read-out is defined affects the correlation. To do so, we use the supervised AlexNet model trained in a supervised way with ImageNet-1K and we calculate the correlation with human perception using the different distance definitions, described in Section 2, using the features of each layer of the model. We tested not only the different distance definitions (summarizing the layer outputs with statistics or not) but also checked what happens when we concatenate the outputs of the three max pooling layers. Fig. 4 shows how different distance measurements correlate with human perception (MOS) at different layers for TID-2013 and val-KADID-10K.

The best correlation is obtained when the whole output is used to calculate the distances, without the use of any statistical descriptor. When statistical descriptors are used, using the spatial Gram Matrix performs worst. Using the spatial mean together with the spatial standard deviation leads to a higher correlation than using only the spatial means because it implies using more information to calculate the distances. With regard to concatenating different layer outputs (such as in LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2020)), it does not have a big effect. However, by doing it we can obtain an upper limit of the correlation and it improves the results of the last layer, when if one does not concatenate the previous layer features, the correlation goes down.

This result suggests that image quality algorithms based on deep learning models should use the full output of the layers without using any statistical descriptor. Also, using a concatenation of the outputs of different layers seems to have no important benefits while increasing the computational complexity.

3.4. Learning environment and data statistics

Finally, we check how human perception correlates with distances from networks trained with different data. We fixed the architecture that obtained a better correlation (AlexNet) and the training objective, supervised. We also keep fixed the way we perform the read-out to calculate the distance (euclidean, $d^l(x_0, y_0)$) and we did not weigh the features. The only difference between them is the data used to train the models, and we analyze how the correlation depends on training on ImageNet-1k, Places-365, and Cifar-10. Fig. 5 shows how different training data correlate with human perception (MOS) at different layers for TID-2013 and val-KADID-10K.

The best result is obtained with ImageNet-1K, which is around 1 million natural images. Places-365 (10 million place locations images) and Cifar-10 (50K small natural images) achieve less correlation.

It suggests that, as expected, image quality algorithms based on deep learning models should use as many natural and big images as possible.

4. Discussion

After the performed analysis new questions arise. On the one hand, one wonders if using no perceptual data one could obtain a model that performs well in a perceptual task. In our analysis, one of the most correlated goals with perception has been supervised classification. Also, it is the goal that is used to evaluate self-supervised methods. Therefore, we select this goal as a proxy for perception. On the other hand, one wonders if using a simple model as AlexNet trained for classification as a feature extractor could be enough to get the best model to measure distances between images in a perceptual fashion. We fine-tune this model with perceptual data and compare it with state-of-the-art measures. Moreover, it is interesting to see if new multimodal objectives affect in any way the human visual alignment, for example comparing between models trained for supervised classification or with supervised contrastive vision-language goals, such as CLIP (Radford et al., 2021).

4.1. Does multimodal vision-language training improve human visual alignment?

Currently, multimodal training objectives are becoming increasingly important since the release of the CLIP model, trained for contrastive

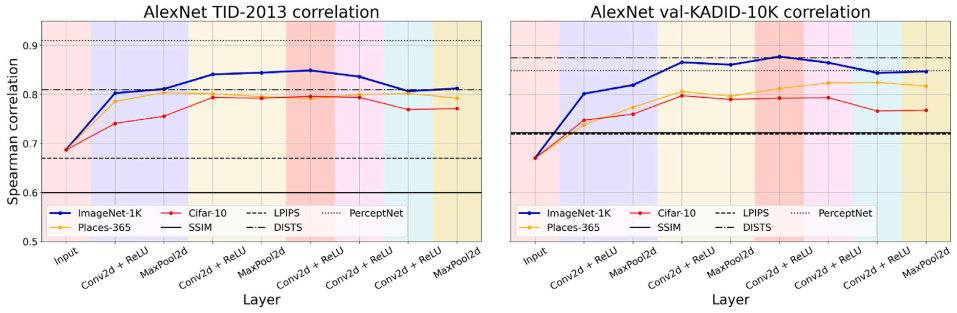


Fig. 5. Environment statistics: Plots display the Spearman correlation with human opinion at different depths of AlexNet with training data of eventually *different statistics* trained for the same computer vision task. We consider databases of different sizes and image natures. We consider the similarity with humans in the TID-2013 database (left) and in the val-KADID-10K database (right). Background colors represent the different areas of AlexNet. The performance of some standard image quality measures in these databases is shown in black for convenient reference.

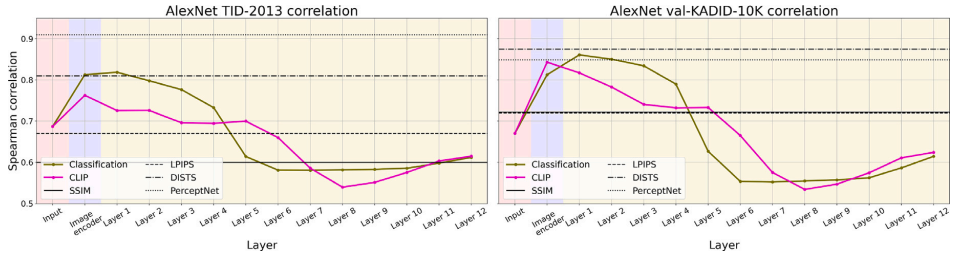


Fig. 6. Multimodal language training (CLIP): Plots display the Spearman correlation with human opinion at different depths of ViT trained for two different goals: image classification and contrastive multimodal image-language objectives (CLIP). We consider the similarity with humans in the TID-2013 database (left) and in the val-KADID-10K database (right). Background colors represent the different areas of ViT. The performance of some standard image quality measures in these databases is shown in black for convenient reference.

image-language objectives (Radford et al., 2021). It has been shown to learn powerful image representations that are useful for zero-shot tasks such as image classification, where it achieves almost state-of-the-art results. However, how this multimodal training affects human visual alignment is not analyzed.

Fig. 6 shows how human perception correlates with the distances between the different layers of a Vision Transformer (ViT). This model has been trained for both image classification and multimodal contrastive image-language tasks, i.e. CLIP. Results show that, for this specific model architecture there are no major differences between both training objectives. However, in the early layers, the model trained for image classification gets a slightly higher correlation. Then, both models show a decrease in correlation in their middle layers and end with really similar correlations in the latest layers. However, the correlation drop is higher in the model trained for supervised image classification.

4.2. Purely functional arguments do not explain perceptual distances

As an illustrative summary of all the considered factors, Fig. 7 (left) plots the Spearman correlation of the networks with human opinion in terms of a common measure of performance. This plot includes all the considered models trained for different *functions*, using different *architectures*, and trained in different *environments*. The y-axis value for each model corresponds to the correlation represented the last layer of our previous plots. The x-axis, ImageNet validation classification accuracy, is a uniform way to quantify the performance in the variety of considered models. There are two main reasons to use this. First, it is one of the most correlated functions with perception in our analysis. Second, it is a widely used measure to evaluate different self-supervised tasks in a common way (Caron et al., 2018; Goyal et al., 2021; Komodakis & Gidaris, 2018; Noroozi & Favaro, 2016; Zamir et al., 2018; Zhang et al., 2016). Therefore, the performance is measured in *direct* and *indirect* ways. In the cases where the function is classification, the interpretation of the x-axis is *direct*, i.e. it corresponds with the model Top-1 accuracy in ImageNet in validation. For other functions, the quantification of the performance is *indirect*: following the self-supervised approach, we assess the success in other visual tasks through its positive impact on classification. For instance, rotation prediction is good if it helps to extract good features for classification. Therefore, the reported performance for a non-classification goal is the performance of a classifier optimized on top of the given representation. In this way, all the functions are evaluated in a uniform way. Particularly, we take the accuracy values from the sources described in Appendix A.

In light of Fig. 7 (left), there is a nontrivial relation between the model performance and the correlation with humans: where very low or very high performances bring a low correlation with human perception, and the best point in correlation terms seems to be on intermediate accuracy. Improvements in the performance (e.g. by using deeper models) do not always lead to more human-like distances. Therefore, optimization of certain *function* is not the only explanation of the human non-Euclidean metric.

These results on the nontrivial (non-monotonic, inverted-U) interaction between the *function* of the nets with other design parameters (mainly the *architecture* and readout strategies) confirm and expand the results in Kumar et al. (2022) and Li et al. (2022) shown for reference in Fig. 7 (center), and (right) respectively. These related works study the emergence of human-like distances (Kumar et al., 2022) and the emergence of human-like CSFs (Li et al., 2022). We considered 29

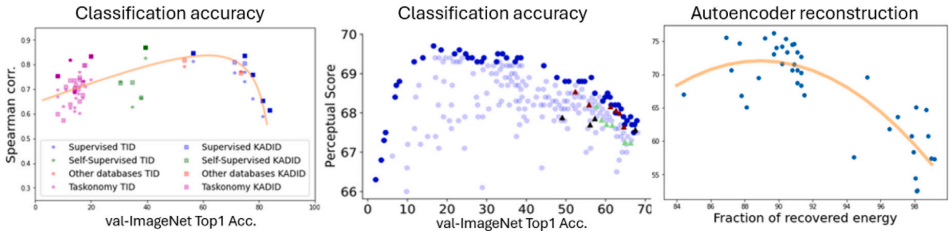


Fig. 7. Nontrivial (inverted-U) relation between human behavior and classification accuracy on Imagenet validation set., *Left: our results* correlation with the human opinion of distortion (both in TID-2013 and KADID-10K) depending on the validation ImageNet classification accuracy for all the different considered networks. The points of the Pareto frontier have been highlighted in darker color. *Center: result in Kumar et al. (2022)* Pareto frontier showing an equivalent correlation between the validation ImageNet classification accuracy and the human behavior (reproduced with permission of the authors), and *Right: results in Li et al. (2022)* similarity with human Contrast Sensitivity Function (CSF) of autoencoders depending on the performance in several image reconstruction tasks (from data in the tables 1-4 of Li et al. (2022)).

different functions while (Kumar et al., 2022) is restricted to classification, and Li et al. (2022) only considers 5 unsupervised functions. Our results suggest a general trend: for *many different* functions, better performance does not necessarily mean more similarity with humans. This is consistent with the information-theoretic considerations made in Sucholutsky and Griffiths (2023) on finding representations robust to few-shot learning.

This result raises an important question: why is the highest correlation with human perception observed at intermediate levels of classification accuracy? One potential explanation is the overfitting of networks to specific tasks and datasets. Supporting this hypothesis, prior works (Kumar et al., 2022; Schrimpf et al., 2018) demonstrated that classification models achieve higher correlations with human perception during intermediate stages of training and become inversely related with neural predictivity once their classification accuracy becomes to high. At this point, the network has likely learned generalizable and useful representations but has not yet become overly specialized to the task or dataset. This suggests that intermediate accuracy may reflect a balance between general feature learning and task-specific optimization, which aligns more closely with the features humans use for visual perception.

4.3. Alexnet features as the best perceptual representation?

After our dissection of the different factors involved in the emergence of a human-like distance, we found that the best model is the AlexNet architecture trained on ImageNet for classification, calculating the distance between the responses with the Euclidean metric. It is important to note that no perceptual data has been used so far for anything other than evaluating the models. However, usually, the state-of-the-art models in image quality do use perceptual data to adjust the parameters. More specifically, some models are trained to directly maximize the correlation with human perception in some perceptual databases. Then, our next step is to use perceptual data to fine-tune the best model we found in a similar way to what LPIPS or DISTS do. We fine-tune the feature relevance of the outputs that we use to calculate the distance by adding between 97 and 385 parameters (depending on which layer we consider, one weight per output feature channel plus a bias). This implies finding a non-Euclidean diagonal metric as opposed to the identity matrix. We found that metric (weights, θ_C, for each channel) by maximizing the correlation with some perceptual databases:

$$\max_{\theta} \rho_s \left(\text{MOS}, \theta_0 + \sum_{C^l} \theta_{C^l} \cdot \sqrt{\sum_{H^l, W^l} (x^l - y^l)^2} \right) \tag{5}$$

We made two versions of the model, each one trained in one database, TID-2008 or train-KADID-10K. We tested them in the same way we did in all the previous experiments (Fig. 8).

The first conclusion is that fine-tuning with TID-2008 when evaluating with TID-2013 results in an increase in the correlation probably because the images are similar. It also occurs when fine-tuning with train-KADID-10K and evaluating with val-KADID-10K. More interesting results happened in cross fine-tuning: fine-tuning with train-KADID-10K gives no substantial changes when evaluating with TID-2013. However, fine-tuning with TID-2008 gives much worse results when evaluating in val-KADID-10K. This is consistent with the fact that KADID-10K shows more variability than TID-2013 and thus can be taken as a more general database to be used. More interestingly, the model that performs fine-tuning with train-KADID-10K obtains better results than state-of-the-art perceptual metrics (LPIPS and DISTS) both in TID-2013 and in val-KADID-10K, even though DISTS was trained to maximize the correlation with this last database. Only a bio-inspired perceptual metric (Hepburn et al., 2020), which was trained for a similar database to TID-2013 can surpass by a small margin the AlexNet model with fine-tuning with train-KADID-10K when evaluating in the same TID-2013.

4.4. AlexNet perceptual analysis

In this section, we analyze in deeper detail the performance of AlexNet (trained for classification in ImageNet) when used to measure human-like distances.

First, we consider the effect of the distortion type. To do that, we calculate how the correlation given by AlexNet changes depending on the distortion type for all the TID-2013 images. Fig. 9 shows how the correlation changes layer by layer for the 24 different TID-2013 distortions. Note that all the plots have the same y-scale for an easy comparison. It shows that the majority of the distortions are extremely well correlated by the network distances and for all the layers, i.e. their correlation curves are almost flat and close to a correlation of one. Only two of the distortions do not have a high correlation, i.e. the network does not predict well the distances: the local block-wise and contrast change distortions.

There are two reasons for this apparently great performance. On the one hand, correlation on separated problems may be artificially high, but the model is unable to explain all the situations at the same time so the aggregated performance drops. On the other hand, the energy of the distortions in the database comes in five different levels which are clearly ranked from less-to-more visible for the observers. This implies that when considering distortion-wise partitions of the data it is easy to predict the human ranking of those distortion levels. This leads to high Spearman correlations even in the input domain.

Next, we analyze the effect of the different images. To do that we compute how the correlation given by AlexNet changes depending on the original image for all the TID-2013 images. Fig. 10 shows how the correlation changes layer by layer for the 25 different TID-2013 original images. As in the previous figure the y-scale is shared

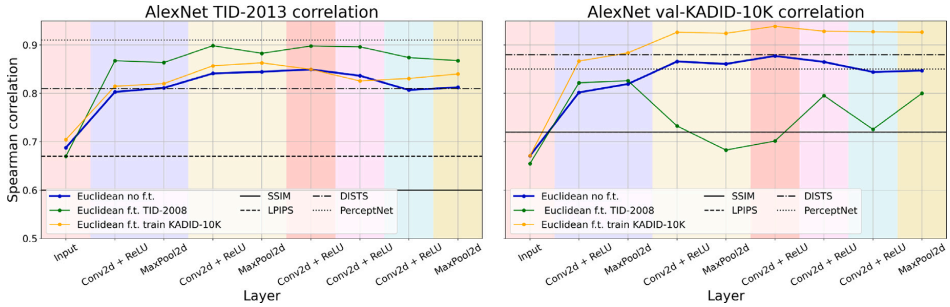


Fig. 8. Architecture 3: fine-tuning or internal metric. Plots display the Spearman correlation with human opinion at different depths of AlexNet with *different fine-tuning strategies*. The different explored options for *internal metric* include: no fine-tuning (or plain Euclidean metric), and fine-tuned metrics for two different databases. We consider the similarity with humans in the TID-2013 database (left) and in the val-KADID-10K database (right). Background colors represent the different areas of AlexNet. The performance of some standard image quality measures in these databases is shown in black for convenient reference.

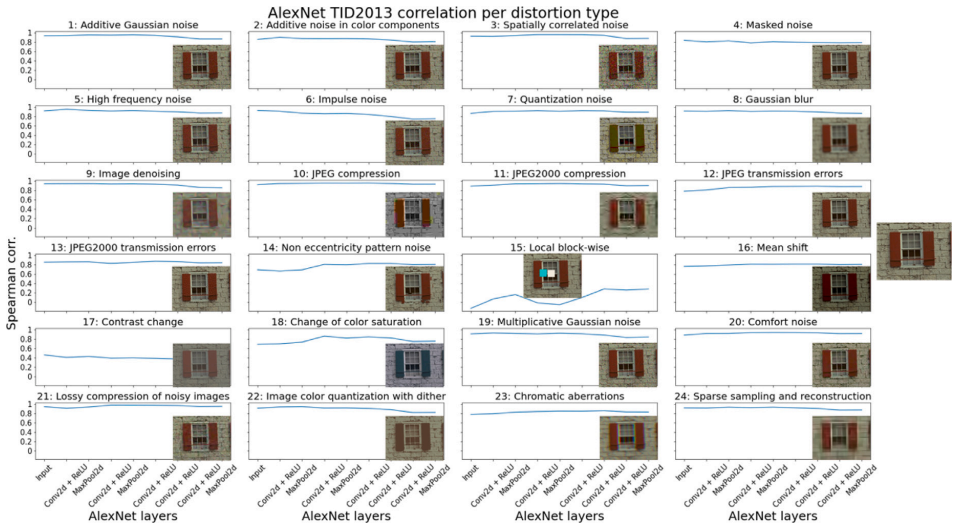


Fig. 9. AlexNet analysis: TID-2013 correlation per distortion. Plots display an example of each TID2013 distortion type applied on one of the reference images (shown at the right) and the Spearman correlation with human opinion at different depths of AlexNet for each of the distortions.

between all the plots. Note that all the images get a low correlation with human distances if the comparison is made in the image space, i.e. just doing RMSE between images. Interestingly, just entering into the network and pass through only one layer already helps to get a high increase in correlation for almost all the images. And, in general, correlations increase with layer depth. There are some interesting cases, such as image number 2 that has almost perfect correlation for all the layers and, in the opposite scenario, image 13 never has more than 0.8 correlation. If we focus on these two images, image number 2 corresponds with an image of a door that has large smooth regions and image number 13 corresponds to a mountain and forest landscape with a lot of textures and high contrast details. Therefore, human opinion in smooth images with large flat areas seems easier to predict for the network than in images with high frequency and high contrast details. Appendix B gives visual illustrations of these facts.

5. Conclusions

In this work, we explore the eventual human-like behavior (or perceptual properties) of deep learning models devoted to computer vision tasks. In particular, we do it through the analysis of the correlation between image distances computed using artificial network features and human subjective distances. We compare the distances inferred from multiple deep learning models along three main design factors, such as *function*, *architecture*, and *environment*, exploring a total of 46 different design conditions (see Table 1). To this end, we use two large image quality databases accepted by the image quality community (Lin et al., 2019; Ponomarenko et al., 2015). We restrict ourselves to off-the-shelf pre-trained models (see *Methods*) to discard bias on architecture design or training procedures. We got the following conclusions:

- **Function:** While all the visual tasks (which could be sensible organization principles for human vision) have good perceptual properties (most have better correlation with humans than SSIM), some have better properties than others. In particular, within the analyzed objectives the best results are obtained by the models trained for supervised goals.
- **Architecture I. Connections:** Again, almost all the considered architectures lead to better correlation than SSIM at some layer. However, note that for architectures with different connectivity

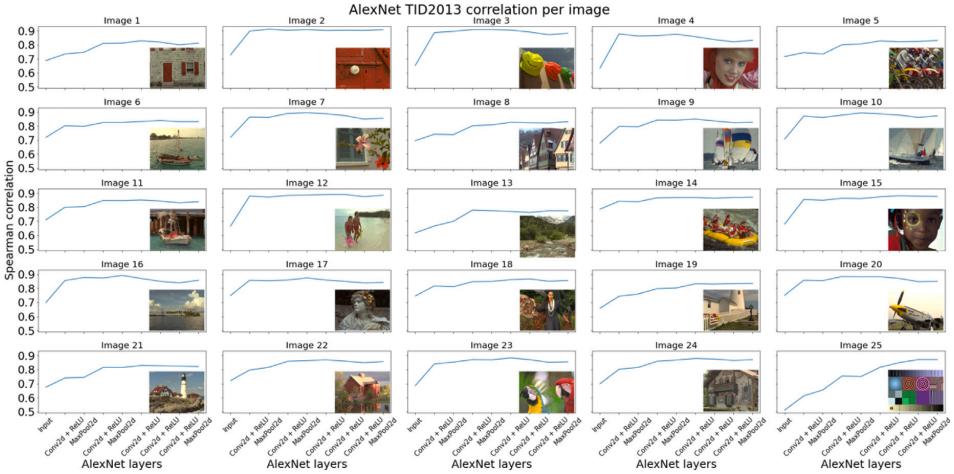


Fig. 10. AlexNet analysis: TID-2013 correlation per image. Plots display different original images from TID-2013 and the Spearman correlation with human opinion at different depths of AlexNet for each of the images.

(blue dots in Fig. 7-left), there is a nontrivial relation between function and architecture: for these points, the correlation between performance and human opinion is negative (for the fixed functional goal of classification). Simpler models have better perceptual behavior than complex models and are better than state-of-the-art image quality metrics such as LPIPS and DISTS using no perceptual data. Also, for these simpler models, there is a direct relation between layer depth and correlation with human vision except for the last layers.

- **Architecture II. Feature description and readout:** Concatenating outputs of different layers (as in LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2020)) or just taking the output of one layer does not have a big effect on the correlation with humans. The use of statistical descriptors such as the channel means and standard deviations, or the Gram matrices (as proposed for texture invariance in style transfer (Gatys, Ecker, & Bethge, 2015)) degrades the correlation with humans.
- **Environment statistics:** Training the models with enough big natural images leads to better results than using smaller/fewer images or using non-natural images with different statistics. The use of a wider class of images (images of diverse nature as in ImageNet as opposed to Places) improves a bit the correlation with humans.

In particular, the best correlation with human opinion is obtained by the fifth layer (not the last layer) of the original AlexNet model trained for supervised classification, with a plain Euclidean metric, no concatenation of layers (no sophisticated read-out from all layers), and with no use of style-transfer-like statistical summaries intended for spatial invariance. To us, this was a surprising result, given the simplicity of that network.

Also, fine-tuning the feature relevance for a particular image quality database with high variability in its images leads to an increase in perceptual correlation. In fact, just maximizing the correlation with AlexNet best model output, we obtain results above state-of-the-art perceptual quality models such us LPIPS or DISTS. Moreover, we analyze which images and distortions AlexNet get the best and worst correlations. We found that AlexNet fails at predicting human-like distances for the local block-wise and contrast changes. Regarding the image type, we found that high-frequency and high-contrast textures are a problem for the network. Both facts suggest that AlexNet does not reproduce the contrast masking phenomenon.

Interestingly, the above results (on function and architecture) confirm some of the findings in papers related to (a) other properties of human vision as for instance *visual illusions* (Bertalmío et al., 2020; Gomez-Villa, Martín, et al., 2020) *contrast sensitivity* (Akbarinia et al., 2023; Li et al., 2022), and *pattern masking* (Martinez-Garcia et al., 2019), and (b) in papers specifically related to perceptual distances (Kumar et al., 2022; Zhang et al., 2018): as in the cited literature, the function-architecture interaction is not trivial, which complicates the interpretation of the goal function as an organizing principle. Specifically, here we also find an inverted-U shape in the reproduction of human behavior as a function of the performance in the task (see Fig. 7). On the other hand, it is intriguing that almost all the considered nets overperform the classic perceptual measure SSIM (Wang et al., 2004) (a *de-facto* standard), so the connection between the explored nets and human behavior is consistently true. However, in the explored nets we also find differences with (Ding et al., 2020; Kumar et al., 2022) in the impact produced by statistical summaries. Which, in principle, are intended to introduce invariance to non-relevant distortions.

Lack of an increasing relation between performance in the task and human behavior, and nontrivial interaction of factors implies that (1) consistently with Poggio (Poggio, 2021) and others (Malo & Hernandez-Camara, 2024), function and architecture are not as separable as thought before, and (2) as opposed to an extended practice in metric learning (Shakhnarovich et al., 2011), modeling biological vision cannot be considered as a pure regression problem to be solved with whatever regression tool regardless of the architecture of this tool. Instead, as pointed out in Martinez-Garcia et al. (2019), Rust and Movshon (2005) the use of biologically sensible architectures is key for a proper explanation of the problem, and why these specific architectures actually emerged is still open to debate (Sterling & Laughlin, 2015).

Finally, our work opens new questions that can be carefully analyzed in future research. On the one hand, one of our key observations is that simpler networks correlate better with human perception than deeper networks. However, simplicity in network design can manifest in multiple ways, such as reducing the number of layers, removing skip connections, or decreasing the number of neurons per layer. We found that networks without residual connections correlated better with human perception, but these networks also tended to be shallower. This raises an open question: which aspect of simplicity – removing residual connections, reducing depth, or decreasing width – plays the

most significant role in aligning artificial networks with human perception? Previous research (Kumar et al., 2022) suggested that both network depth and width can negatively impact human alignment, but how these factors interact remains unclear. Future research that retrains different deep networks should explore how these architectural components interact and contribute to human alignment, providing deeper insights into designing networks that better capture human-like processing.

On the other hand, our findings raise an intriguing question about the relationship between human perception and the training paradigms of artificial neural networks. Specifically, our results, along with findings from other studies (Akbarinia, 2025; Akbarinia, Morgenstern, & Gegenfurtner, 2021; Akbarinia et al., 2023) observed that supervised networks are more human-align than compared to networks trained with self-supervised or unsupervised tasks. This seems to challenge the current view of the computer vision community, which usually adds a pre-train stage using self-supervised or unsupervised learning. The stronger correlation observed between supervised classification networks and human perception could arise because supervised tasks take advantage of the same statistical regularities that drive human visual learning. For example, accurately classifying objects requires learning features such as shape, color, and texture, which are also fundamental to human perception. In this sense, supervised training on high-level tasks may result in internal representations that resemble those learned through self-supervised or unsupervised processes in humans. Additionally, current self-supervised and unsupervised models may not fully capture the richness of human learning, which involves complex, multimodal interactions with the world. In contrast, neural networks are typically trained on static datasets, which may limit their alignment with biological perception. Importantly, the observed advantage of supervised networks may also reflect task and dataset dependence. Future work could explore more on the task-dependence of the human perception alignment to better understand the relationship between different learning paradigms and human vision.

CRediT authorship contribution statement

Pablo Hernández-Cámara: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Investigation, Formal analysis, Data curation, Conceptualization. **Jorge Vila-Tomás:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Data curation, Conceptualization. **Valero Laparra:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Jesús Malo:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by MICIIN/FEDER/UE under Grants PID2020118071GB-I00, PDC2021-121522-C21 (funded by MCIN/AEI/10.13039/501100011033 and the EU NextGenerationEU/PRTR) and Grant PID2023-152133NB-I00; in part by Spanish MIU under Grant FPU21/02256; and in part by Generalitat Valenciana, Spain under Projects GV/2021/074, CIPROM/2021/056 and CIAPOT/2021/9. The authors gratefully acknowledge the computer resources at Artemisa and the technical support provided by the European Union through the 2014–2020 ERDF Operative Programme of Comunitat Valenciana, project IDIFEDER/2018/048.

Appendix A

Table A.2 includes a summary of all the tested models, the way they have been trained, information about from where we downloaded them and their Top-1 ImageNet accuracy (except for the models trained in Cifar-10 and Places-365 in which we report their accuracy in their corresponding training database). To obtain the accuracy for the models that have not been trained for ImageNet classification, a linear classifier is added after the model and trained to classify the ImageNet database.

Appendix B

We show some visual illustrations of the analysis performed in Section 4.4. To do that, we select the AlexNet layer that gets the maximum TID-2013 correlation, the third convolution. Then, we pass through the network all the image pairs for image 2 and image 13, which we found are easy and difficult images.

Clearly, results from image 2 correlate much better than results from image 13, as stated by their correlations and shown in Fig. B.11. Interestingly, we can select some points that illustrate the failures and successes of the model. On the one hand, failure is represented by images with similar model distances but very different human distances (MOS), i.e. two points where the network completely fails to see the distortions in a human-like way. Fig. B.12 shows an example from image 13. On the other hand, success is represented by pairs of images where the model predicts their distances in a human-like way, i.e. similar human distances (MOS) get similar model distances. Fig. B.13 shows this behavior for two points from image 2.

Table A.2
Summary of tested models, their training goal, their source and their Top 1 ImageNet classification accuracy (except for the models trained in Cifar-10 and Places-365 in which we report their accuracy in their corresponding training database).

n	Architecture	Training goal	Source	Top 1 class. accuracy
1	AlexNet	RotNet	Facebook VISSL[a]	ImageNet: 39.5%
2	AlexNet	Jigsaw	Facebook VISSL[a]	ImageNet: 34.8%
3	AlexNet	Colorization	Facebook VISSL[a]	ImageNet: 30.4%
4	AlexNet	DeepCluster	Facebook VISSL[a]	ImageNet: 37.9%
5	AlexNet	ImageNet class.	TorchVision[b]	ImageNet: 56.5%
6	VGG-16	ImageNet class.	TorchVision[b]	ImageNet: 71.3%
7	DenseNet-121	ImageNet class.	TorchVision[b]	ImageNet: 74.7%
8	ResNet-50	ImageNet class.	TorchVision[b]	ImageNet: 74.9%
9	EfficientNet-B0	ImageNet class.	TorchVision[b]	ImageNet: 77.7%
10	ConvNeXt-Tiny	ImageNet class.	TorchVision[b]	ImageNet: 81.3%
11	ViT-b16–224	ImageNet class.	TorchVision[b]	ImageNet: 83.8%

(*continued on next page*)

Table A.2 (*continued*).

n	Architecture	Training goal	Source	Top 1 class. accuracy
12	ViT-b16–224	Multimodal CLIP	HuggingFace[f]	ImageNet: 76.8%
13	AlexNet	Places class.	MIT CSAIL[c]	Places-365: 53.2%
14	AlexNet	Cifar-10 class.	Raschka DL Models[d]	Cifar-10: 73.5%
15	ResNet-50	Object class.	Taskonomy[e]	ImageNet: 19.0%
16	ResNet-50	Scene class.	Taskonomy[e]	ImageNet: 20.0%
17	ResNet-50	Semantic segment.	Taskonomy[e]	ImageNet: 17.5%
18	ResNet-50	Curvature	Taskonomy[e]	ImageNet: 17.0%
19	ResNet-50	3D Key-points	Taskonomy[e]	ImageNet: 16.5%
20	ResNet-50	Occlusion edges	Taskonomy[e]	ImageNet: 15.5%
21	ResNet-50	Point match	Taskonomy[e]	ImageNet: 14.0%
22	ResNet-50	2.5D Segment.	Taskonomy[e]	ImageNet: 16.5%
23	ResNet-50	Distance estim.	Taskonomy[e]	ImageNet: 14.0%
24	ResNet-50	Cam pose (fix)	Taskonomy[e]	ImageNet: 14.0%
25	ResNet-50	Colorization	Taskonomy[e]	ImageNet: 13.0%
26	ResNet-50	Normals	Taskonomy[e]	ImageNet: 17.0%
27	ResNet-50	Layout	Taskonomy[e]	ImageNet: 16.5%
28	ResNet-50	Cam pose (non fix)	Taskonomy[e]	ImageNet: 12.5%
29	ResNet-50	2D Segment.	Taskonomy[e]	ImageNet: 12.5%
30	ResNet-50	Vanishing Pts.	Taskonomy[e]	ImageNet: 14.0%
31	ResNet-50	Denoising	Taskonomy[e]	ImageNet: 7.5%
32	ResNet-50	In-painting	Taskonomy[e]	ImageNet: 10.0%
33	ResNet-50	2D Key-points	Taskonomy[e]	ImageNet: 7.5%
34	ResNet-50	Auto-encoding	Taskonomy[e]	ImageNet: 9.0%
35	ResNet-50	Z-Depth	Taskonomy[e]	ImageNet: 14.0%
36	ResNet-50	2D Edges	Taskonomy[e]	ImageNet: 10.0%
37	ResNet-50	Scratch	Taskonomy[e]	ImageNet: 6.5%

[a] https://github.com/facebookresearch/vissl
[b] https://pytorch.org/vision/stable/models.html
[c] https://github.com/CSAILVision
[d] https://github.com/rasbt/deeplearning-models
[e] http://taskonomy.stanford.edu/
[f] https://huggingface.co/models

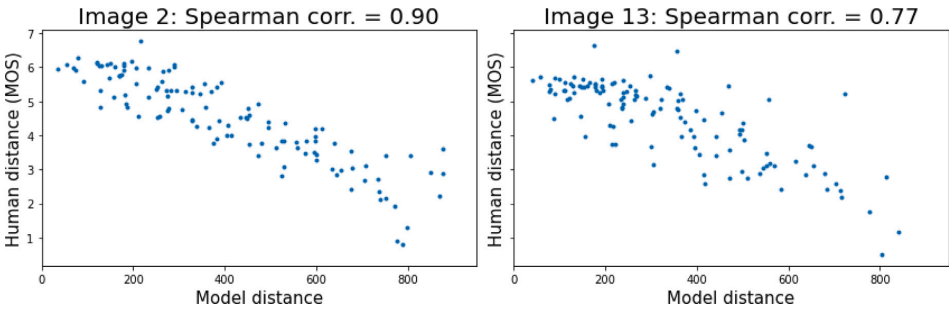


Fig. B.11. AlexNet third convolution results: TID-2013 images 2 and 13. Plots display the human distances and AlexNet third convolution distances for all the image distortions for images 2 and 13 from TID-2013.

Fig. B.12. Example of AlexNet failure. Two distorted versions (left and right) from original image number 13 (center) have similar model distances but completely different human distances.

Fig. B.13. Example of AlexNet success. Two distorted versions (left and right) from original image number 2 (center) have similar model distances and similar human distances.

Data availability

The datasets analyzed during the current study are available from the following links:

KADID-10k (Reference data)
TID2013 (Reference data)

References

(2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, *16*(7), 974–981.
Akbarinia, Arash (2025). Exploring the categorical nature of colour perception: Insights from artificial networks. *Neural Networks*, *181*, Article 106758.
Akbarinia, Arash, Morgenstern, Yaniv, & Gegenfurtner, Karl R. (2021). Contrast sensitivity is formed by visual experience and task demands. *Journal of Vision*, *21*(9), 1996–1996.
Akbarinia, Arash, Morgenstern, Yaniv, & Gegenfurtner, Karl R. (2023). Contrast sensitivity function in deep networks. *Neural Networks*, *164*, 228–244.
Atick, J. J., Li, Z., & Redlich, A. N. (1992). Understanding retinal color coding from first principles. *Neural Computation*, *4*(4), 559–572.
Atick, J. J., & Redlich, A. N. (1992). What does the retina know about natural scenes? *Neural Computation*, *4*(2), 196–210.
Barlow, Horace (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In *NPL Symposium on the Mechanization of Thought Process: vol. 10*, (pp. 535–539).
Barlow, Horace (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, *12*(3), 241.
Bertalmío, Marcelo, Gomez-Villa, Alex, Martín, Adrián, Vazquez-Corral, Javier, Kane, David, & Malo, Jesús (2020). Evidence for the intrinsically nonlinear nature of receptive fields in vision. *Scientific Reports*, *10*(1), 16277.
Bowers, Jeffrey S., Malhotra, Gaurav, et al. (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 1–74.
Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., et al. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Computational Biology*, *15*(4), Article e1006897.
Caron, Mathilde, Bojanowski, Piotr, Joulin, Armand, & Douze, Matthijs (2018). Deep clustering for unsupervised learning of visual features. In *European conference on computer vision* (pp. 132–149).
Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, & Fei-Fei, Li (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).
Ding, Keyan, Ma, Kede, Wang, Shiqi, & Simoncelli, Eero P. (2020). Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(5), 2567–2581.
Dosovitskiy, Alexey, et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR2021*.
Funke, Christina M., Borowski, Judy, Stosio, Karolina, Brendel, Wieland, Wallis, Thomas S. A., & Bethge, Matthias (2021). Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, *21*(3), 16–16.
Gatys, Leon A., Ecker, Alexander S., & Bethge, Matthias (2015). A neural algorithm of artistic style.
Gatys, Leon A., Ecker, Alexander S., & Bethge, Matthias (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2414–2423).
Geirhos, R., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*, 665–673.
Geirhos, Robert, Meding, Kristof, & Wichmann, Felix A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. arXiv:2006.16736.
Geirhos, Robert, Rubisch, Patricia, Michaelis, Claudio, Bethge, Matthias, Wichmann, Felix A., & Brendel, Wieland (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *int. conf. learn. repr.*. https://arxiv.org/abs/1811.12231.
Gomez-Villa, A., Bertalmio, M., & Malo, J. (2020). Visual information flow in wilson-cowan networks. *Journal of Neurophysiology*.
Gomez-Villa, A., Martín, A., Vazquez-Corral, J., Bertalmío, M., & Malo, J. (2020). Color illusions also deceive CNNs for low-level vision tasks: Analysis and implications. *Vision Research*, *176*, 156–174.
Goyal, Priya, Duval, Quentin, Reizenstein, Jeremy, Leavitt, Matthew, Xu, Min, Lefaudeux, Benjamin, et al. (2021). VISSL. https://github.com/facebookresearch/vissl.
He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, & Sun, Jian (2016). Deep residual learning for image recognition. In *Proceedings of computer vision and pattern recognition* (pp. 770–778).
Hepburn, Alexander, Laparra, Valero, Malo, Jesús, McConville, Ryan, & Santos-Rodriguez, Raul (2020). Perceptnet: A human visual system inspired neural network for estimating perceptual distance. In *2020 IEEE international conference on image processing* (pp. 121–125).
Hepburn, Alexander, Laparra, Valero, Santos-Rodríguez, Raúl, Ballé, Johannes, & Malo, Jesus (2022). On the relation between statistical learning and perceptual distances. In *International conference on learning representations*.
Hong, H., Yamins, D. L., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*, 613.
Huang, Gao, Liu, Zhuang, Van Der Maaten, Laurens, & Weinberger, Kilian Q (2017). Densely connected convolutional networks. In *Proceedings of the computer vision and pattern recognition* (pp. 4700–4708).
Kastryulin, Sergey, Zakirov, Dzhamil, & Prokopenko, Denis (2019). PyTorch Image Quality: Metrics and measure for image quality assessment. Open-source software available at https://github.com/photosynthesis-team/piq.
Kastryulin, Sergey, Zakirov, Jamil, Prokopenko, Denis, & Dylov, Dmitry V. (2022). Pytorch image quality: Metrics for image quality assessment. http://dx.doi.org/10.48550/ARXIV.2208.14818.
Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific Reports*, *6*, 32672.
Komodakis, Nikos, & Gidaris, Spyros (2018). Unsupervised representation learning by predicting image rotations. In *International conference on learning representations*.
Kriegeskorte, Nikolaus (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*(1), 417–446.
Krizhevsky, Alex, Hinton, Geoffrey, et al. (2009). Learning multiple layers of features from tiny images.
Krizhevsky, Alex, Sutskever, Ilya, & Hinton, Geoffrey E. (2012). ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th neural information processing systems* (pp. 1097–1105).
Kumar, Manoj, Houlsby, Neil, Kalchbrenner, Nal, & Cubuk, Ekin Dogus (2022). Do better ImageNet classifiers assess perceptual similarity better? *Transactions of Machine Learning Research*.
Laparra, V., Muñoz Marí, J., & Malo, J. (2010). Divisive normalization image quality metric revisited. *Journal of the Optical Society of America A*, *27*(4), 852–864.
Li, Qiang, Gomez-Villa, Alex, Bertalmío, Marcelo, & Malo, Jesús (2022). Contrast sensitivity functions in autoencoders. *Journal of Vision*, *22*(6), 8–8.
Lin, Hanhe, Hosu, Vlad, & Saupe, Dietmar (2019). KADID-10k: A large-scale artificially distorted IQA database. In *2019 tenth international conference on quality of multimedia experience* (pp. 1–3). IEEE.
Liu, Zhuang, et al. (2022). A convnet for the 2020s. In *IEEE cVPR 2022* (pp. 11976–11986).
Malo, J. (2020). Spatio-chromatic information available from different neural layers via Gaussianization. *Journal of Mathematical Neuroscience*, *10*(18), 10.1186/s13408–020–00095–8.
Malo, J., & Hernandez-Camara, P. (2024). A separate theory-on-top level may be inspiring, but it is neither separate nor enough. *Journal of Physiology*, *602*(9), 1919.
Malo, J., Pons, A. M., & Artigas, J. M. (1997). Subjective image fidelity metric based on bit allocation of the human visual system in the DCT domain. *Image and Vision Computing*, *15*(7), 535–548.
Malo, J., & Simoncelli, E. (2015). Geometrical and statistical properties of vision models obtained via maximum differentiation. In *Proc. SPIE electronic imaging*. International Society for Optics and Photonics, 93940L–93940L.
Marr, David, & Poggio, Tomaso (1976). From understanding computation to understanding neural circuitry. AI Memo No. AIM-357.
Martinez, M., Cyriac, P., Batard, T., Bertalmío, M., & Malo, J. (2018). Derivatives and inverse of cascaded linear+nonlinear neural models. *PLoS One*, *13*(10), 1–49.

Martinez-Garcia, Marina, Bertalmío, Marcelo, & Malo, Jesús (2019). In praise of artifice reloaded: Caution with natural image databases in modeling vision. *Frontiers in Neuroscience*, *13*.

Noroozi, Mehdi, & Favaro, Paolo (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision* (pp. 69–84).

Poggio, Tomaso (2021). From Marr's vision to the problem of human intelligence. *CBMM Memo*, (118).

Ponomarenko, Nikolay, Jin, Lina, Ieremeiev, Oleg, Lukin, Vladimir, Egiazarian, Karen, Astola, Jaakko, et al. (2015). Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, *30*, 57–77.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *Int'l Journal of Computer Vision*, *40*(1), 49–71.

Radford, Alec, et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.

Raschka, Sebastian (2023). Deep learning models. https://github.com/rasbt/deeplearning-models,

Rust, Nicole C., & Movshon, J. Anthony (2005). In praise of artifice. *Nature Neuroscience*, *8*(12), 1647–1650.

Schrimpf, Martin, Kubilius, Jonas, Hong, Ha, Majaj, Najib J, Rajalingham, Rishi, Issa, Elias B, et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? Article 407007, BioRxiv.

Shakhnarovich, G., Batra, D., Kulis, B., & Weinberger, K. (2011). Beyond mahalanobis: supervised large-scale learning of similarity. In *NIPS workshop on metric learning*.

Simonyan, Karen, & Zisserman, Andrew (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd international conference on learning representations* (pp. 1–14).

Sterling, P., & Laughlin, S. (2015). *Principles of neural design*. MIT Press.

Sucholutsky, Ilia, & Griffiths, Thomas L. (2023). Alignment with human representations supports robust few-shot learning. arXiv preprint arXiv:2301.11990.

Tan, Mingxing, & Le, Quoc (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114).

TorchVision maintainers and contributors (2016). TorchVision: Pytorch's computer vision library. https://github.com/pytorch/vision,

Wang, Zhou, Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wang, Pei, Li, Yijun, & Vasconcelos, Nuno (2021). Rethinking and improving the robustness of image style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 124–133).

Watson, Andrew B. (Ed.), (1993). *Digital images and human vision*. Cambridge, MA, USA: MIT Press.

Watson, A. B, & Malo, J. (2002). Video quality measures based on the standard spatial observer. In *Proc. IEEE int. conf. im. proc. (ICIP 2002)*: *vol. 3*, (pp. III–41). IEEE.

Wichmann, F. A., Janssen, D. H. J., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., et al. (2017). Methods and measurements to compare men against machines. *Electronic Imaging*, 36–45(10).

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*, 356–365.

Yamins, Daniel, Hong, Ha, Cadieu, Charles F., Solomon, Ethan A., Seibert, Darren, & DiCarlo, James J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*, 8619–8624.

Zamir, Amir R., et al. (2018). Taskonomy: Disentangling task transfer learning. In *IEEE cVPR*.

Zhang, Richard, Isola, Phillip, & Efros, Alexei A. (2016). Colorful image colorization. In *European conference on computer vision* (pp. 649–666).

Zhang, Richard, Isola, Phillip, Efros, Alexei A, Shechtman, Eli, & Wang, Oliver (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 586–595).

Zhou, Bolei, Lapedriza, Agata, Khosla, Aditya, Oliva, Aude, & Torralba, Antonio (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, Bolei, Lapedriza, Agata, Xiao, Jianxiong, Torralba, Antonio, & Oliva, Aude (2014). Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*: *vol. 27*, Curran Associates, Inc.