Computational Modeling of Human Visual Function using Psychophysics, Deep Neural Networks, and Information Theory

PhD Thesis - Compendium of Publications

Author : Qiang Li

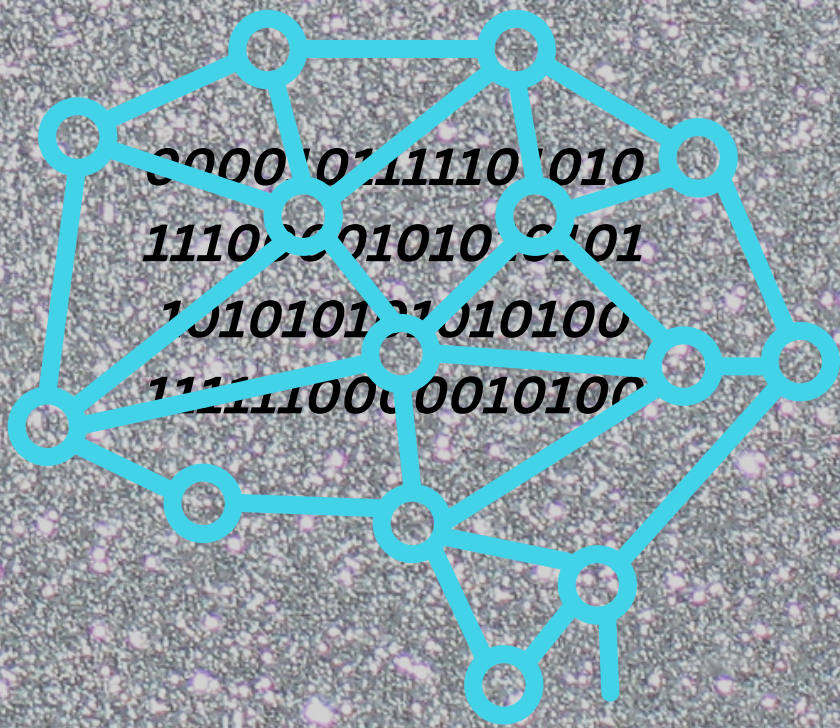
Supervisor : Jesús Malo

January 2023

Faculty of Physics

Doctorate Program in Neuroscience

Universitat de València



Computational Modeling of Human Visual Function using Psychophysics, Deep Neural Networks, and Information Theory

Author: Qiang Li

Faculty of Physics
Image Processing Laboratory
Parc Cientific, University of Valencia, Valencia, Spain

Supervisor: Jesús Malo

Submitted in fulfillment of the requirements
for the degree of Doctor in Neuroscience
by the Doctorate Program in Neuroscience of the Universitat de Valencia

January 2023

Jesús Malo López, Catedrático del Departamento de Óptica y Optometría y Ciencias de la Visión de la Facultat de Física de la Universitat de València

INFORMA

de que la memoria **Computational Modeling of Human Visual Function using Psychophysics, Deep Neural Networks, and Information Theory**, se ha realizado bajo su dirección en el Image Processing Lab de la Universitat de València por el neuroingeniero **Qiang Li** y constituye su *Tesis* para optar al grado de **Doctor en Neurociencia**.

Y para que así conste, en cumplimiento de la legislación vigente, presenta en la Universitat de València la referida Tesis Doctoral.

Valencia 6 de enero de 2023

JESUS|MALO|
LOPEZ

Firmado digitalmente
por JESUS|MALO|LOPEZ
Fecha: 2023.01.06
21:09:51 +01'00'

Director de la Tesis y Tutor

NOTE TO THE READER

According to the University of Valencia Doctorate Regulation[1] this PhD dissertation is presented as a **compendium of publications**. This regulation requires at least three publications in international journals containing the results of the conducted research. This thesis includes three papers and also describes work that has recently been submitted to a scientific journal and is still under review. Furthermore, in accordance with the aforementioned regulation and with the aim to foster the language of the University of Valencia in research and education activity, this PhD dissertation starts with two abstracts in English and Spanish. In addition, a conclusion dissertation in English and Spanish are included at the end of the Thesis. Following the regulations, the main body basically included two parts:

PART I: Extended summary: (1) goals, (2) introduction, (3) methods and results, and (4) conclusions.
PART II: Appendix with journal publications (published and under review).

---

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.
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.

Contents

Acknowledgement

> Be true to yourself, help others, make each day your masterpiece, make friendship a fine art, drink deeply from good books - especially the Bible, build a shelter against a rainy day, give thanks for your blessings and pray for guidance every day.
>
> *John Wooden*

Time flies. I started working on my dissertation about three years ago, but looking back, it was a short time, but I had a very comfortable and rich experience in Valencia, Spain. These last three years have been difficult not only for me but for everyone. The world has changed, and we will never come back again. Even though neither they nor I realized it at the time, many different people have had an influence on my life throughout the years by providing me with encouragement and assistance. This is something that cannot be denied. There are far too many for us to be able to list each and every one of them here. I want to express my gratitude to everyone who is going to assist me. I'd like to take a moment to recognize and thank a few very important people.

I owe the deepest gratitude to my advisor, Jesús Malo, whose insight, dedication, patience, and unfazed optimism will serve as a model for any scientific enterprise I undertake. His ability to see through the details to the important looming questions has been truly inspiring. We spend a lot of time together. To be honest, I learned a lot of skills from him, and I really want to express my gratitude to him. My time in Valencia was made pleasurable by many extraordinary people. I would like to acknowledge all the people who give me support in both life and study. Finally, I have been very fortunate to know all my coauthors, and we did some great work together during my studying days. They have provided invaluable ideas and support during these years.

Thanks to all my friends, I never feel alone during my weekends and holidays. We usually hang out to have a launch or dinner, and we also talk a lot about our lab life and share our experiences. More importantly, we also travel together during holidays. All of these make me not only see the outside world but also meet some great people.

Finally, I'd like to thank my parents for their unwavering support throughout my long academic journey, as well as all members of my family, but especially my parents, for their love, lessons, and support throughout all these years.

Abstract

Visual perception is a key to understanding how the brain works because most of the information is processed by the early visual system and then sent to the high-level cognitive perception brain regions. The brain functions as a self-organizing, bio-dynamic, and chaotic system that receives outside information and then decomposes it into pieces of information that can be processed efficiently and independently.

Those with an interest in computers, image processing, or biological vision will find natural image statistics to be of great value. As natural images form the basic stimuli in most vision-oriented tasks, it is of great importance to understand their unique properties and statistical structure. In this thesis, I show a number of results that look at how the statistical properties of natural images and how people see them are related.

The work connects psychophysics, deep neural networks, and information theory to perceptual vision systems to explore how vision processes information from the outside world and how the information communicated drives functional connectivity between visual regions and even higher-level brain regions.

This Thesis (compendium of publications) specifically addresses the following scientific questions, where each goal corresponds to a related journal publication.

Goal 1: Quantifying the information flow in the noisy brain is crucial to understand its functional connectivity. Mutual information or its extension to multiple nodes, such as total correlation, could capture the information shared by multiple brain regions. The role of redundancy in the brain is still not explored up to the last consequences due to the practical problems in its estimation, and how the redundancy explains basic cognitive functions remains unclear. In this thesis, we try to propose new answers to the question of communication between brain regions.

Goal 2: In addition to quantifying small-scale functional connectivity with total correlation, we want to extended our research to include large-scale functional connectivity. On the one hand, functional networks derived from total correlation can be different from the ones that one obtains using other measures so it can reveal yet-to-be-known relations. On the other hand, anomalies in the networks derived from total correlation can be useful to find some potential biomarkers for brain diseases.

Goal 3: In particular we want to quantify functional connectivity in the early vision system using new estimates of total correlation because (1) as opposed to other brain regions, there are well understood analytical models of the early visual pathway (e.g. based on linear transforms, divisive normalization and pooling), (2) these visual regions have been extensively studied from information-theoretic perspectives, and (3) these biological vision models have strong connections with current artificial neural networks that have gained a lot of attention nowadays. Our aim is obtaining an analytical description of the functional connectivity. Analytical results for the visual areas not only provide higher insight into this fundamental brain function but also, as by product, they could represent a realistic and configurable ground-truth scenario to test empirical estimators of information to be used in other (less understood) brain regions.

Goal 4: Today, using deep neural networks to study biological vision is a hot research topic in the neuroscience community, and deep neural networks indeed accelerate both artificial

intelligence and neuroscience development. There are multiple artificial architectures trained for specific goals, and some of them do outperform humans on certain vision tasks. However we need to be cautious in using these models in biological vision given the functional difference between artificial deep neural networks and natural networks in the visual brain. Therefore, we want to explore the similarity and differences between some low-level visual behavior in humans and in deep neural networks devoted to vision. In particular we will try to understand the functional origin of the bandwidth limits of early human vision (i.e. the spatio-temporal and chromatic contrast sensitivity) using autoencoders, and we will compare information degradation in biological visual areas and in standard artificial networks devoted to vision.

Based on the above goals, here we will quickly summarize each associated paper:

The *first goal-related* publication appears in the Neural Networks [1] and we find that interaction information and total correlation do provide an explicit explanation for quantifying information flow among multiple brain regions. The differences between mutual information, interaction information, and total correlation are discussed in this article. In addition, we proposed a new strategy for calculating the interaction information between three variables by using total correlation and conditional mutual information in neuroscience in parallel with other labs. Furthermore, it is unclear how to effectively implement it in real-world scenarios, and we attempt to address these issues using two information theory estimators, RBIG and CorEX. To estimate functional connectivity in the brain using the aforementioned three higher-order information-theoretic approaches, we provided data from both simulation experiments and actual neural studies. We found that interaction information and total correlation were both robust in their ability to capture redundancy information for multivariate variables, suggesting that this method may be applicable to the study of both established and as-yet-unidentified functional brain connections. Our research shows that we could use this high-order information theory metric to unravel some meaningful neuroscience problems.

The *second goal-related* publication appears in the Entropy journal [2] where we extended the above notion in order to infer a large-scale connection network based on total correlation and demonstrated the potential use of such networks as biomarkers of changes in the brain. We applied the concept of total correlation in order to capture these multivariate, large-scale connections between different brain regions. Through the use of experimental testing, it has been demonstrated that the aforementioned processes are effective in re-creating multivariate relationships in the brain. In this investigation, the overall correlation was calculated with the help of CorEx. The CorEx approach is able to accurately capture the intricacies of functional connectivity when more than simply a pair of regions in the brain are being investigated at the same time. Moreover, we tested the approach using data from large-scale fMRI scans taken while the subjects were in a resting condition. We realized that it was not possible to identify multivariable relationships by relying solely on pairwise correlation and mutual information values. The total correlation is a useful method that enables us to cluster multivariate relationships, which may be understood in a broader sense. Total correlation measures are an essential tool for figuring out the extensive functional connection that exists between different parts of the brain. We have shown that total correlation may be used to assess functional connectivity in a real neural dataset as well as uncover biomarkers that can be utilized to diagnose brain illnesses, as we indicated earlier.

The *third goal-related* publication is under review in the Neural Networks [3]. Total Correlation is used to describe the functional connectivity in the visual pathway analytically. The connectivity between the nodes, within the cortex, and the eventual top-down feedback can all be adjusted in our neural model, which consists of three layers (retina, LGN, and V1 cortex). We derive analytical results for the three-way Total Correlation and for all possible

pairwise Mutual Information measures in this multivariate setting (three nodes with multidimensional signals). Simulation and analytical analysis results demonstrate that three-way Total Correlation captures the impact of distinct intra-cortical inhibitory connections, while pairwise Mutual Information does not. The presented analytical framework can also be used to validate Total Correlation empirical estimators. Accordingly, once a reliable estimator has been identified, the behavior with non-Gaussian signals can be investigated. Despite the fact that the Gaussian assumption cannot be made, the empirical results with RBIG from natural signals show the same tendencies. Furthermore, we also use real fMRI recordings to look at the functional connections in real brain visual regions, for instance, V1, V2, V3, and V4.

The *fourth goal* focused on the study of similarities and differences between human vision and deep-nets for vision appears in the Journal of Vision [4] and is the subject of additional *on-going work* strictly derived from the other publications in this Thesis. While the JoV uses a psychophysical approach, the *on-going* work takes an information theory perspective.

First, from a psychophysical perspective, we reevaluate the importance of low-level vision tasks in explaining the contrast sensitivity functions (CSFs) in light of 1) the recent trend of using artificial neural networks for studying vision and 2) the current understanding of retinal image representations. As a first contribution, we show that autoencoders, a popular type of convolutional neural networks (CNNs), can learn to perform some low-level vision tasks (such as retinal noise and optical blur removal) with human-like CSFs in the spatial and temporal dimensions, but not others (such as chromatic adaptation or pure reconstruction after simple bottlenecks). Second, we show experimentally that, for some functional goals (at low abstraction level), deeper CNNs that are better at achieving the quantitative goal are actually worse at replicating human-like phenomena (such as the CSFs). Consistent with a growing body of research, our findings add a note of caution about CNNs in vision science, arguing that their oversimplification of visual processing and reliance on unrealistic architectures for goal optimization may prevent them from being fully utilized in the study of human vision.

Second, while the parallels and contrasts between the brain and convolution neural networks are currently investigated on several levels, there are very few studies that explain them from an information-theoretical standpoint. In this *on-going* work, we quantify the degradation of information along pre-trained AlexNet and VGG16 and in biological visual systems.

Keywords: Perception; Human vision system; Deep neural networks; Contrast sensitivity functions; Divisive normalization; Information theory; Functional connectivity; Total correlation; Natural images statistics; Large-scale connectivity; Biomarkers;

Resumen

[1] La percepción visual es clave para entender cómo funciona el cerebro, porque la mayor parte de la información es procesada por el sistema visual primitivo y enviada después a las regiones cerebrales de percepción cognitiva de alto nivel. El cerebro funciona como un sistema autoorganizado, biodinámico y caótico que recibe información exterior y luego la descompone en trozos de información que pueden procesarse de forma eficiente e independiente.

Las personas interesadas en la informática, el tratamiento de imágenes o la visión biológica encontrarán de gran utilidad la estadística de imágenes naturales. Dado que las imágenes naturales constituyen los estímulos básicos en la mayoría de las tareas orientadas a la visión, es de gran importancia comprender sus propiedades únicas y su estructura estadística. En esta tesis, muestro una serie de resultados que analizan cómo se relacionan las propiedades estadísticas de las imágenes naturales y cómo las ven las personas.

El trabajo conecta la psicofísica, las redes neuronales profundas y la teoría de la información con los sistemas de visión perceptiva para explorar cómo la visión procesa la información del mundo exterior y cómo la información comunicada impulsa la conectividad funcional entre las regiones visuales e incluso las regiones cerebrales de nivel superior.

Esta Tesis (compendio de publicaciones) aborda específicamente las siguientes cuestiones científicas, donde cada objetivo corresponde a una publicación en revista relacionada.

Objetivo 1: Cuantificar el flujo de información en el cerebro ruidoso es crucial para comprender su conectividad funcional. La información mutua o su extensión a múltiples nodos, como la correlación total, podría captar la información compartida por múltiples regiones cerebrales. El papel de la redundancia en el cerebro no se ha explorado hasta las últimas consecuencias debido a los problemas prácticos que plantea su estimación, y sigue sin estar claro cómo explica la redundancia las funciones cognitivas básicas. En esta tesis, intentamos proponer nuevas respuestas a la cuestión de la comunicación entre regiones cerebrales.

Objetivo 2: Además de cuantificar la conectividad funcional a pequeña escala con la correlación total, queremos ampliar nuestra investigación para incluir la conectividad funcional a gran escala. Por un lado, las redes funcionales derivadas de la correlación total pueden ser diferentes de las que se obtienen utilizando otras medidas, por lo que pueden revelar relaciones aún por conocer. Por otro lado, las anomalías en las redes derivadas de la correlación total pueden ser útiles para encontrar algunos biomarcadores potenciales de enfermedades cerebrales.

Objetivo 3: En particular, queremos cuantificar la conectividad funcional en el sistema de visión temprana utilizando nuevas estimaciones de la correlación total porque (1) a diferencia de otras regiones cerebrales, existen modelos analíticos bien comprendidos de la vía visual temprana (por ejemplo, basados en transformaciones lineales, normalización divisoria y agrupación), (2) estas regiones visuales han sido ampliamente estudiadas desde perspectivas de teoría de la información, y (3) estos modelos biológicos de visión tienen fuertes conexiones con las actuales redes neuronales artificiales que han ganado mucha atención en la actualidad. Nuestro objetivo es obtener una descripción analítica de la conectividad funcional. Los resultados analíticos para las áreas visuales no sólo proporcionan una mayor comprensión de esta función cerebral fundamental sino que también, como

[1] Traducido del inglés por Qiang Li

producto derivado, podrían representar un escenario realista y configurable de verdad básica para probar los estimadores empíricos de información que se utilizarán en otras regiones cerebrales (menos comprendidas).

Objetivo 4: Hoy en día, el uso de redes neuronales profundas para estudiar la visión biológica es un tema de investigación candente en la comunidad neurocientífica, y las redes neuronales profundas aceleran de hecho el desarrollo tanto de la inteligencia artificial como de la neurociencia. Existen múltiples arquitecturas artificiales entrenadas para objetivos específicos, y algunas de ellas superan a los humanos en determinadas tareas de visión. Sin embargo, debemos ser cautos a la hora de utilizar estos modelos en la visión biológica, dada la diferencia funcional entre las redes neuronales profundas artificiales y las redes naturales del cerebro visual. Por lo tanto, queremos explorar las similitudes y diferencias entre algunos comportamientos visuales de bajo nivel en humanos y en redes neuronales profundas dedicadas a la visión. En particular, intentaremos comprender el origen funcional de los límites de ancho de banda de la visión humana temprana (es decir, la sensibilidad al contraste espacio-temporal y cromático) utilizando autocodificadores, y compararemos la degradación de la información en áreas visuales biológicas y en redes artificiales estándar dedicadas a la visión.

Basándonos en los objetivos anteriores, a continuación resumiremos rápidamente cada artículo.

La *primera publicación relacionada con objetivos* aparece en Neural Networks [1] y descubrimos que la información de interacción y la correlación total sí proporcionan una explicación explícita para cuantificar el flujo de información entre múltiples regiones cerebrales. Las diferencias entre la información mutua, la información de interacción y la correlación total se discuten en este artículo. Además, proponemos una nueva estrategia para calcular la información de interacción entre tres variables utilizando la correlación total y la información mutua condicional en neurociencia en paralelo con otros laboratorios. Además, no está claro cómo aplicarla eficazmente en escenarios del mundo real, e intentamos abordar estas cuestiones utilizando dos estimadores de la teoría de la información, RBIG y CorEX. Para estimar la conectividad funcional en el cerebro utilizando los tres enfoques teóricos de la información de orden superior mencionados, aportamos datos tanto de experimentos de simulación como de estudios neuronales reales. Descubrimos que tanto la información de interacción como la correlación total eran robustas en su capacidad de capturar información redundante para variables multivariantes, lo que sugiere que este método puede ser aplicable al estudio de conexiones cerebrales funcionales tanto establecidas como aún no identificadas. Nuestra investigación demuestra que podríamos utilizar esta métrica de la teoría de la información de alto orden para desentrañar algunos problemas significativos de la neurociencia.

La *segunda publicación relacionada con el objetivo* aparece en la revista Entropy [2], donde ampliamos la noción anterior para inferir una red de conexiones a gran escala basada en la correlación total y demostramos el uso potencial de dichas redes como biomarcadores de cambios en el cerebro. Aplicamos el concepto de correlación total para captar estas conexiones multivariantes a gran escala entre diferentes regiones cerebrales. Mediante el uso de pruebas experimentales, se ha demostrado que los procesos mencionados son eficaces para recrear relaciones multivariantes en el cerebro. En esta investigación, la correlación global se calculó con ayuda de CorEx. El enfoque CorEx es capaz de captar con precisión las complejidades de la conectividad funcional cuando se investigan al mismo tiempo más de un par de regiones del cerebro. Además, probamos el enfoque utilizando datos de escáneres fMRI a gran escala tomados mientras los sujetos estaban en estado de reposo. Nos dimos cuenta de que no era posible identificar relaciones multivariables basándonos únicamente en los valores de correlación por pares y de información mutua. La correlación total es

un método útil que nos permite agrupar relaciones multivariables, que pueden entenderse en un sentido más amplio. Las medidas de correlación total son una herramienta esencial para averiguar la extensa conexión funcional que existe entre las distintas partes del cerebro. Hemos demostrado que la correlación total puede utilizarse para evaluar la conectividad funcional en un conjunto de datos neuronales reales, así como para descubrir biomarcadores que pueden utilizarse para diagnosticar enfermedades cerebrales, como hemos indicado anteriormente.

La *tercera publicación relacionada con el objetivo* está en revisión en la revista Neural Networks [3]. La correlación total se utiliza para describir analíticamente la conectividad funcional en la vía visual. La conectividad entre los nodos, dentro de la corteza, y la eventual retroalimentación descendente pueden ajustarse en nuestro modelo neural, que consta de tres capas (retina, LGN y corteza V1). Derivamos resultados analíticos para la Correlación Total de tres vías y para todas las medidas posibles de Información Mutua por pares en este entorno multivariante (tres nodos con señales multidimensionales). Los resultados de la simulación y del análisis analítico demuestran que la Correlación Total de tres vías capta el impacto de las distintas conexiones inhibitorias intracorticales, mientras que la Información Mutua por pares no lo hace. El marco analítico presentado también puede utilizarse para validar los estimadores empíricos de Correlación Total. Por consiguiente, una vez que se ha identificado un estimador fiable, se puede investigar el comportamiento con señales no gaussianas. A pesar de que no se puede hacer la suposición gaussiana, los resultados empíricos con RBIG a partir de señales naturales muestran las mismas tendencias. Además, también utilizamos grabaciones fMRI reales para observar las conexiones funcionales en regiones visuales cerebrales reales, por ejemplo, V1, V2, V3 y V4.

El *cuarto objetivo* centrado en el estudio de las similitudes y diferencias entre la visión humana y las redes profundas para la visión aparece en el Journal of Vision [4] y es objeto de un *trabajo en curso* adicional estrictamente derivado de las otras publicaciones de esta Tesis. Mientras que el JoV utiliza un enfoque psicofísico, el trabajo en curso adopta una perspectiva de teoría de la información.

En primer lugar, desde una perspectiva psicofísica, reevaluamos la importancia de las tareas de visión de bajo nivel para explicar las funciones de sensibilidad al contraste (CSFs) a la luz de 1) la reciente tendencia a utilizar redes neuronales artificiales para estudiar la visión y 2) la comprensión actual de las representaciones de la imagen retiniana. Como primera contribución, mostramos que los autocodificadores, un tipo popular de redes neuronales convolucionales (CNNs), pueden aprender a realizar algunas tareas de visión de bajo nivel (como la eliminación del ruido retiniano y del desenfoque óptico) con CSFs similares a las humanas en las dimensiones espacial y temporal, pero no otras (como la adaptación cromática o la reconstrucción pura después de simples cuellos de botella). En segundo lugar, mostramos experimentalmente que, para algunos objetivos funcionales (a bajo nivel de abstracción), las CNNs más profundas que son mejores para alcanzar el objetivo cuantitativo son en realidad peores para replicar fenómenos similares a los humanos (como los CSFs). En consonancia con un creciente cuerpo de investigación, nuestros hallazgos añaden una nota de precaución sobre las CNN en la ciencia de la visión, argumentando que su simplificación excesiva del procesamiento visual y la dependencia de arquitecturas poco realistas para la optimización de objetivos pueden impedir que se utilicen plenamente en el estudio de la visión humana.

En segundo lugar, aunque los paralelismos y contrastes entre el cerebro y las redes neuronales de convolución se investigan actualmente a varios niveles, hay muy pocos estudios que los expliquen desde el punto de vista de la teoría de la información. En este trabajo, cuantificamos la degradación de la información a lo largo de AlexNet y VGG16 pre-entrenadas y en sistemas visuales biológicos.

Chapter 1

Objectives and Organization of the Thesis

1.1 Research Objectives

1. **Objective 1** Solving the problems of pair-wise measures of neural functional connectivity by using multivariate information theory measures, for instance, interaction information and total correlation. In applied neuroscience research, interaction information and total correlation are neglected and little explored compared to mutual information. Interaction information and total correlation do not directly describe quantification of information flow between brain regions in neuroscience. The goal is to illustrate the distinctions between the neuroscience concepts of mutual information, interaction information, and total correlation. On the other hand, how to appropriately use it in real-world scenarios

2. **Objective 2** The development of techniques to estimate total correlation can be applied to quantify large-scale functional connectivity and biomarkers. We applied total correlation to large-scale functional connectivity and brain disease.

3. **Objective 3** The development of an analytical but realistic neural scenario to check functional connectivity measures based on information theory. As an analytical tool, total correlation can be used to characterize the functional connectivity of the visual pathway. Our neural model consists of three layers, allowing us to modify the connectivity between the nodes, the connectivity within the cortex, and the top-down feedback that is ultimately implemented. Further, we analyze the functional connections between real V1, V2, V3, and V4 visual regions of the brain using fMRI recordings.

4. **Objective 4** The similarities and differences between the brain and convolutional neural networks are explored at various levels. There haven't been many investigations into this topic that use psychophysical and information theoretical explanations. First, modeling low-level visual properties, CSF with autoencoder, and when trained to perform fundamental low-level vision tasks, it may create CSFs in the spatio-chromatic and temporal dimensions similar to those of humans. Second, we compared the similarities and differences between pre-trained AlexNet, VGG16 and visual systems and tried to understand how the visual information processing between pre-trained neural networks and real visual systems work. We also explored how the redundancy altered in visual and artificial neural networks. It has significant meaning for us to understand visual and artificial neural networks from an information-theoretical view.

1.2 Organization of the Thesis

The thesis basically includes two parts. ***PART I*** consists of goal description, introduction, methodology and results, and conclusions, respectively. ***PART II*** mainly includes scientific publications. The following is a more detailed structure:

1. ***PART I***
    - Abstract (English, and Spanish)
    - Chapter 1: Objectives and Organization of the Thesis
    - Chapter 2: Introduction: Background
    - Chapter 3: Methodology and Results
    - Chapter 4: Conclusions in English
    - Chapter 5: Conclusions in Spanish
2. ***PART II***
    - Appendix with Scientific Publications

Chapter 2

Introduction: Background

> Neuroscience over the next 50 years is going to introduce things that are mind-blowing.
>
> *David Eagleman*

In this brief introduction, we will go over some of the most foundational methods and concepts used in our study of human vision, including psychophysics, deep neural networks, and information theory.

2.1 Color Vision and Contrast Sensitivity Functions

2.1.1 Color Vision: Opponent Colors Theory

The theory of opponent colors will be applied to a publication paper in the Journal of Vision [4], specifically when we synthesize achromatic and chromatic sinusoidal waves with varying frequencies. As a result, we will quickly review the fundamental concepts of opponent color theory here.

The trichromatic hypothesis is a useful tool for explaining how different types of cone receptors may sense light of varying wavelengths [5, 6]. On the other hand, the opponent process theory provides an explanation for how these cones connect to the nerve cells in our brains that are responsible for determining how we actually experience colors. The trichromatic hypothesis, on the other hand, describes how color vision takes place at the receptor level, and the opponent process theory explains how color vision takes place at the brain level. According to the opponent process theory, white and black, red and green, and yellow and blue are the three channels that make up the three pairs of opponent color channels [6–8]. This theory postulates that the manner in which humans perceive colors is governed by three contrasting systems. Each pair of opponent colors works to suppress the other.

When the outputs of all three types of cones are added together $(L+M+S)$, an achromatic response is produced. Due to the fact that the cone signals are differentiated from one another, it is possible to build red-green $(L-M+S)$ and yellow-blue $(L+M-S)$ opponent signals. As a result of translating *LMS* signals into opponent signals, the color information sent over

the three channels is no longer related to each other. This makes signal transmission more effective and makes noise less of a problem [9].

2.1.2 Color Vision: Color Constancy

The theory of color constancy will be applied to a publication paper in the Journal of Vision [4], specifically when we explore color constancy in deep neural networks. As a result, we will quickly review the fundamental concepts of color constancy here.

Vision is the process through which we see color, and color constancy refers to surfaces in a scene that are unaffected by changes in illumination [10]. When used in a wide variety of natural settings, digital cameras and machine vision systems face a significant challenge when it comes to determining how the illumination affects their images. The human visual system is consistent, can adjust to different levels of illumination, and is a very significant cognitive mechanism for the processing of color information. There are a wide number of illumination-independent descriptors that have been produced as a response to varying levels of physical or biological activity, and every one of these descriptors has proved successful on certain vision tests. In recent years, there has been a big jump in the number of people working on deep neural networks. These networks are very good at color constancy [11–13].

2.1.3 Spatial Vision: Contrast Sensitivity Functions

The concept of contrast sensitivity functions will be applied to a publication paper in the Journal of Vision [4], specifically when we try to reconstruct it from deep neural networks. As a result, we will quickly review the fundamental concepts of contrast sensitivity functions here.

The threshold reaction to contrast (sensitivity is the inverse of threshold) is used to build a contrast sensitivity function [14–16]. This response can be expressed as a function of either spatial or temporal frequency. Primate vision systems have a high degree of sensitivity to both spatial and temporal frequency, which is an extremely crucial component of the cognitive mechanisms underlying vision [17]. Too many psychophysical experiments are done to study the qualities of human vision, and from these experiments [14–16], the human contrast sensitivity functions are drawn, as shown in Fig. 1.

It provides a conceptual illustration of typical contrast sensitivity functions for luminance (black–white) and chromatic (red–green and yellow–blue at constant luminance) contrasts at constant luminance [14, 16]. The luminance contrast sensitivity function has a band-pass characteristic, and its peak sensitivity is located somewhere around 5 cycles per degree. Additionally, the low-pass characteristics of the chromatic mechanisms show that edge detection and enhancement do not take place along these dimensions [16]. Because there are fewer S cones in the retina, the blue–yellow chromatic CSF has a lower cutoff frequency than the red–green chromatic CSF. This is because blue and yellow are complementary colors. It is also interesting to note that the luminance contrast sensitivity function (CSF) is noticeably higher than the chromatic contrast sensitivity functions. This suggests that the visual system is more sensitive to small changes in luminance contrast than in chromatic contrast [14–16].

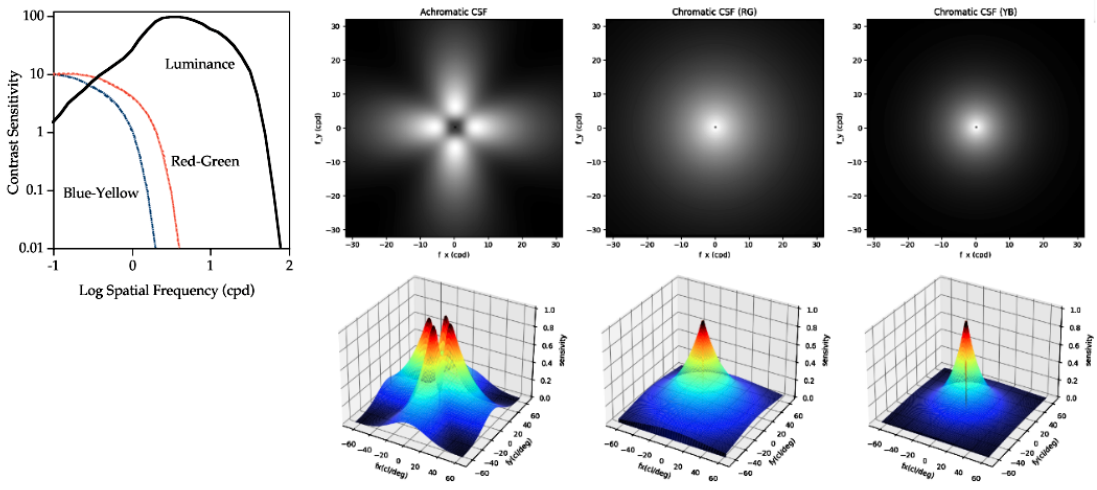


Figure 1: **Spatial contrast sensitivity functions for luminance and chromatic contrast.** The graphic shows the spatial contrast sensitivity functions for luminance and chromatic contrast in one-dimensional, two-dimensional, and three-dimensional angles.

2.2 Redundancy and Linear Representation in Visual Brain

The redundancy reduction and linear transform will be used in the paper that is under review in the Neural Networks [3]. Data processing inequality will be used, specifically when we investigate information flow in the biological linear plus nonlinear vision model. In order to do this, we will give a quick overview of some of the most important concepts about redundancy and the linear transform.

Natural images are redundant, and in order to achieve efficient processing of natural image information, a sequence of decorrelations will be applied from the retinal image to the primary visual brain [18, 19]. Because of the information bottleneck in the human visual system and the efficient coding hypothesis, the neurons will adjust to the statistical features of natural images to maximize the encoding of useful information while ignoring information that is not relevant to the task at hand [20, 21]. Principal component analysis is a dimension reduction and decorrelation approach that is used in digital image processing [22]. It transforms a correlated multivariate distribution into orthogonal linear combinations of the original variables. This technique is typically used to decrease redundancy in modeling the visual system. In comparison to principal component analysis (PCA), independent component analysis (ICA) is a demixing technique that is more effective at decreasing redundancy and may separate a multivariate signal into additive subcomponents. Both of the approaches described above have already seen widespread use to minimize duplication in natural images represented by neural representation [23–26].

2.3 Divisive Normalization

The normalization and gain control will be applied in the paper that is under review in the Neural Networks [3], specifically when we explore biological linear plus nonlinear visual model. In order to do this, we will give a quick overview of some of the most important ideas in divisive normalization.

There are many linear and nonlinear models used to describe how the visual system works. The linear parts of these models focus mostly on the early stages of color processing, receptive

fields, and contrast sensitivity functions of the visual system. The nonlinear parts are often modeled with divisive normalization, which is a canonical neural computation [27, 28]. The process of normalization involves computing a ratio that compares the response of an individual neuron to the activity that is totaled across a group of neurons.

Not only has divisive normalization already been widely used to model properties of the visual system and even some high-level visual functions [29], but it has also been applied to the evaluation of image quality [30, 31], the representation of images [32], and the development of deep neural networks, for instance, AlexNet [33]. The matrix computation for the divisive normalization can be expressed as a formula as follows:

$$\mathbf{z} = f(\mathbf{e}) = \text{sign}(\mathbf{e}) \cdot \kappa \cdot \frac{|\mathbf{e}|^{\gamma}}{b + H \cdot |\mathbf{e}|^{\gamma}} \tag{2.1}$$

where nonlinear signal, $\mathbf{z}$, results from a divisive normalization transform, $f(\cdot)$, of the outputs of the linear receptive fields at the previous intermediate layer, $\mathbf{e}$. Note that the division, the exponent, and the absolute values in $f(\cdot)$ are Hadamard (element-wise) operations [34], and the matrix H in the denominator represents the interaction between the neurons of the previous cortical layer $\mathbf{e}$.

2.4 Assessing Biological Plausibility through Image Quality

The image quality assessment will be applied to evaluate our biological model's performance in the paper that is under review in the Neural Networks [3], specifically when we explore model performance compared with human vision. In order to do this, we will give a quick overview of some of the most important ideas in image quality assessment.

Image quality assessment is an open problem for digital equipment and it's also connected to digital image or video compression, transmission, and decompression, with the elimination of coding and spatial redundancy, which use some of the characteristics of the human visual system [22, 35–37]. Image quality will directly affect the good and the bad of visual sense because the goal of image quality assessment is to predict the quality of an image as perceived by human observers. Therefore, there is a large body of image quality models that were developed based on some general or specific properties of statistics or human vision perceptual functions [38]. The natural images presented and processed in human vision systems and feature-based computing in each of the early vision layers were done, then transmitted information to the cortex for deep processing. There are four main stream image quality assessment approaches so far: full-reference, reduced-reference, and no-reference image quality assessment approaches [39, 40], respectively, and the Just-noticeable difference (JND), which is one of the important metrics, is defined as the smallest intensity change in an image that can be noticed by the human vision system [41, 42]. As we mentioned earlier, the assessment of image quality is also connected to the human vision system (HVS). It is imperative that the JND be greater than any discernible level of distortion.

2.5 Deep Neural Networks and the Visual Cortex

Visual neuroscientists have been studying convolutional neural networks as models of visual processing to better understand human information processing because it is a shared language

between machine vision and biological vision [43–45]. However, these models are usually trained in a supervised manner on object classification, but supervised learning is usually considered biologically unrealistic. Therefore, unsupervised learning is gradually getting more attention in the neuroscience field because it works more approximately the brain at some levels. We try to address these scientific problems in our paper, which was published in the Journal of Vision [4].

As we mentioned in the beginning, today, with the rise of deep learning, a lot of great achievements have been made, and many amazing deep learning architectures and related algorithms are still in development [46–48]. One of the major effect fields is computer vision [46, 49–51], and to be honest, deep learning accelerates machine vision development, and it also improves industry efficiency because artificial intelligence has already widely fusion into our daily lives.

The intersection between deep learning and neuroscience is also getting more and more attention because many computer vision approaches come from the brain. Some previous studies have already proved that brains and deep networks share qualitative similarities and differences [52–55]. In the beginning, the primary neural network was inspired by the human vision system, and it could model some specific low-level vision properties, mainly based on the way of feature processing in the human vision system, for instance, perceptrons [56], neocognitrons [57], and so on. Later, some improved neural networks were proposed based on neuron properties and hierarchical human vision systems, and all of them have achieved good performance on some computer vision tasks, such as classification [58, 59], because these neural networks are more approximations of real vision systems [60]. The big milestone event is AlexNet [33]. It competed in the ImageNet Large Scale Visual Recognition Challenge, and it achieved great performance compared to other previous neural networks [59]. From here, convolution neural networks have made leaps and bounds, and they have also achieved great performance on various computer vision tasks, for instance, classification [59], segmentation [61], saliency prediction [62], and so on. Later researchers proposed recurrent neural networks, which are considered memory mechanism, and pushed convolution neural networks to more functionality [63, 64]. As we mentioned above, using deep learning to study biological vision is a major way to inspire visual neuroscientists to understand the brain through deep neural networks [65–67]. Such as, can we use deep learning to advance neuroscience [68]? At the same time, it can inspire computer scientists to develop more realistic, functional, low-cost deep neural networks. For example, can we learn from the brain to improve deep learning?

2.6 Functional Connectivity via Information Theory

2.6.1 Information Theory in Neuroscience

The information theory will be applied to a publication paper in the Neural networks [1], Entropy [2], and Neural Networks [3] under review, specifically when we explore high-order information communicated in the human brain. In order to do this, we will give a quick overview of some of the most important ideas in information theory, such as mutual information and total correlation.

Information theory has a long and distinguished role in cognitive science and neuroscience. One of the important ways to study the brain is the quantity of information flow or share between synapses, neurons, and brain regions [69]. With the development of neuroimaging

techniques and theories, it has become possible to specify the relationship between information quantities and brain activation, allowing us to fully study brain functions associated with information processing [70, 71].

As we mentioned above, quantifying the information that is coupled between neurons or brain regions has rapidly become an important area of study in the field of neuroscience. Information theory is usually used to figure out how information is passed between neurons or parts of the brain. This can be compared to the information that is conveyed in biological systems such as the stimulus–response, neurons–neurons, or brain regions–brain regions function, as well as to the highest potential information transfer for effective and low-cost coding of information. These sorts of comparisons are essential since they validate the assumptions that are made during any neurophysiological investigation. Signals from functional magnetic resonance imaging can be used to create a network representation of the functional connections found in the human brain. Mutual information (MI), often known as a measure of non-directional connectivity [72–74], is one of the functional connectivity measures that are utilized in the analysis of fMRI data. It permits the estimation of both linear and non-linear statistical relationships between time series and can be used to discover functional coupling [75]. Additionally, it enables the assessment of functional coupling. MI analysis may be useful in understanding and quantifying the nonlinear transmission of information within the brain because neural dynamics almost certainly comprises a great number of highly nonlinear processes [76].

Beyond mutual information, Total Correlation is a generalization of mutual information that has been studied extensively in the fields of probability theory and information theory. [77]. It is also known as the multivariate constraint or multiinformation [78]. Here we recall the definitions of the descriptors compared in this work (*Mutual Information* [79] and *Total Correlation* [77]), in terms of *Entropy*:

$$I(\mathbf{x},\mathbf{y}) = h(\mathbf{x}) + h(\mathbf{y}) - h(\mathbf{x},\mathbf{y}) \tag{2.2}$$

$$T(\mathbf{x},\mathbf{y},\mathbf{z}) = \left(\sum_{i=1}^{n} h(x_i) + h(y_i) + h(z_i)\right) - h(\mathbf{x},\mathbf{y},\mathbf{z}) \tag{2.3}$$

where vector $\mathbf{x} \in \mathbb{R}^n$ is same as vector y, and z, and $h(\cdot)$ stands for the (univariate or joint) entropy of the corresponding (scalar or vector) variables. From Eq.2.3, we find that total correlation can be estimated through marginal entropy and joint entropy, and marginal entropy could be easy to estimate, but joint entropy is hard to directly estimate, but there are some novelty estimators, for instance, Rotation-based Iterative Gaussianization (RBIG) [80], Correlation Explanation (CorEx) [81], and the Matrix-based Rényi's entropy [82]. We have already used some of the above estimators in our studies to solve some neuroscience problems.

2.6.2 Functional Connectivity

The functional connectivity will be applied to a publication paper in the Neural networks [1], Entropy [2], and Neural Networks [3] under review, specifically when we explore information communicated in the human brain. As a result, we will go over the fundamental concepts of functional connectivity quickly in the following section.

The brain regions are connected together at a structural and functional level to drive high-level

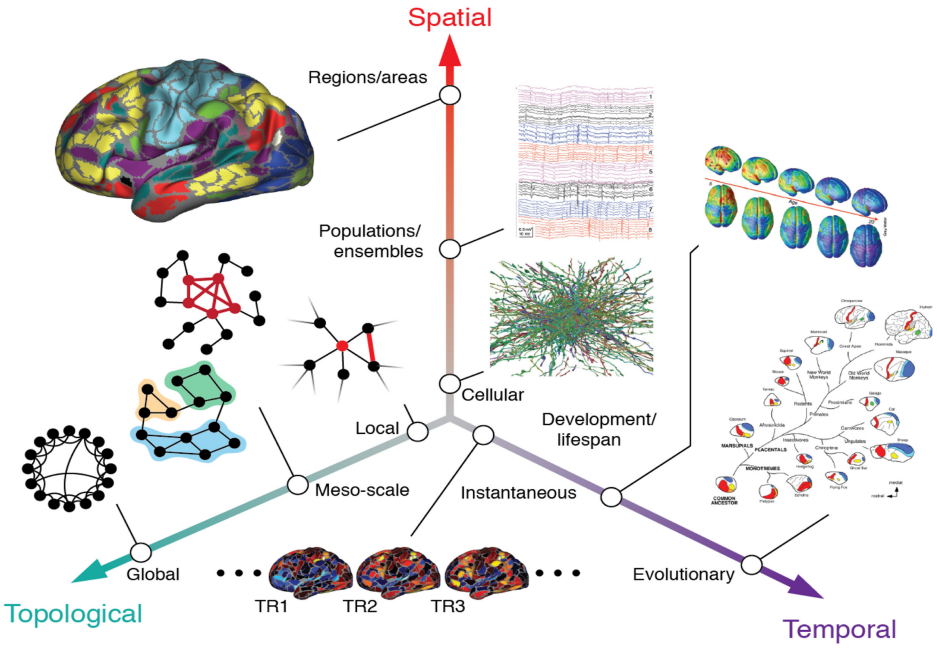


Figure 2: **Connectivity at multiple scales is functional.** The graph illustrates that functional connectivity happens at multiple scales in the human brain. The three main categories are spatial, temporal, and topological, respectively. Spatial functional connectivity is mainly estimated based on spatial brain regions, but temporal functional connectivity mainly considers connectivity alteration with time involvement, for instance, brain development and so on. Topological functional connectivity focuses on how each brain region interacts with one another and quantifies that interaction using graph theories. The figure is adapted from [83].

cognitive behaviors [83, 84] (see Figure. 2). The structure of connectivity exists physically, and it is primarily through fibers that connect wired brain regions. However, beyond structural connectivity, there are a lot of mental diseases related to altered functional connectivity, and since functional connectivity is mainly defined via statistics, it plays a significant role in understanding both basic cognitive mechanisms and some mental diseases. As previously stated, the way statistical approaches are defined has a direct impact on functional connectivity metrics, and there is a large body of metrics used to measure functional connectivity between brain regions [85]. One common metric is Pearson correlation, and it opens the door to quantifying information coupled between brain regions [86]. However, based on their own application limitations, information theory concepts gradually gain researchers' attention. At the same time, information theory begins to connect to neuroscience research, allowing us to find more evidence about information coupled or flowing between or among brain regions [1, 3]. Under the information theory framework, mutual information is one of the most popular metrics to quantify indirect information shared between brain regions because it captures both linear and nonlinear information compared to Pearson correlation. In the meantime, researchers are interested in the direction of information flow or how brain regions cause each other. Therefore, some directed information flow metrics have been developed, for instance, Granger causality [87], transfer entropy [88, 89], and so on.

As we mentioned before, if most of the information goes into the brain through vision, then how does the brain solve vision? One of the important ways is that information is hierarchically transmitted to other brain regions, which keeps their causal effects separate from each other. The data we used is from fMRI, and fMRI measures brain activity by detecting changes associated with blood flow, and this technique relies on the fact that cerebral blood flow and neuronal activation are coupled [85]. When an area of the brain is in use, blood flow to that region also increases. It has already been widely used to study brain functions and some mental diseases [90–92].

Graphical Summary of the Introduction

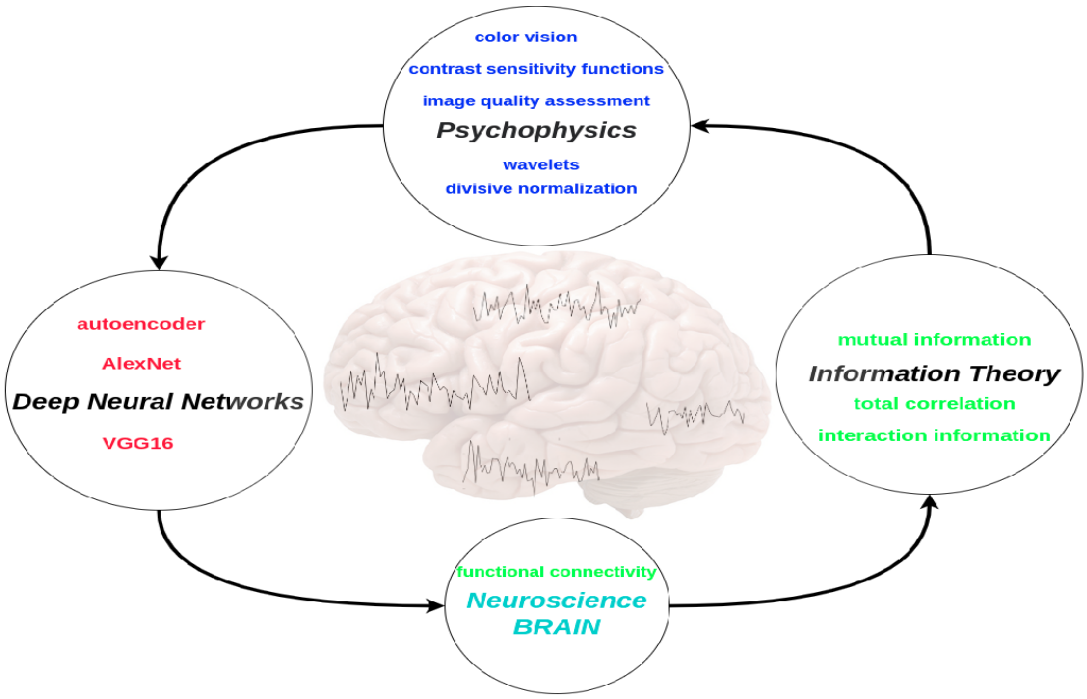


Figure 3: **A brief summary of the introductory concepts** To help the reader more easily and clearly grasp the important concepts of visual functions. Each concept listed in the circle is introduced in the introduction section.

Chapter 3

Methodology and Results

3.1 Functional Connectivity Inference using Multivariate Information Measures

> Without the physical world, Ideas will not exist.
>
> *Joey Lawsin*

3.1.1 Open Issues in Using Information Theory for Connectivity

In 1948, Shannon introduced information theory as the first proposal for resolving issues of information transmission and compression across a noisy communication channel between a sender and a receiver [93]. Learning how information is transferred between neurons, whether on a microscopic or macroscopic scale, is a perennially pressing scientific question. Therefore, understanding brain function requires investigating information exchange across brain areas and how cell activity couples together to drive cognitive activities. Mutual information is a typical information-theoretic method used to quantitatively explain the dependent connection between two random variables [79, 93]. Neuroimaging, an imaging method used to examine brain function and neurological disorders, has long made use of mutual information to describe the interaction between various neurons and gain insight into neural activity [94–96]. The estimation of mutual information from observations may be done in a number of different ways, each with their own set of advantages. Binned techniques, continuous methods, such as Kernel Density Estimation (KDE) [97], naive non-parametric-nearest neighbor (KNN) [98] and various versions of the KNN estimators [99], Gaussian copula-based estimators [100, 101], and so on are all examples of common estimators. To better understand how information is exchanged across different parts of the brain, mutual information has recently emerged as a prominent method for estimating pair-wise functional connectivity [86, 100, 102].

However, while using the aforementioned mutual information estimators, we ran across two issues. It is difficult, if not impossible, to reliably estimate mutual information from a small sample size in many real-world contexts. Finally, we'd want to look at the connections between neurons beyond the pair-wise level, but the aforementioned issues can't be remedied by using generic mutual information.

Total correlation (also known as multi-mutual information) is one expansion of the mutual information theory offered for addressing the aforementioned multivariate variable connection [77]. Redundancy in data may be quantified by calculating the total correlation between a group of variables. It is the disciplines of image statistics and machine learning [103–106], rather than neurology, that do the vast majority of research on total correlation. In biological research, the idea that multivariate variables depend on information is slowly gaining ground [1–3, 107–110].

3.1.2 Methodology and Our Proposal

The primary objective of the **first paper** was to define the variations among mutual information, interaction information, and total correlation as they apply to the field of neuroscience. We propose to apply multivariate measures as opposed to bivariate measures. In addition, we developed a new strategy for calculating the interplay information among three variables by combining total correlation with conditional mutual information. However, how to use it effectively in real life. We provided data from both synthetic and anatomically real neural experiments to support our use of the aforementioned three higher-order information-theoretic methods for functional connectivity estimation in the brain. We found that interaction information and total correlation were both strong at capturing redundancy information for multivariate variables, and that this may allow them to capture both well-known and as-yet-undiscoverable functional brain connections.

3.1.3 Results

In the **first paper**, our results demonstrated that total correlation was able to represent a more nuanced kind of reliance than mutual information. At the same time, it provides a new avenue for research into picture statistics and deep learning tasks including representation learning, ensemble learning, and model distillation, among others. Second, organizing statistical dependencies is challenging for the reasons we've already given. As a result, we used greedy clustering and graph theory to depict functional connectivity between (Region of Interests) ROIs using both resting-state and task-related fMRI data. We demonstrated that high-order information-theoretic models may represent both established and novel forms of functional connectivity in the brain.

Graphical Summary of Connectivity from Information Theory

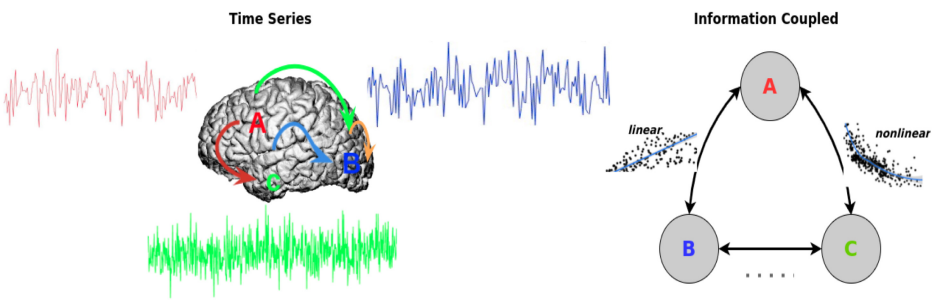


Figure 4: **Neural time series of three inter-brain regions and how they are statistically dependent.** (Left figure) Using a complex system like the brain, we attempt to estimate the underlying information dependencies (Right figure), taking into account linear and nonlinear dependencies.

3.2 Large Scale Functional Connectivity Inference from Total Correlation

> The history of life is written in terms of negative entropy.
>
> *James Gleick*

3.2.1 Methodology and Our Proposal

In our **second paper**, we propose to use Total Correlation to infer a large-scale (whole-brain) connection network, and we demonstrate the potential of such networks as biomarkers of brain changes. In particular, Total Correlation is estimated by the application of Correlation Explanation (CorEx). To begin, we establish that the overall correlation and clustering result estimations generated by CorEx are reliable when compared to the truth, and through the synthetic experiments with known ground truth, we are then ready to address the study of real data.

3.2.2 Results

We conducted tests on standard resting-state fMRI datasets to determine the efficacy of the total correlation. First, the larger open fMRI datasets yield an inferred large-scale connectivity network that is consistent with previous neuroscience research but, intriguingly, can estimate additional relations beyond pair-wise areas. Lastly, we show that connection graphs based on total correlation are a good way to find neurological disorders. Second, our research showed that using only pairwise correlation and mutual information values renders the detection of multivariate relationships impossible. In a broader sense, total correlation is the only method that allows for the clustering of multivariate connections. Finding complex functional connections between brain regions is thus critically important, and total correlation measurements play a pivotal role in this process.

Graphical Summary of Large-Scale Connectivity

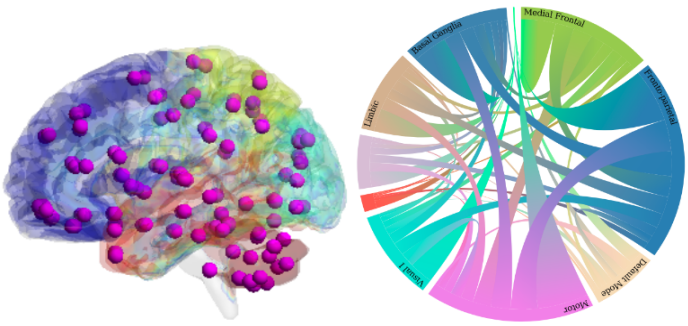


Figure 5: **Conceptual scheme of large scale functional connectivity in the human brain.** The left figure indicates the whole brain's functional areas. The right figures illustrate large-scale functional connectivity in the brain.

3.3 Analytical Results on Functional Connectivity in Visual Areas

> All models are wrong, but some are useful.
>
> *George E. P. Box*

3.3.1 Methodology and Our Proposal

In our **third paper**, which is still being reviewed, we provide an analytical illustration of the benefits of using Total Correlation to characterize the functional connectivity in the visual pathway. In order to do so, we need an analytical network that we can control. Our neurological model consists of three distinct regions (retina, LGN, and V1) and allows for the modification of top-down feedback and inter-nodal connectivity. This analytical framework may also be used to validate empirical estimators of Total Correlation. When an estimator has been shown to be reliable in this control situation, it becomes possible to study its behavior in more general situational settings, for which the analytical conclusions are, in principle, no longer applicable. As such, (a) we use natural pictures to investigate the impact of connection and feedback in the analytic retina-cortex network, and (b) we use real fMRI data to evaluate the functional connectivity in V1, V2, V3, and V4.

3.3.2 Results

Analytical findings demonstrate that whereas the three-way Total Correlation is able to capture the influence of distinct intra-cortical inhibitory connections, the pairwise Mutual Information is unable to do so. Moreover, we demonstrate that, in feasible models of the retina-LGN-V1 that include nonlinearities owing to intra-cortical connection and top-down feedback, Total Correlation is a more accurate descriptor of connectivity than Mutual Information. TC excels where MI fails because it is more attuned to network connections. However, empirical estimates verify that the analytic insights obtained for Gaussian signals also apply to natural inputs. Our TC-scores for responses recorded in visual cortices V1, V2, V3, and V4 show that feedback linkages are stronger in V1, V2, and V3 than in V2, V3, and V4.

Graphical Summary of Analytical Models with Varying Connectivities

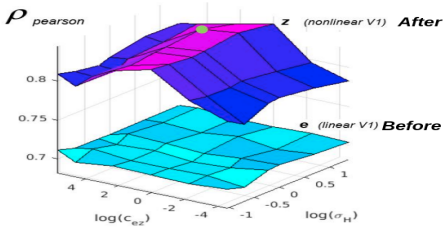


Figure 6: **Normalization and gain of control in visual models.** The correlation with human opinion for different cortical connectivity values. The surface shows the correlation before and after divisive normalization, we see that divisive normalization is critical.

3.4 Similarities and Differences between Biological Vision and Deep Nets

> The sad thing about artificial intelligence is that it lacks artifice and therefore intelligence.
>
> *Jean Baudrillard*

In this section, we investigate the similarities and differences between biological vision and deep networks in two ways. The first study takes a psychophysical approach trying to understand the origin of the Contrast Sensitivity Functions and was published in the Journal of Vision (JoV). The second study takes an information-theoretic approach and is being prepared to be sent to Frontiers in Neuroscience.

3.4.1 Contrast Sensitivity Functions in Autoencoders

Open Issues in Using Autoencoders in Vision

Autoencoders are artificial networks that do these two transformations: they convert the signal into an "inner representation," and then using a decoder, they convert the "inner representation" back into the input domain. Namely:

$$x \xrightarrow{CNN_\theta(x)} y \tag{3.1}$$

Since x and y are in the image space, the encoding and decoding processes are not made explicit in Eq. 3.1. No assumptions about the autoencoder's internal representation are made here. In autoencoders, the picture domain is both the input and the output, so the fundamental target function (reconstruction error) is specified there. Convolutional autoencoders are our main interest since they have been shown to display human-like behaviors as for instance in computing distances between images [111], or in being deceived by visual illusions [112, 113]. In particular, our goal is to study the emergence of CSFs, as the authors in [113] suggested that spatio-chromatic illusions could come from CSF-like filters in the artificial networks.

Different computational aims will be discussed to see if CSFs emerge, and CSF characterization may also be determined in this picture domain with the right stimulus. But for now it is sufficient to assume that the parameters θ are learned to adjust for the blur and noise introduced in the signal during the picture collection process. Given a pristine picture, x_c, the distorted version would be sent into the neural network as input: $x = H \cdot x_c + n_r$, where H is an optical blurring operator and n_r is the noise associated with the response of the LMS photodetectors at the retina. The brain has no concept of H or n_r. At this point, the network is trying to extrapolate x_c from x. This may be one of the outcomes of the biological processing after retinal detection, as shown by accurate models of LGN cells [114].

Methodology and Our Proposal

In the goal of our **forth paper** is reexamining the importance of simple visual tasks in providing an explanation for CSFs. We do this in light of (1) the recent rise in the use of

artificial neural networks in vision research (which include nonlinear behaviors not considered in classic explanations [115–117]), and (2) the current understanding of how images are corrupted in the retina [118]. To do so, we did experiments in which different convolutional neural network autoencoders were trained on natural scenes and cartoon scenes to solve different low-level vision goals, such as compensating for retinal distortions, changes in lighting, information loss after simple bottlenecks (or pure reconstruction after bottlenecks), and combinations of these.

Results

For the most part, we can still make sense of the CSFs by thinking about simple visual tasks, but we found that architecture is not irrelevant, so the classical implementation and computational levels or Marr and Poggio are not that independent. To begin, we demonstrate that autoencoders, a widely used kind of convolutional neural network, may learn to execute certain essential low-level vision tasks (such as retinal noise and optical blur removal) while failing at others, resulting in CSFs that are eerily similar to humans' (such as chromatic adaptation or pure reconstruction after simple bottlenecks). The CSFs are reproduced with a root mean square error of 11% of the maximum sensitivity by the best CNN (among the collection of basic designs evaluated for amplification of the retinal signal). You may use this as an example. Our second contribution is experimental proof that, for certain functional purposes (at low abstraction level), more depth in a convolutional neural network (CNN) does not always result in greater performance when it comes to recreating human-like behavior. The ability to test this theory in a lab setting allowed us to make this discovery (such as the CSFs). This result (for the analyzed networks) does not necessarily contradict previous research that demonstrates the advantages of deeper nets when replicating higher-level vision goals via modeling. Our results add to the increasing body of literature warning against the widespread use of CNNs in vision science. This is because it is possible that the modeling and comprehension of human vision may be hindered if unrealistic designs or simplified units are used in the process of goal optimization.

Graphical Summary of the Definition of CSFs for Autoencoders

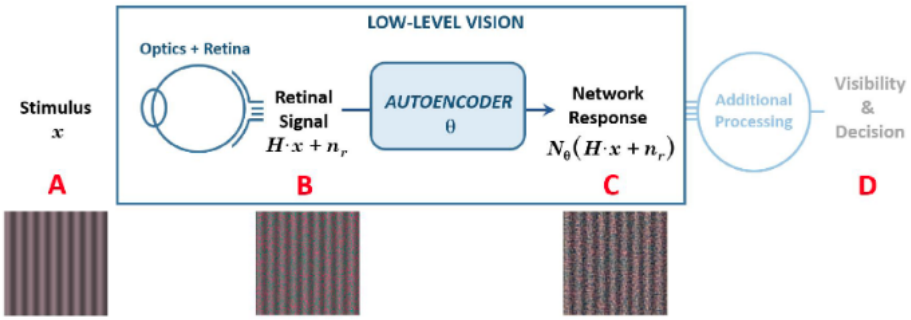


Figure 7: **The CSF is here defined as a frequency-dependent attenuation factor in a system to develop elementary visual tasks.** In this diagram, we see how the visual signal is altered as it travels from the input stimulus (A) to the degraded signal due to optical blur and retinal noise (B) to the process of the early neural path where the output is still in a spatial LMS representation (C), modeled here by autoencoders, and other mechanisms that compute a decision on the visibility of an object (D). Assuming a linear process from A to C, the standard filter model for the CSF assumes that the amplitude of the response at point C determines whether or not gratings are visible to humans. In human psychophysics (without access to C), the observer decides on visibility, and attenuation factors are calculated from thresholds.

3.4.2 Data Processing Inequality in Biological and Artificial Vision Nets (on-going work)

As we mentioned at the beginning of this section, this part of the work is still under preparation to be submitted to Frontiers in Neuroscience. Therefore, it has no associated publication in PART II yet and results have to be expanded to be conclusive. However, we decided to include it in as a part of the PhD given its relation with the analysis of information flow in the visual areas (in Neur. Nets. under review [3]) and with the analysis of the human-like behavior of some artificial nets (in the JoV paper [4]). As a result, compared to the succinct summary of other sections, here we will provide more details for this on-going work.

Open Issues in using Information Theory to Compare the Brain and Deep Nets

Given the capacity of the visual brain, redundancy reduction seems key to process information quickly and efficiently, and a lot of research has been done to figure out its role [119–121]. As stated before in this Thesis, the pair-wise mutual information may be a limited tool given the complicated multi-node interactions that may happen amongst multiple brain regions. In parallel, understanding the black box of deep learning is always a key and hot scientific question. The usual way is by visualizing hidden feature representation in convolutional neural networks [122, 123]. Here we will do it in a different, more quantitative way: by measuring information in CNNs. Shannon information supply us a potential chance to do that and there already large body related studies that investigated it [124–126]. The autoencoder, which is one of the classical CNN architectures, has attracted a lot of research on information compression, information bottleneck, and information decompression [127, 128]. Tishby is a pioneer who explored the alteration of mutual information during training autoencoders, and first opened the black box of Deep Neural Networks (DNNs) from an information-theoretical perspective [125, 126, 129]. As previously stated, there is a substantial body of research on measuring information change in CNNs, but almost all of these studies only used mutual information to quantify it, and as we go down in the layers of a DNN, the mutual information between the layer and the input can only decrease since the network is being thought of as a Markov chain in which information is transmitted from one node to the next. As we know, mutual information can only measure pair-wise information. It has limitations in scenarios where there are complicated interactions among multiple nodes.

Our Proposal: Check the data processing inequality in the Visual Brain and in Standard Deep Nets subject to the Same Visual Stimuli

Here, we propose to use total correlation to quantify redundancy and information flow among different layers in CNNs and among different cortical regions in the visual brain. In both cases (visual brain and artificial nets) we will use the same information estimates based on RBIG [80] because RBIG has been shown to work according to analytical predictions in plausible vision models: from toy 3-dimensional scenarios [113, 130], to 30-dimensional scenarios [32], up to to 512-dimensional scenarios [3]. In both cases (visual brain and artificial nets) we will use the same stimuli as input to the system using the database of the Algonauts Project [131], using their experimental signals from humans and, in our case, geting the corresponding signals from the artificial networks. Estimating mutual information and total correlation between each layer of deep neural networks and each visual region of cortex using the same stimuli and information estimation has not been done before. This could open an interesting window to compare similarities between deep neural networks and biological vision pathways from an information-theoretical perspective.

Materials I: Signals from the Visual Brain (regions V1, V2, V3, and V4)

In order to compare information flow in artificial neural networks with biological visual pathways, here we used data from The Algonauts Project 2021 Challenge, in which biological visual neural signals were collected from V1, V2, V3, and V4 when 10 human participants viewed a rich set of over 1,000 short video clips depicting everyday events [131]. See Figure. 8 for the location of V1, V2, V3, and V4.

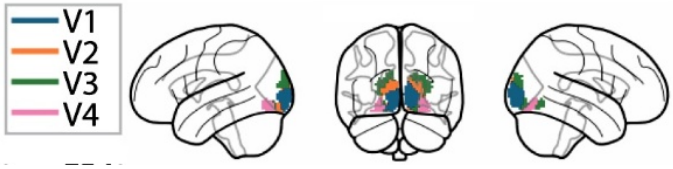


Figure 8: **The low-level and mid-level visual cortex.** The visual response of V1, V2, V3, and V4 to video clips. The figure is adapted from [131].

Materials II: Deep Networks in Vision

AlexNet[1]**:** Here we employed a deep CNN (referred to as the "AlexNet") to extract hierarchical visual information. In the Large Scale Visual Recognition Challenge 2012[2], the model had been pre-trained to produce the best object recognition results [59]. Alex net is an 8-layer convolutional neural network used to classify images and objects, with the first five layers being convolutional and the latter three being fully linked. See the architecture in the Figure. 9. Each convolutional plus nonlinear layer and the dimensionality reduction operations filter out some of the incoming information, and it is interesting to quantify the amout of neglected information along the way.

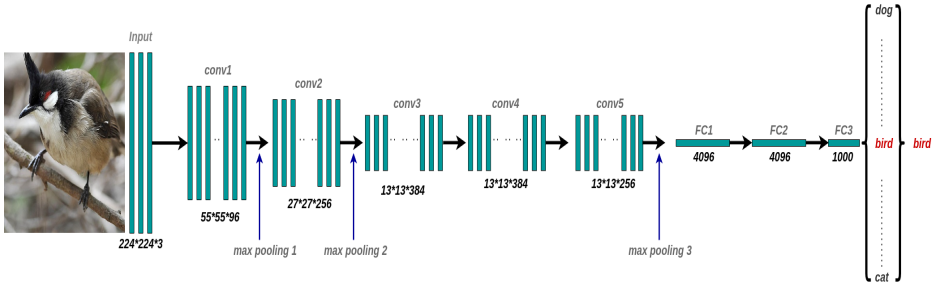


Figure 9: **The AlexNet architecture.** The convolutional layers and max pooling stages are labeled in different ways. The numbers indicate the size of the feature maps. The last vector refers to a fully connected layer with 1000 classes and the prediction comes from the class with higher probability value.

VGG16[3]**:** A more advanced version of the popular CNN family, VGG16 is widely regarded as one of the most effective computer vision models available today. To classify 1000 images into 1000 distinct categories, VGG16 uses an object detection and classification algorithm with an accuracy of 92.7% [132]. See the architecture in the Figure. 10.

[1]https://www.mathworks.com/matlabcentral/fileexchange/59133-deep-learning-toolbox-model-for-alexnet-network

[2]https://www.image-net.org/challenges/LSVRC/2012/

[3]https://www.mathworks.com/matlabcentral/fileexchange/61733-deep-learning-toolbox-model-for-vgg-16-network?s_tid=prof_contriblnk

The VGG16 network architecture contains a total of 21 layers, consisting of 13 convolutional layers, 5 max pooling layers, and 3 dense layers, but only 16 layers with weights. There are 64 filters in Conv-1, 128 in Conv-2, 256 in Conv-3, 512 in Conv-4, and 512 in Conv-5. After a series of convolutional layers, three Fully-Connected layers are applied, the first two of which have 4096 channels each, and the third of which conducts 1000-way ILSVRC classification and so it has 1000 channels (one for each class). In our experiments we *qualitatively* chose the most similar abstract features in the AlexNet and VGG16 (see Figure. 12).

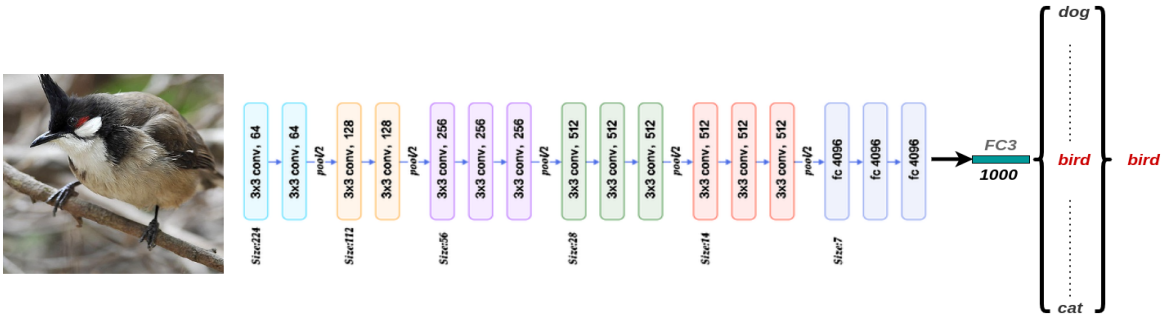


Figure 10: **The VGG16 architecture.** The numbers indicate the size of the feature maps. The last vector refers to fully a connected layer with 1000 classes and the predictions come from the class with higher probability.

DeepDream: The DeepDream is a standard way of seeing the features that a neural network has learned [133, 134]. To accomplish this, an initialization image is sent through the network, and the gradient of the image relative to the activations of a specific layer is computed. After that, the image is iteratively changed to make these activations stronger (see Figure. 11 and Figure. 12 for the result for AlexNet and VGG16 using the Matlab implementation of the method[4]). This illustrates the kind of images that optimally excite certain neurons (or layers) of these networks.

These illustrations are *kind of consistent* with studies that claim that the visual brain and deep networks share some functional similarities as for instance progressively more abstract features [65, 67, 135].

Methods I: Information Transmitted in Deep Nets with Inner Noise

Following [32, 42], we are going to study the information lost in the deep nets, given the fact that the inner representation may be subject to a certain amount of noise. Just for the sake of the illustration, following [32], we assumed that the amplitude of the noise is 5% of the amplitude of the responses in each layer. Noise is a relevant factor in the loss of information because in the transform $x \rightarrow y$, the information about the input x shared by the noisy response, y is [32]:

$$I(x,y) = \sum_i h(y_i) - TC(y) - h(n) \tag{3.2}$$

where $h(\cdot)$ stands for the (univariate or joint) entropy of the corresponding (scalar or vector) variables, $TC(y)$ refers to total correlation of y, and $h(n)$ indicates the joint entropy of the noise in the response. When the energy budget is restricted the entropy of the response

[4]https://www.mathworks.com/help/deeplearning/ref/deepdreamimage.html#d124e47961

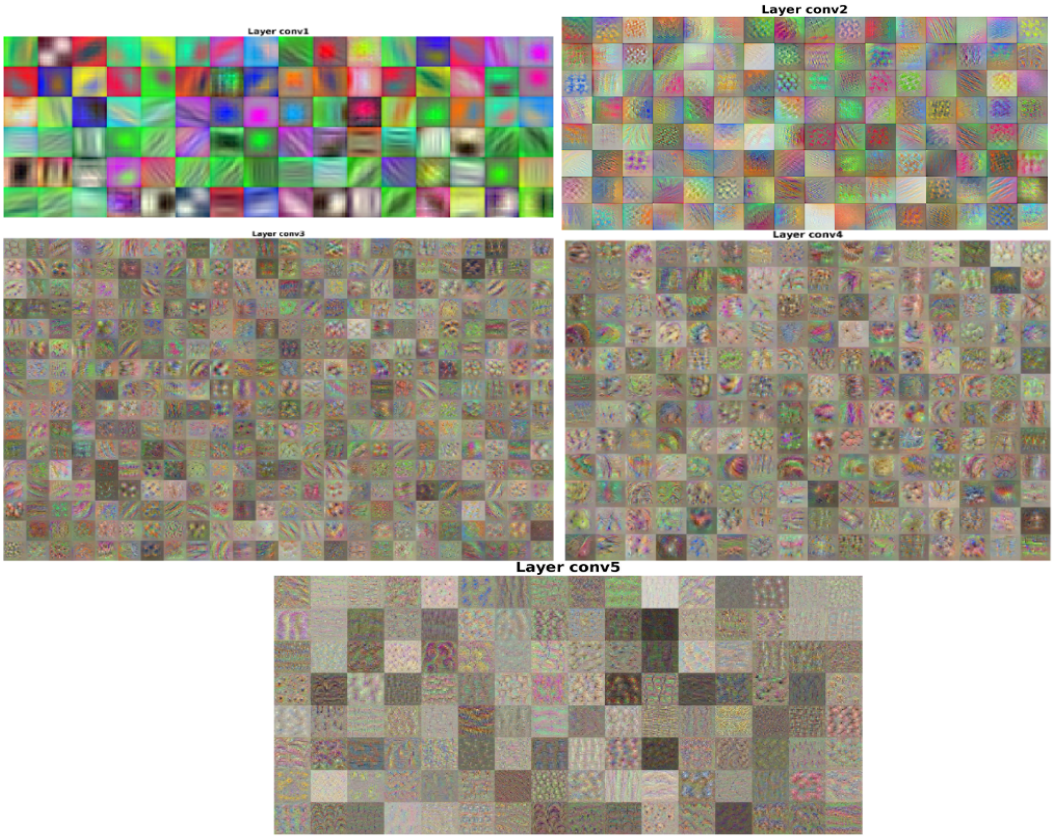


Figure 11: **DeepDream features in AlexNet.** These features were computed through the current Matlab implementations of AlexNet and the DeepDream method (see previous footnotes).

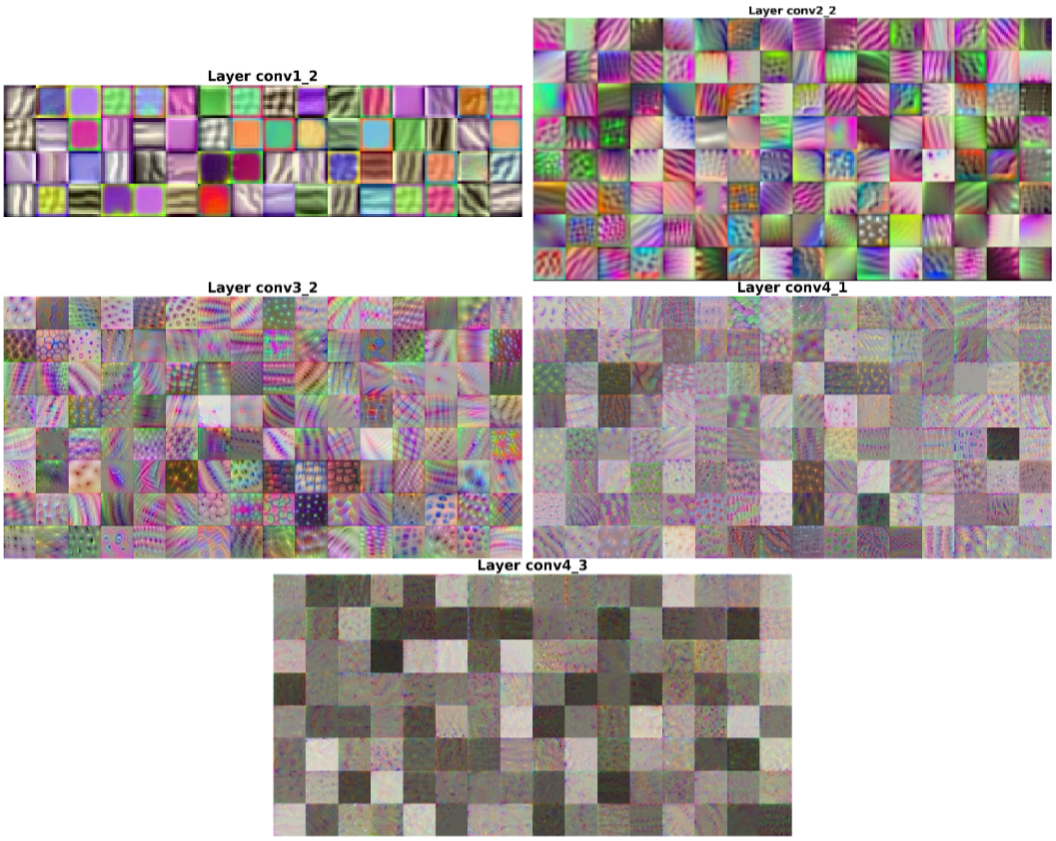


Figure 12: **DeepDream features in VGG16.** These features were computed through the current Matlab implementations of VGG16 and the DeepDream method (see previous footnotes).

cannot increase arbitrarily and hence redundancy TC and noise n contribute to the loss of information.

Methods II: Information Measured with Rotation-Based Iterative Gaussianization

The RBIG is a cascade of L *nonlinear+linear* layers, and the l-th layer is made of marginal Gaussianizations, $\Psi^{(l)}(x^{(l)})$, followed by a rotation, $R^{(l)}$. Each of such layers is applied on the output of the previous layer:

$$x^{(l+1)} = R^{(l)} \cdot \Psi^{(l)}(x^{(l)}) \tag{3.3}$$

For a big enough number of layers, this invertible architecture is able to transform any input PDF, $p(x^{(0)})$, into a zero-mean unit-covariance multivariate Gaussian even if the chosen rotations are random [80] (see Figure. 13).

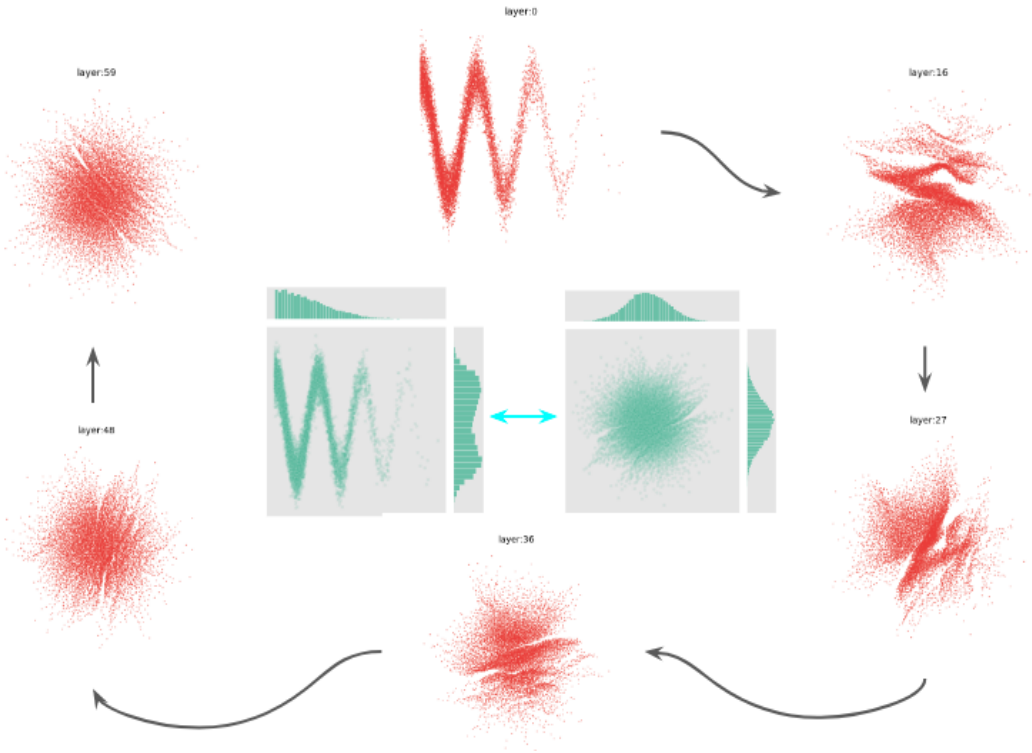


Figure 13: **RBIG transforms arbitrary data distribution into a Gaussian distribution.** The *layer* : 0 refers initial data distribution and each dimensional distribution is non-Gaussian (middle left green figure) then after 60 layers PCA plus marginal Gaussian, the final data distribution become to Gaussian distribution in each dimensional (middle right green figure).

Theoretical convergence to a Gaussian is obtained when the number of layers tends to infinity. However, in practical situations early stopping criteria can be proposed taking into account the uncertainty associated with a finite number of samples [80]. Convergence even with random rotations implies that both elements of the transform are straightforward: univariate equalizations and random rotations. The differentiable and invertible nature of RBIG make it a member of the Normalizing Flow family [136]. Within this general family, differentiable transforms with the ability to remove all the structure of the PDF of the input data are referred to as *density destructors* [137]. By density destruction, the authors in [137] mean a transform of the input PDF into a unit-covariance Gaussian or into a d-cube aligned with the axes. The considered Gaussianization [80, 130, 138] belongs to this family by definition. *Total*

correlation describes the redundancy within a vector, i.e., the information shared by the univariate variables [77, 78]. Note that strong relations between variables indicates a rich structure in the data. Density destruction together with differentiability is useful to estimate the total correlation within a vector $TC(x^{(0)})$. Imagine that the considered RBIG transforms the PDF of the input $x^{(0)}$ into a Gaussian through the application of L layers (L individual transforms). As the redundancy of the Gaussianized signal, $g_x(x^{(0)}) = x^{(L)}$, is zero, the redundancy of the original signal, $TC(x^{(0)})$, should correspond to the cumulative sum of the individual variations, $\Delta TC^{(l)}$ with $l = 1, \dots L$, that take place along the L layers of RBIG, while converting the original variable x, into the Gaussianized variable $g_x(x)$. Interestingly, the individual variation in each RBIG layer only depends on (easy to compute) univariate negentropies, therefore, after the L layers of RBIG, the total correlation is [80]:

$$TC(x) = \sum_{l=1}^{L} \Delta TC(x^{(l-1)}, x^{(l)}) = \sum_{l=1}^{L} J_m(x^{(l)}) \tag{3.4}$$

where the marginal negentropy of a d-dimensional random vector is given by a set of d univariate divergences $J_m(v) = \sum_{i=1}^{d} D_{\mathrm{KL}}(p(v_i)|\mathcal{N}(0,1))$. Therefore, using RBIG, the challenging problem of estimating one d-dimensional joint PDF to compute $TC(x)$ reduces to solve $d \times L$ univariate problems. RBIG estimate of total correlation has been shown to be better than previously reported estimates of TC [130], and it has been successfully used in neuroscience to quantify the redundancy reduction over the visual pathway [32, 113, 130].

Preliminary results

We separately examined the functionalities of artificial neural networks and biological neural networks, and we compared the flow of information in pre-trained deep convolutional neural networks, such as AlexNet and VGG16, with the human visual system, such as V1-V2-V3-V4. It opens the window to comparing similarities between deep neural networks and biological vision pathways from an information-theoretical perspective. The videos are fed into pre-trained deep neural networks, and we selected 16 frames in each video. We have a total of 16×1000 samples. Then we extract the response in each layer of artificial neural networks; after we save the response in each layer and consider the dimensioning problems and related computing efficiency, we read the corresponding responses, resample to put all of them in the same spatial resolution (6×6), and then we concatenate all the responses in each layer. After this, we are ready to estimate the information measures in deep networks. The information representation is estimated based on mutual information and total correlation (see Table. 1, Table. 2, and Table. 3).

$\mathbf{I(V_i, V_j)}$ (in *bits/visual region*)	V_1	V_2	V_3	V_4
V_1	$\mathbf{2.4} \pm 0.3$	$\mathbf{1.3} \pm 0.2$	$\mathbf{1.0} \pm 0.2$	$\mathbf{0.8} \pm 0.1$
V_2	1.3	$\mathbf{2.0} \pm 0.2$	$\mathbf{1.2} \pm 0.2$	$\mathbf{0.7} \pm 0.1$
V_3	1.0	1.2	$\mathbf{1.7} \pm 0.3$	$\mathbf{0.8} \pm 0.1$
V_4	0.8	0.7	0.8	$\mathbf{2.2} \pm 0.3$
$\mathbf{TC(V_i)}$ (in *bits/visual region*)	V_1	V_2	V_3	V_4
	$\mathbf{3.6} \pm 0.3$	$\mathbf{3.2} \pm 0.2$	$\mathbf{3.0} \pm 0.2$	$\mathbf{3.5} \pm 0.3$

Table 1: **The mutual information and total correlation in biological visual pathways. $\mathbf{I(V_i, V_j)}$** between pairs of areas, $\mathbf{TC(V_i)}$ in each area.

$\mathbf{I(L_i, L_j)}$ (in *bits/layer*)	$Layer_1$	$Layer_2$	$Layer_3$	$Layer_4$
$Layer_1$	**2.3** $\pm$ 0.3	**1.3** $\pm$ 0.3	**0.8** $\pm$ 0.1	**0.6** $\pm$ 0.2
$Layer_2$	1.3	**1.9** $\pm$ 0.1	**1.3** $\pm$ 0.3	**0.6** $\pm$ 0.1
$Layer_3$	0.8	1.3	**2.0** $\pm$ 0.2	**1.8** $\pm$ 0.1
$Layer_4$	0.6	0.6	1.8	**2.0** $\pm$ 0.2
$\mathbf{TC(L_i)}$ (in *bits/layer*)	$Layer_1$	$Layer_2$	$Layer_3$	$Layer_4$
	3.6 $\pm$ 0.1	**5.6** $\pm$ 0.3	**6.0** $\pm$ 0.3	**5.8** $\pm$ 0.2

Table 2: **The mutual information and total correlation in artificial neural networks (AlexNet).** $\mathbf{I(L_i, L_j)}$ between pairs of layers, $\mathbf{TC(L_i)}$ in each layer.

$\mathbf{I(L_i, L_j)}$ (in *bits/layer*)	$Layer_2$	$Layer_4$	$Layer_6$	$Layer_8$
$Layer_2$	**2.0** $\pm$ 0.3	**1.5** $\pm$ 0.2	**0.1** $\pm$ 0.1	**0.03** $\pm$ 0.2
$Layer_4$	1.5	**1.9** $\pm$ 0.2	**0.1** $\pm$ 0.3	**0.02** $\pm$ 0.3
$Layer_6$	0.1	0.1	**0.05** $\pm$ 0.2	**0.01** $\pm$ 0.1
$Layer_8$	0.03	0.02	0.01	**0.05** $\pm$ 0.1
$\mathbf{TC(L_i)}$ (in *bits/layer*)	$Layer_2$	$Layer_4$	$Layer_6$	$Layer_8$
	0.33 $\pm$ 0.1	**0.38** $\pm$ 0.2	-	-

Table 3: **The mutual information and total correlation in artificial neural networks (VGG16).** $\mathbf{I(L_i, L_j)}$ between pairs of layers, $\mathbf{TC(L_i)}$ in each layer.

There hasn't been much research on explaining the similarities between the brain and deep neural networks via information before. Most of the research focuses on fitting deep neural networks with real neural signals, then predicting neural response with well-fitted models. Here, we try to explore this question from an information-theoretical views, and it would be open a door for us to study brain and deep neural networks via information. From an information-theoretical perspective, we found that the human brain and deep neural networks do share information representation similarities. Meanwhile, we found that neural information processing in the brain also matches the properties of data processing inequality, for instance, see Table. 1, Table. 2, and Table. 3.

Further Research Directions

Here we only checked information flow in the pre-trained AlexNet and VGG16, and we collected neural response in each layer of neural networks with fed videos, and it could be extended to other more deep neural networks with different architectures. On the other hand, as we mentioned above, we used weights to freeze neural networks, and we could get different results if we retrained deep neural networks with new image or video datasets, then explored how the weights affect information representation in deep neural networks. Meanwhile, we could also measure information representation in neural networks with different visual tasks. Furthermore, we used fMRI to investigate the circulation of data in the visual cortex; future research might investigate the use of other modalities' datasets (e.g., EEG, MEG) to better estimate the extent to which data is represented in each visual region. All in all, we can explore information theory in deep neural networks by exploring their architecture and

functional goals, which would help us understand information altered in neural networks and the brain.

Graphical Summary of Information Transmission in the Brain and DNNs

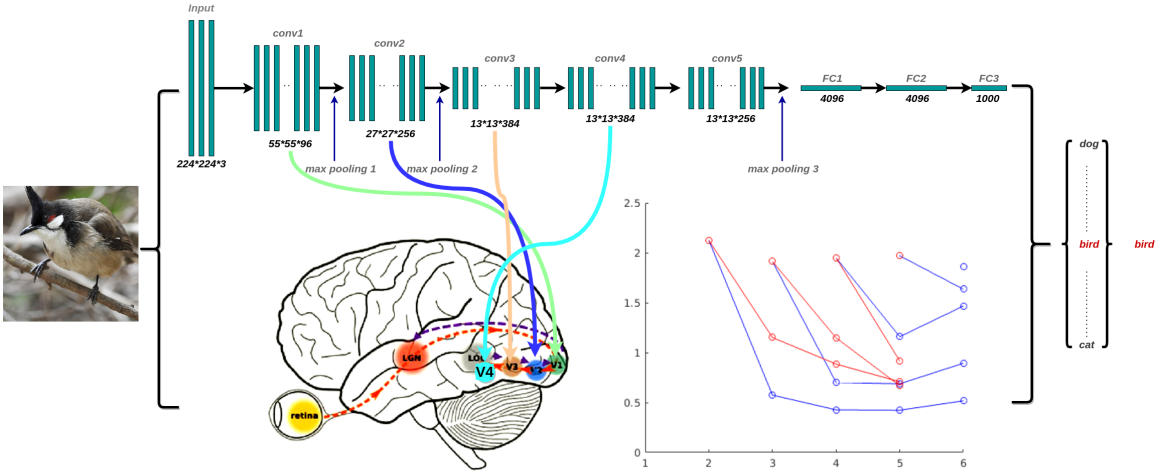


Figure 14: **Information flow in biological vision pathways and artificial neural networks.** The information representation between deep neural networks and biological vision was evaluated from an information-theoretical perspective. The information measured in the visual brain and deep networks was shown with a red and blue curve, respectively.

Chapter 4

Conclusions

The thesis investigated brain functions through psychophysics, deep neural networks, and information theory. On the one hand, we were very interested in quantifying information flow among brain regions in general and among layers of the visual system in particular. We proposed to measure that flow via multi-node concepts, as for instance, total correlation, as opposed to the conventional (just pair-wise) mutual information. As a consequence of this advantage, we proposed the use of total correlation to measure functional connectivity (both in the whole brain, and in the visual regions). We found that total correlation maybe more sensitive to the connectivity in networks than mutual information. As a result, we think that this new metric of functional connectivity could be applied in the future to answer other problems in neuroscience. On the other hand, some low-level visual functions were studied through the analysis of psychophysics and deep neural networks. In particular we studied the basic bandwidth of the human vision: the spatio-temporal and chromatic Contrast Sensitivity Functions (CSFs). We tried to explore possible functional explanations for such psychophysical bottleneck using current autoencoders which can be trained for different goals. We found that while some functions (like the enhancement of the retinal signal) are more likely to be behind the CSFs than others (like chromatic adaptation or like pure reconstruction after dimensionality reduction), the architecture of the algorithm that optimizes the goal is also relevant and it cannot be separated from the functional discussion.

Finally, following the above work on information theory in neuroscience (three papers on total correlation) and on deep nets devoted to vision (one paper on autoencoders for chromatic video), we presented promising results on the combination of these concepts: the quantification of the information loss along natural and artificial visual networks.

4.1 Contributions

- **Overcoming the pairwise limitations of mutual information in assessing functional connectivity through multi-node alternatives like total correlation**. It is always a key problem to quantify the flow of information in biological systems, and information theories have quickly become important tools in investigating information coupling among brain regions. Mutual information is a common metric that is used in functional connectivity, but it only measures shared information between two sets while relations in the brain involve more than two nodes. In order to overcome this limitation we proposed concepts like interaction information and total correlation. In addition, we proposed a novel method for determining the interaction information between three variables by making use of total correlation and conditional mutual information. On the other hand, how to correctly apply it in real-world scenarios remains a mystery.

Using the higher-order information-theoretic approaches, we estimated functional connectivity in the brain using simulation experiments as well as real neural studies. We discovered that interaction information and total correlation were both robust, and they could be used to capture both well-known and yet-to-be-discovered functional brain connections. This contribution was published in *Neural Networks* [1].

- **Derivation of large-scale functional connectomes and neural biomarkers using total correlation**. We investigated large-scale functional relationships (among hundreds of brain regions) using total correlation. Moreover, we proposed the use of the networks derived from total correlation as eventual biomarkers of brain disease. We found that total correlation identifies networks that differ from the ones obtained by pairwise approaches. As a result, it has the potential to lead to new biomarkers in a variety of brain diseases. This contribution was published as *feature paper* in *Entropy* [2].

- **Analytical results on functional connectivity among visual areas via total correlation and mutual information**. We used a plausible four-layer neural model of the retina, the LGN and the V1 cortex as an analytical scenario to control the connectivity between the layers and within V1. Variants of this model include nonlinearities and feedback. We provide analytical results for all possible pairwise mutual information measures and total correlation measures among the four layers of this model. The effect of individual inhibitory connections within the cortex is captured by the multi-node total correlation but not by the pairwise mutual information, Moreover, in case of feedback, total correlation is more sensitive to the connectivity than mutual information. The above is seen both in the analytical results and in the numerical experiments. The presented analytical framework can also be used to verify empirical estimators of the total correlation. Interestingly, empirical results for natural images (which do not follow the Gaussian assumption of the analytical results) follow the trends of the analytical expressions. Finally, we used fMRI recordings to examine the functional connections between real visual regions V1, V2, V3, and V4. This contribution is under review in *Neural Networks* [3].

- **Emergence of human-like psychophysical bottlenecks in autoencoders devoted to vision**. We show that a common type of convolutional neural network (the autoencoder), has the potential to develop human-like contrast sensitivity in the spatio-temporal and chromatic dimensions when trained to perform some fundamental low-level vision tasks (such as removing retinal noise and optical blur) but not others (like chromatic adaptation or pure reconstruction after simple bottlenecks). As an illustration, the best video-enhancer autoencoder (among the simple architectures that have been contemplated) is able to reproduce the CSFs with an 11% RMSE error. Our second finding is that we provide experimental evidence for the fact that, for certain functional goals (at low abstraction level), deeper CNNs that are better at reaching the quantitative goal are actually worse at replicating human-like phenomena. This is in line with a growing body of research that questions the blind use of deep-learning models in vision science. In particular, if oversimplified units or unrealistic architectures are used in goal optimization, results may be misleading. This contribution was published in the *Journal of Vision* [4].

- **Data processing inequality in the visual brain and in artificial neural networks**. Some researchers have pointed out the qualitative similarity between progressively deeper layers of networks such AlexNet and VGG and progressively deeper visual regions such as V1-V4 in the cortex [65, 67, 135]. However, accurate estimations of mutual information and total correlation between each layer of deep neural networks

and each visual region of cortex using the same stimuli and the same information estimates has not been done before. This could be a new way to compare deep neural networks and biological vision. We presented experiments in which standard pre-trained artificial networks are stimulated with the same images shown to the human observers that participated in the Algonauts project [131]. Our preliminary results using Gaussianization estimates among different layers show similarities and differences between the visual brain and the artificial networks. Both systems progressively disregard visual information along the way, but the rate at which this happens is different: it seems that the visual brain nodes share a bigger fraction of the initial information than the artificial layers. This is still on-going (unpublished) work, but we think it perfectly fits in the PhD because of two reasons: (1) its novelty, and (2) it is connected with the estimations of information-theoretic measures in V1-V4 areas of our third publication [3], and with our analysis of artificial nets devoted to vision in our fourth publication [4].

4.2 Future Works

- In the last part of this thesis, we only showed total correlation application in the visual system and functional connectivity with fMRI. In the future, we hope to incorporate the functional connectivity associations discovered through total correlation into existing graph neural networks for readable brain illness detection. This will enable medical professionals to look on the subgraphs that provide the most information that contributes to the overall diagnosis (e.g., autism patients or health-control groups). In order to evaluate and improve the qualitative results that have been shown here, it will be necessary to extend the analytical results to a greater number of nodes and to build quantitative metrics that can quantify differences between graphs. Both of these things are required. Estimates of pairwise linkages are created using the linear correlation coefficient in all of the newly provided methods, which generally disregard high-order dependencies. Furthermore, it is still not widely applied in neuroscience applications, such as spiking neural data and so on. Future work will involve extending total correlation and exploring the possibility of applying total correlation to mental diseases.

- Although total correlation can reduce the pair-wise problem of other connectivity measures, it does not quantify the synergy information and instead just measures the redundant information. In the first place, taking into account higher-order interactions involving more than three variables makes it possible to quantify the informative content in terms of its complementary and overlapping relationships. The term "redundancy" refers to the information that is shared across variables, whereas "synergy" refers to the statistical interactions that can be found collectively in the whole but not in the parts when considered in isolation. Therefore, we can extend total correlation to more deep levels, and through combined dual total correlation, which allows us to instantly measure redundancy and synergy in the multiple brain networks.

- Moreover, we should think about adding total correlation to deep neural networks, either as a metric or a loss function, and then using this to optimize the training of deep neural networks.

- Most models only focus on feedforward connectivity and do not consider lateral and feedback connectivity. Many experiments have already proved that lateral and feedback connectivity exist in neural pathways. Therefore, we should consider that sensory-driven information flow is put into context by lateral interaction with related

feature representations and top-down signals to be re-entered at different stages in the hierarchy, and implement feedforward, lateral, and feedback together when we go to model visual information processing. On the other hand, we also need to optimize normalization gain control in the model, and one of the most important components in the model should be divisive normalization. There are too many flexible parameters that need to be optimized in order to maximally psycho-physically optimize behaviors.

- We should not ignore the differences between deep neural networks and biological vision because they can help us optimize deep neural network architectures and then make them more functionally similar to the brain. Finally, we should check deep neural networks' generality ability with some low-level visual tasks when we compare them to human vision, and not just keep our eyes on some high-level visual tasks, for example, classification, segmentation, and so on. While the architecture of CNN can be used to simulate high-level human visual systems, it is unknown whether or not it can also be used to model low-level visual functions such as color perception, orientation, brightness reduction, and adaptation. Last but not least, the reliability of saliency prediction can be verified by adding further variety deterioration to natural images that were previously clean as well as images that have been generated using psychophysical methods. As was previously said, we need human saliency prediction data on damaged images to serve as a baseline against which the accuracy of the model can be tested. This may be done by comparing the model's predictions to human predictions. And continuing in that direction would be quite exciting for the years to come. In conclusion, when researching the properties of artificial neural networks, it is important to take psychophysical techniques into consideration. The numerous psychophysical experiments on visual attention mechanisms that have been carried out by scholars in the field of visual neuroscience have been of significant use to the discipline as a whole. Despite the fact that this area has the potential to improve the robustness of artificial neural networks, there is a dearth of research being conducted in it. Incorporating the findings of psychophysical investigations into an artificial neural network would be fascinating, but it would be even more fascinating if this led to an increase in our knowledge of both the human visual system and the inner workings of the network. This would be the case because it would reveal how the network itself functions. In conclusion, we proposed that the subfields of visual neuroscience and artificial neural networks might gain something from the further investigation of CNNs via the lens of psychophysical methods.

- Finally, we should try to estimate a new total correlation estimator and improve the accuracy of total correlation performance. we need to construct a new estimate for total correlation because the two estimators that we utilized in our earlier research had some limits. This is the reason why we feel it is necessary to do so. For us to be able to easily and accurately estimate total correlation and dual total correlation, it is important for us to design a new efficient and less compute-intensive cost estimator. After we have developed methods for measuring redundancy and synergy, we will be able to quantify the information representation in artificial neural networks as well as biological neural networks by using the aforementioned metrics. The similarities and contrasts between artificial neural networks and biological neural networks are a major problem for unearthing the mystery behind both the human brain and artificial neuron networks, as we indicated in the preceding section. Only the AlexNet and VGG16 architectures of deep neural networks were investigated in this study; however, it is possible that other architectures, such as recurrent neural networks, might also be utilized. In addition, the present models have the weights frozen, and then they extract the features from each layer. This may cause some bias when we compare the responses from artificial neural networks and the brain. As a result, on the one hand, it

might be necessary to retrain the neural networks using a dataset that contains several modalities. On the other hand, we need to measure the information representation during the training and validating phases, and this will be a more fascinating task than working with static neural networks. Last but not least, it will be very interesting for us to investigate the representation of spatial statistics and color information in each layer of neural networks. All of these aspects are very important for comprehending low-level and high-level human visual functions, as well as those of artificial neural networks.

Acknowledgement

The research activities leading to this thesis were supported by the these spanish/european grants from GVA/AEI/FEDER/EU: MICINN PID2020-118071GB-I00, MICINN PDC2021-121522-C21, and GVA Grisolía-P/2019/035.

Journal Publications

- **Qiang Li**, Functional connectivity inference from fMRI data using multivariate information measures, Neural Networks, Volume 146, 2022, Pages 85-97. doi:10.1016/j.neunet.2021.11.016.
- **Qiang Li**, Greg Ver Steeg, Shujian Yu and Jesús Malo. "Functional Connectome of the Human Brain with Total Correlation" Entropy 24, no. 12: 1725, 2022.
- **Qiang Li**, Greg Ver Steeg, and Jesús Malo. "Functional Connectivity in Visual Areas from Total Correlation." arXiv preprint arXiv:2208.05770 (2022). [Under Review of Neural Networks Journal]
- **Qiang Li**, Alex Gomez-Villa, Marcelo Bertalmío, Jesús Malo; Contrast sensitivity functions in autoencoders. Journal of Vision 2022;22(6):8. doi: https://doi.org/10.1167/jov.22.6.8.

Related articles

- **Qiang Li**, Jesús Malo. Data Processing Inequality in Biological Vision and Convolution Neural Networks. [To be submitted to Frontiers in Neuroscience]

Related conference publications

- **Qiang Li**, Emmanuel Johnson, Jose Juan Esteve-Taboada, Valero Laparra, Jesús Malo. Computing Variations of Entropy and Redundancy under Nonlinear Mappings not Preserving the Signal Dimension: Quantifying the Efficiency of V1 Cortex, Entropy 2020: The Scientific Tool of the 21st Century, Porto, Portugal. doi:10.3390/Entropy2021-09813.
- **Qiang Li**, Measuring Functional Connectivity of Human Intra-Cortex Regions with Total Correlation, Entropy 2020: The Scientific Tool of the 21st Century, Porto, Portugal. doi:10.3390/Entropy2021-09797.
- Jesus Malo, Benyamin Kheravdar, **Qiang Li**. Visual Information Fidelity with better Vision Models and better Mutual Information Estimates, VSS2021, Virtual. doi:10.1167/jov.21.9.2351.
- **Qiang Li**, Pablo Hernández Cámara, Jorge Vila Tomás, Valero Laparra, Jesús Malo. Basic Psychophysics of Deep Networks Trained via Maximum Differentiation, The 9th Iberian Conference on Perception (CIP 2022), In-person, 6/27th-6/29th, Barcelona, Spain.
- Jorge Vila Tomás, Pablo Hernández Cámara, Alexander Hepburn, **Qiang Li**, Valero Laparra, Jesús Malo. Basic Psychophysics of Deep Networks Trained for Subjective Image Distortion, The 9th Iberian Conference on Perception (CIP 2022), In-person, 6/27th-6/29th, 2022, Barcelona, Spain.
- Pablo Hernández Cámara, Jorge Vila Tomás, **Qiang Li**, Valero Laparra, Jesús Malo. Basic Psychophysics of Deep Networks Trained for Image Segmentation, The 9th Iberian Conference on Perception (CIP 2022), In-person, 6/27th-6/29th, Barcelona, Spain.

Chapter 5

Conclusiones

[1] La tesis investigó las funciones cerebrales a través de la psicofísica, las redes neuronales profundas y la teoría de la información. Por un lado, estábamos muy interesados en cuantificar el flujo de información entre regiones cerebrales en general y entre capas del sistema visual en particular. Propusimos medir ese flujo mediante conceptos multinodales, como por ejemplo la correlación total, frente a la información mutua convencional (sólo por pares). Como consecuencia de esta ventaja, propusimos el uso de la correlación total para medir la conectividad funcional (tanto en todo el cerebro como en las regiones visuales). Descubrimos que la correlación total puede ser más sensible a la conectividad en las redes que la información mutua. Como resultado, pensamos que esta nueva métrica de la conectividad funcional podría aplicarse en el futuro para responder a otros problemas en neurociencia. Por otro lado, se estudiaron algunas funciones visuales de bajo nivel mediante el análisis de la psicofísica y las redes neuronales profundas. En particular, estudiamos el ancho de banda básico de la visión humana: las funciones de sensibilidad al contraste (CSF) espaciotemporales y cromáticas. Intentamos explorar posibles explicaciones funcionales para este cuello de botella psicofísico utilizando autocodificadores actuales que pueden entrenarse para diferentes objetivos. Descubrimos que, aunque algunas funciones (como la mejora de la señal retiniana) tienen más probabilidades de estar detrás de las CSF que otras (como la adaptación cromática o la reconstrucción pura tras la reducción de la dimensionalidad), la arquitectura del algoritmo que optimiza el objetivo también es relevante y no puede separarse de la discusión funcional.

Por último, tras los trabajos anteriores sobre teoría de la información en neurociencia (tres artículos sobre correlación total) y sobre redes profundas dedicadas a la visión (un artículo sobre autocodificadores para vídeo cromático), presentamos resultados prometedores sobre la combinación de estos conceptos: la cuantificación de la pérdida de información a lo largo de redes visuales naturales y artificiales.

5.1 Contribuciones

- **Superando las limitaciones pairwise de la información mutua en la evaluación de la conectividad funcional a través de alternativas multi-nodo como la correlación total**. Siempre es un problema clave para cuantificar el flujo de información en los sistemas biológicos, y las teorías de la información se han convertido rápidamente en herramientas importantes en la investigación de acoplamiento de la información entre las regiones del cerebro. La información mutua es una métrica común que se

[1] Traducido del inglés por Qiang Li

utiliza en la conectividad funcional, pero sólo mide la información compartida entre dos conjuntos, mientras que las relaciones en el cerebro implican a más de dos nodos. Para superar esta limitación propusimos conceptos como información de interacción y correlación total. Además, propusimos un método novedoso para determinar la información de interacción entre tres variables haciendo uso de la correlación total y la información mutua condicional. Por otro lado, sigue siendo un misterio cómo aplicarlo correctamente en escenarios del mundo real. Utilizando los enfoques teóricos de la información de orden superior, estimamos la conectividad funcional en el cerebro mediante experimentos de simulación, así como estudios neuronales reales. Descubrimos que tanto la información de interacción como la correlación total eran robustas y podían utilizarse para capturar conexiones funcionales cerebrales tanto conocidas como por descubrir. Esta contribución se publicó en *Neural Networks* [1].

- **Derivación de conectomas funcionales a gran escala y biomarcadores neuronales usando correlación total**. Investigamos las relaciones funcionales a gran escala (entre cientos de regiones cerebrales) utilizando la correlación total. Además, propusimos el uso de las redes derivadas de la correlación total como eventuales biomarcadores de enfermedades cerebrales. Descubrimos que la correlación total identifica redes que difieren de las obtenidas mediante enfoques por pares. Como resultado, tiene el potencial de conducir a nuevos biomarcadores en una variedad de enfermedades cerebrales. Esta contribución se publicó como *feature paper* en *Entropy* [2].

- **Resultados analíticos sobre la conectividad funcional entre las áreas visuales a través de la correlación total y la información mutua**. Utilizamos un modelo neuronal plausible de cuatro capas de la retina, el LGN y la corteza V1 como escenario analítico para controlar la conectividad entre las capas y dentro de V1. Las variantes de este modelo incluyen no linealidades y retroalimentación. Proporcionamos resultados analíticos para todas las posibles medidas de información mutua por pares y medidas de correlación total entre las cuatro capas de este modelo. El efecto de las conexiones inhibitorias individuales dentro de la corteza es capturado por la correlación total multinodo, pero no por la información mutua por pares, Además, en caso de retroalimentación, la correlación total es más sensible a la conectividad que la información mutua. Esto se observa tanto en los resultados analíticos como en los experimentos numéricos. El marco analítico presentado también puede utilizarse para verificar los estimadores empíricos de la correlación total. Curiosamente, los resultados empíricos para imágenes naturales (que no siguen el supuesto gaussiano de los resultados analíticos) siguen las tendencias de las expresiones analíticas. Por último, utilizamos grabaciones fMRI para examinar las conexiones funcionales entre las regiones visuales reales V1, V2, V3 y V4. Esta contribución está siendo revisada en *Neural Networks* [3].

- **Emergencia de cuellos de botella psicofísicos similares a los humanos en autocodificadores dedicados a la visión**. La modelización de funciones visuales de bajo nivel mejorará la capacidad de generalización de las redes neuronales profundas. Combinar la psicofísica y las redes neuronales profundas será una vía interesante para investigar las funciones visuales y, al mismo tiempo, beneficiará la comprensión de las redes neuronales profundas. Como primera contribución, en el JOV, mostramos que un tipo común de red neuronal convolucional (el autoencoder), tiene el potencial de desarrollar una sensibilidad al contraste similar a la humana en las dimensiones espacio-temporal y cromática cuando se entrena para realizar algunas tareas fundamentales de visión de bajo nivel (como la eliminación del ruido retiniano y el desenfoque óptico) pero no otras (como la adaptación cromática o la reconstrucción pura tras simples cuellos de botella). A modo de ilustración, el mejor autoencoder de mejora de vídeo (entre las arquitecturas simples que se han contemplado) es capaz de reproducir los LCR con

un error RMSE del 11%. Nuestro segundo hallazgo es que proporcionamos pruebas experimentales del hecho de que, para ciertos objetivos funcionales (a bajo nivel de abstracción), las CNN más profundas que son mejores para alcanzar el objetivo cuantitativo son en realidad peores para reproducir fenómenos similares a los humanos. Esto está en consonancia con un creciente cuerpo de investigación que cuestiona el uso ciego de modelos de aprendizaje profundo en la ciencia de la visión. En particular, si se utilizan unidades excesivamente simplificadas o arquitecturas poco realistas en la optimización de objetivos, los resultados pueden ser engañosos. Esta contribución se publicó en el *Journal of Vision* [4].

- **Desigualdad en el procesamiento de datos en el cerebro visual y en las redes neuronales artificiales**. Algunos investigadores han señalado la similitud cualitativa entre capas progresivamente más profundas de redes como AlexNet y VGG y regiones visuales progresivamente más profundas como V1-V4 en la corteza [65, 67, 135]. Sin embargo, hasta ahora no se habían realizado estimaciones precisas de la información mutua y la correlación total entre cada capa de redes neuronales profundas y cada región visual de la corteza utilizando los mismos estímulos y las mismas estimaciones de información. Esta podría ser una nueva forma de comparar las redes neuronales profundas y la visión biológica. Presentamos experimentos en los que redes artificiales estándar preentrenadas son estimuladas con las mismas imágenes mostradas a los observadores humanos que participaron en el proyecto Algonauts [131]. Nuestros resultados preliminares utilizando estimaciones de gaussianización entre diferentes capas muestran similitudes y diferencias entre el cerebro visual y las redes artificiales. Ambos sistemas prescinden progresivamente de la información visual a lo largo del camino, pero el ritmo al que esto ocurre es diferente: parece que los nodos del cerebro visual comparten una fracción mayor de la información inicial que las capas artificiales. Se trata de un trabajo aún en curso (no publicado), pero creemos que encaja perfectamente en el doctorado por dos razones: (1) su novedad, y (2) está conectado con las estimaciones de las medidas de la teoría de la información en las áreas V1-V4 de nuestra tercera publicación [3], y con nuestro análisis de las redes artificiales dedicadas a la visión en nuestra cuarta publicación [4].

5.2 Trabajo Futuro

- En la última parte de esta tesis, sólo mostramos la aplicación de la correlación total en el sistema visual y la conectividad funcional con fMRI. En el futuro, esperamos incorporar las asociaciones de conectividad funcional descubiertas a través de la correlación total en las redes neuronales de grafos existentes para la detección legible de enfermedades cerebrales. Esto permitirá a los profesionales de la medicina fijarse en los subgrafos que proporcionen más información que contribuya al diagnóstico global (por ejemplo, pacientes con autismo o grupos de control de salud). Para evaluar y mejorar los resultados cualitativos que se han mostrado aquí, será necesario ampliar los resultados analíticos a un mayor número de nodos y construir métricas cuantitativas que puedan cuantificar las diferencias entre gráficos. Ambas cosas son necesarias. Las estimaciones de los vínculos entre pares se crean utilizando el coeficiente de correlación lineal en todos los métodos recién proporcionados, que por lo general no tienen en cuenta las dependencias de alto orden. Además, todavía no se aplica ampliamente en aplicaciones de neurociencia, como los datos neuronales de spiking, etc. El trabajo futuro consistirá en ampliar la correlación total y explorar la posibilidad de aplicarla a las enfermedades mentales.

- Aunque la correlación total puede reducir el problema de pares de otras medidas de conectividad, no cuantifica la información de sinergia y en su lugar sólo mide la información redundante. En primer lugar, tener en cuenta las interacciones de orden superior en las que intervienen más de tres variables permite cuantificar el contenido informativo en términos de sus relaciones complementarias y de solapamiento. El término "redundancia" hace referencia a la información que comparten las variables, mientras que "sinergia" se refiere a las interacciones estadísticas que pueden encontrarse colectivamente en el conjunto, pero no en las partes cuando se consideran de forma aislada. Por lo tanto, podemos ampliar la correlación total a niveles más profundos, y mediante la correlación total dual combinada, lo que nos permite medir instantáneamente la redundancia y la sinergia en las múltiples redes cerebrales.

- Además, deberíamos pensar en añadir la correlación total a las redes neuronales profundas, ya sea como una métrica o una función de pérdida, y luego usar esto para optimizar el entrenamiento de las redes neuronales profundas.

- La mayoría de los modelos sólo se centran en la conectividad feedforward y no tienen en cuenta la conectividad lateral y de retroalimentación. Muchos experimentos ya han demostrado que la conectividad lateral y de retroalimentación existe en las vías neuronales. Por lo tanto, deberíamos considerar que el flujo de información impulsado por los sentidos se contextualiza mediante la interacción lateral con representaciones de rasgos relacionados y señales descendentes que se reintroducen en diferentes etapas de la jerarquía, e implementar el feedforward, el lateral y el feedback conjuntamente cuando vayamos a modelar el procesamiento de la información visual. Por otro lado, también necesitamos optimizar el control de la ganancia de normalización en el modelo, y uno de los componentes más importantes del modelo debería ser la normalización divisoria. Hay demasiados parámetros flexibles que necesitan ser optimizados para optimizar al máximo los comportamientos psicofísicos.

- No debemos ignorar las diferencias entre las redes neuronales profundas y la visión biológica porque pueden ayudarnos a optimizar las arquitecturas de las redes neuronales profundas y hacerlas más similares funcionalmente al cerebro. Por último, deberíamos comprobar la capacidad de generalidad de las redes neuronales profundas con algunas tareas visuales de bajo nivel cuando las comparamos con la visión humana, y no quedarnos sólo con algunas tareas visuales de alto nivel, por ejemplo, la clasificación, la segmentación, etc. Aunque la arquitectura de las CNN puede utilizarse para simular sistemas visuales humanos de alto nivel, se desconoce si también puede utilizarse para modelar funciones visuales de bajo nivel, como la percepción del color, la orientación, la reducción del brillo y la adaptación. Por último, pero no por ello menos importante, la fiabilidad de la predicción de la saliencia puede verificarse añadiendo un mayor deterioro de la variedad a imágenes naturales previamente limpias, así como a imágenes generadas mediante métodos psicofísicos. Como se ha dicho anteriormente, necesitamos datos de predicción de la saliencia humana en imágenes deterioradas para que sirvan de referencia con la que se pueda comprobar la precisión del modelo. Esto puede hacerse comparando las predicciones del modelo con las predicciones humanas. Y continuar en esa dirección sería muy emocionante para los próximos años. En conclusión, al investigar las propiedades de las redes neuronales artificiales, es importante tener en cuenta las técnicas psicofísicas. Los numerosos experimentos psicofísicos sobre los mecanismos de la atención visual que han llevado a cabo los estudiosos del campo de la neurociencia visual han sido de gran utilidad para la disciplina en su conjunto. A pesar de que esta área tiene el potencial de mejorar la robustez de las redes neuronales artificiales, hay una escasez de investigación que se lleva a cabo en ella. Incorporar los resultados de las investigaciones psicofísicas a

una red neuronal artificial sería fascinante, pero lo sería aún más si ello condujera a un mayor conocimiento tanto del sistema visual humano como del funcionamiento interno de la red. Esto sería así porque revelaría cómo funciona la propia red. En conclusión, hemos propuesto que los subcampos de la neurociencia visual y las redes neuronales artificiales podrían obtener algún beneficio de la investigación de las CNN a través de métodos psicofísicos.

- Por último, debemos tratar de estimar un nuevo estimador de correlación total y mejorar la precisión del rendimiento de la correlación total. necesitamos construir una nueva estimación para la correlación total porque los dos estimadores que utilizamos en nuestra investigación anterior tenían algunos límites. Esta es la razón por la que creemos que es necesario hacerlo. Para que podamos estimar con facilidad y precisión la correlación total y la correlación total dual, es importante que diseñemos un nuevo estimador de costes eficiente y que consuma menos recursos informáticos. Una vez que hayamos desarrollado métodos para medir la redundancia y la sinergia, podremos cuantificar la representación de la información en las redes neuronales artificiales, así como en las redes neuronales biológicas, utilizando las métricas mencionadas. Las similitudes y contrastes entre las redes neuronales artificiales y las redes neuronales biológicas son un problema importante para desentrañar el misterio que se esconde tras el cerebro humano y las redes neuronales artificiales, como hemos indicado en la sección anterior. En este estudio sólo se investigaron las arquitecturas AlexNet y VGG16 de redes neuronales profundas; sin embargo, es posible que también se utilicen otras arquitecturas, como las redes neuronales recurrentes. Además, los modelos actuales tienen los pesos congelados y luego extraen las características de cada capa. Esto puede causar cierto sesgo cuando comparamos las respuestas de las redes neuronales artificiales y el cerebro. Como resultado, por un lado, podría ser necesario volver a entrenar las redes neuronales utilizando un conjunto de datos que contenga varias modalidades. Por otro lado, necesitamos medir la representación de la información durante las fases de entrenamiento y validación, y esta será una tarea más fascinante que trabajar con redes neuronales estáticas. Por último, pero no por ello menos importante, nos resultará muy interesante investigar la representación de las estadísticas espaciales y la información del color en cada capa de las redes neuronales. Todos estos aspectos son muy importantes para comprender las funciones visuales humanas de bajo y alto nivel, así como las de las redes neuronales artificiales.

Agradecimiento

Las actividades de investigación que han dado lugar a esta tesis han contado con el apoyo de estas becas españolas/europeas de la GVA/AEI/FEDER/UE: MICINN PID2020-118071GB-I00, MICINN PDC2021-121522-C21, and GVA Grisolía-P/2019/035.

Bibliography

[1] Q. Li, "Functional connectivity inference from fmri data using multivariate information measures," *Neural Networks*, vol. 146, pp. 85–97, 2022, ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2021.11.016.

[2] Q. Li, G. V. Steeg, S. Yu, and J. Malo, "Functional connectome of the human brain with total correlation," *Entropy*, vol. 24, no. 12, 2022. DOI: 10.3390/e24121725.

[3] Q. Li, G. Ver Steeg, and J. Malo, "Functional connectivity in visual areas from total correlation," *ArXiV https://arxiv.org/abs/2208.05770*, Aug. 2022.

[4] Q. Li, A. Gomez-Villa, M. Bertalmío, and J. Malo, "Contrast sensitivity functions in autoencoders," *Journal of Vision*, vol. 22, no. 6, pp. 8–8, May 2022.

[5] A. Stockman and L. T. Sharpe, "The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype," *Vision Research*, vol. 40, no. 13, pp. 1711–1737, 2000, ISSN: 0042-6989.

[6] A. Kitaoka, "Illusion and color perception," *J. Color Sci. Assoc. Japan*, vol. 29, pp. 150–151, 2005.

[7] R. Shapley and M. J. Hawken, "Color in the cortex: Single- and double-opponent cells," *Vision Research*, vol. 51, no. 7, pp. 701–717, 2011, Vision Research 50th Anniversary Issue: Part 1.

[8] A. Flachot and K. R. Gegenfurtner, "Color for object recognition: Hue and chroma sensitivity in the deep features of convolutional neural networks," *Vision Research*, vol. 182, pp. 89–100, 2021, ISSN: 0042-6989.

[9] D. Foster, I. Marin-Franch, and S. Nascimento, "Coding efficiency of cie color spaces," in *Proc. 16th Color Imag. Conf.*, Soc. Imag. Sci. Tech., 2008, pp. 285–288.

[10] I. Marin-Franch and D. Foster, "Number of perceptually distinct surface colors in natural scenes," *Journal of Vision*, vol. 10, no. 9, pp. 9–9, Sep. 2010.

[11] A. Flachot, A. Akbarinia, H. H. Schütt, R. W. Fleming, F. A. Wichmann, and K. R. Gegenfurtner, "Deep neural models for color discrimination and color constancy," *CoRR*, vol. abs/2012.14402, 2020. arXiv: 2012.14402.

[12] J. Vazquez-Corral, C. Párraga, R. Baldrich, and M. Vanrell, "Color constancy algorithms: Psychophysical evaluation on a new dataset," *Journal of Imaging Science and Technology*, vol. 53, no. 3, pp. 31105-1-31105–9, 2009.

[13] D. H. Foster, "Color constancy," *Vision Research*, vol. 51, no. 7, pp. 674–700, 2011, Vision Research 50th Anniversary Issue: Part 1, ISSN: 0042-6989. DOI: https://doi.org/10.1016/j.visres.2010.09.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0042698910004402.

[14] F. Campbell and J. Robson, "Application of Fourier analysis to the visibility of gratings," *Journal of Physiology*, vol. 197, pp. 551–566, 1968.

[15] D. H. Kelly, "Motion and vision. ii. stabilized spatio-temporal threshold surface," *J. Opt. Soc. Am.*, vol. 69, no. 10, pp. 1340–1349, 1979.

[16] K. T. Mullen, "The CSF of human colour vision to red-green and yellow-blue chromatic gratings," *J. Physiol.*, vol. 359, pp. 381–400, 1985.

[17] A. B. Watson, "High frame rates and human vision: A view through the window of visibility," *SMPTE Motion Imaging Journal*, vol. 122, no. 2, pp. 18–32, 2013.

[18] E. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Ann. Rev. Neurosci.*, vol. 24, no. 1, pp. 1193–1216, 2001.

[19] M. U. Gutmann, V. Laparra, A. Hyvärinen, and J. Malo, "Spatio-chromatic adaptation via higher-order canonical correlation analysis of natural images," *PloS ONE*, vol. 9, no. 2, e86481, 2014.

[20] J. Malo, "Redundancy reduction in the human visual system: New formulation and applications to image and video coding," PhD Thesis, School of Physics, Univ. de Valencia, Jul. 1999, ISBN: 9781303906640.

[21] J. Perge, K. Koch, R. Miller, P. Sterling, and V. Balasubramanian, "How the optic nerve allocates space, energy capacity, and information," *J. Neurosci.*, vol. 29, no. 24, pp. 7917–7928, 2009.

[22] B. R. Hunt, "Digital image processing," *Proceedings of the IEEE*, vol. 63, pp. 693–708, 1975.

[23] J. Lindsey, S. A. Ocko, S. Ganguli, and S. Deny, "The effects of neural resource constraints on early visual representations," in *Int. Conf. Learn. Repr., ICLR 19*, 2019.

[24] G. Buchsbaum, A. Gottschalk, and H. B. Barlow, "Trichromacy, opponent colours coding and optimum colour information transmission in the retina," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 220, no. 1218, pp. 89–113, 1983. DOI: 10.1098/rspb.1983.0090.

[25] P. Hancock, R. Baddeley, and L. Smith, "The principal components of natural images," *Network*, vol. 3, pp. 61–70, 1991.

[26] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[27] M. Carandini and D. Heeger, "Normalization as a canonical neural computation," *Nat. Rev. Neurosci.*, vol. 13, no. 1, pp. 51–62, 2012.

[28] S. Lyu, "Divisive Normalization: Justification and Effectiveness as Efficient Coding Transform.," *Nips*, pp. 1–9, 2010, ISSN: 1530-888X. DOI: 10.1162/NECO_a_00197.

[29] M. Burg *et al.*, "Learning divisive normalization in primary visual cortex," *PLoS Comput. Biol.*, vol. 17, no. 6, e1009028, 2021.

[30] V. Laparra, J. Muñoz, and J. Malo, "Divisive normalization image quality metric revisited," *JOSA A*, vol. 27, no. 4, pp. 852–864, 2010.

[31] A. Hepburn, V. Laparra, J. Malo, and R. Santos, "Perceptnet: A human visual system inspired neural network for estimating perceptual distance," in *2020 IEEE Int. Conf. Im. Proc. (ICIP)*, 2020, pp. 121–125.

[32] J. Malo, "Spatio-chromatic information available from different neural layers via gaussianization," *J. Math. Neurosci.*, vol. 10, no. 18, 10.1186/s13408-020-00095–8, 2020.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2012, pp. 1097–1105.

[34] M. Martinez, P. Cyriac, T. Batard, M. Bertalmıo, and J. Malo, "Derivatives and inverse of cascaded linear+nonlinear neural models," *PLoS ONE*, vol. 13, no. 10, doi:10.1371/journal.pone.0201326, 2018.

[35] G. K. Wallace, "The jpeg still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, 1992.

[36] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *Int. Conf. Learn. Repres. (ICLR)*, 2017.

[37] J. Malo and E. Simoncelli, "Nonlinear image representation for efficient perceptual coding," *IEEE Trans.Im.Proc.*, vol. 15, no. 1, pp. 68–80, 2006.

[38] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Trans. Inf. Theory*, vol. 20, pp. 525–536, 1974.

[39] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Img. Proc.*, vol. 15, no. 2, pp. 430–444, Feb. 2006, ISSN: 1057-7149. DOI: 10.1109/TIP.2005.859378.

[40] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[41] A. Pons, J. Malo, J. Artigas, and P. Capilla, "Image quality metric based on multidimensional contrast perception models," *Displays*, vol. 20, no. 2, pp. 93–110, 1999.

[42] A. Berardino, V. Laparra, J. Balle, and E. Simoncelli, "Visibility of eigen-distortions of hierarchical models," *CoSyNe*, Feb. 2017.

[43] F. A. Wichmann *et al.*, "Methods and measurements to compare men against machines," *Electronic Imaging*, 36–45(10), 2017.

[44] C. Firestone, "Performance vs. competence in human–machine comparisons," *Proc. Nat. Acad. Sci.*, vol. 117, no. 43, pp. 26 562–26 571, 2020.

[45] C. M. Funke, J. Borowski, K. Stosio, W. Brendel, T. S. A. Wallis, and M. Bethge, "Five points to check when comparing visual perception in humans and machines," *J. Vision*, vol. 21, no. 3, pp. 16–16, 2021.

[46] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[47] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[48] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. 19th NIPS*, ser. NIPS'06, Canada: MIT Press, 2006, pp. 153–160.

[49] S. Priyanka and Y. Wang, "Fully symmetric convolutional network for effective image denoising," *Applied Sciences*, vol. 9, no. 4, 2019.

[50] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018, pp. 8174–8182.

[51] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Img. Proc.*, vol. 26, no. 7, pp. 3142–3155, 2017.

[52] A. Akbarinia and R. Gil-Rodríguez, "Deciphering image contrast in object classification deep networks," *Vision Research*, vol. 173, pp. 61–76, 2020, ISSN: 0042-6989. DOI: https://doi.org/10.1016/j.visres.2020.04.015.

[53] A. Akbarinia, Y. Morgenstern, and K. R. Gegenfurtner, "Contrast sensitivity is formed by visual experience and task demands," *Journal of Vision*, 2021.

[54] S. Cadena *et al.*, "Deep convolutional models improve predictions of macaque V1 responses to natural images," *PLoS Comput. Biol.*, vol. 15, no. 4, e1006897, 2019.

[55] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "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*, 2019.

[56] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA, USA: MIT Press, 1969.

[57] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, pp. 193–202, 2004.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. Comp. Vis. Patt. Recogn. (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[59] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015.

[60] N. Kriegeskorte, "Deep neural networks: A new framework for modeling biological vision and brain information processing," *Ann. Rev. Vis. Sci.*, vol. 1, no. 1, pp. 417–446, 2015.

[61] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, Feb. 2021. DOI: 10.1109/TPAMI.2021.3059968.

[62] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, Aug. 2019. DOI: 10.1109/TPAMI.2019.2935715.

[63] W. S. Mcculloch and W. H. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, 1943.

[64] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, D. E. Rumelhart and J. L. Mcclelland, Eds., Cambridge, MA: MIT Press, 1986, pp. 318–362.

[65] D. Yamins and J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nat. Neurosci.*, vol. 19, pp. 356–365, 2016.

[66] U. Güçlü and M. A. J. van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," *Journal of Neuroscience*, vol. 35, no. 27, pp. 10 005–10 014, 2015.

[67] C. Cadieu *et al.*, "Deep neural networks rival the representation of primate it cortex for core visual object recognition," *PLoS Comput. Biol.*, vol. 10, no. 12, e1003963, 2014.

[68] B. Richards and e. al., "A deep learning framework for neuroscience," *Nature Neuroscience*, vol. 22, pp. 1761–1770, Oct. 2019.

[69] S. Strong, R. Koberle, R. de Ruyter van Steveninck, and W. Bialek, "Entropy and information in neural spike trains," *Phys. Rev. Lett.*, vol. 80, pp. 197–200, 1998.

[70] A. Dimitrov, A. Lazar, and J. Victor, "Information theory in neuroscience," *J. Comput. Neurosci.*, vol. 30, no. 1, pp. 1–5, 2011.

[71] N. Timme and C. Lapish, "A tutorial for information theory in neuroscience," *eNeuro*, vol. 5, no. 3, ENEURO.0052–18.2018, 2018.

[72] L. Paninski, "Estimation of Entropy and Mutual Information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003, ISSN: 0899-7667. DOI: 10.1162/089976603321780272. eprint: 0402594v3 (arXiv:cond-mat).

[73] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, vol. 69, no. 6, p. 16, 2004, ISSN: 1063651X. DOI: 10.1103/PhysRevE.69.066138. eprint: 0305641 (cond-mat).

[74] B. Chai, D. Walther, D. Beck, and L. Fei-fei, "Exploring functional connectivities of the human brain using multivariate information analysis," in *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds., vol. 22, Curran Associates, Inc., 2009, pp. 270–278.

[75] J. Lizier, J. Heinzle, A. Horstmann, J. Haynes, and M. Prokopenko, "Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity," *J. Comput. Neurosci.*, vol. 30, no. 1, pp. 85–107, 2011.

[76] J. L. Guerrero-Cusumano, "Measures of dependence for the multivariate t distribution with applications to the stock market," *Communications in Statistics - Theory and Methods*, vol. 27, no. 12, pp. 2985–3006, 1998, ISSN: 03610926. DOI: 10.1080/03610929808832268.

[77] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of research and development*, vol. 4, no. 1, pp. 66–82, 1960.

[78] M. Studeny and J. Vejnarova, "The multi-information function as a tool for measuring stochastic dependence," in *Learning in graphical models*, M. I. Jordan, Ed. Kluwer, Jan. 1998, pp. 261–298.

[79] T. M. Cover and J. A. Thomas, *Elements of Information Theory, 2nd Edition*. Wiley, 2006.

[80] V. Laparra, G. Camps-Valls, and J. Malo, "Iterative gaussianization: From ICA to random rotations," *IEEE Trans. Neural Networks*, vol. 22, no. 4, pp. 537–549, 2011.

[81] G. Ver Steeg and A. Galstyan, "Discovering structure in high-dimensional data through correlation explanation," *Advances in Neural Information Processing Systems*, vol. 1, Jun. 2014.

[82] S. Yu, L. G. S. Giraldo, R. Jenssen, and J. C. Principe, "Multivariate extension of matrix-based rényi's α-order entropy functional," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 11, pp. 2960–2966, 2019.

[83] R. F. Betzel and D. S. Bassett, "Multi-scale brain networks," *NeuroImage*, vol. 160, pp. 73–83, 2017, Functional Architecture of the Brain, ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2016.11.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811916306152.

[84] M. Cole, G. Yang, J. Murray, G. Repovs, and A. Anticevic, "Functional connectivity change as shared signal dynamics," *Journal of Neuroscience Methods*, vol. 259, Nov. 2015. DOI: 10.1016/j.jneumeth.2015.11.011.

[85] K. J. Friston, "Functional and effective connectivity: A review," *Brain Connectivity*, vol. 1, 2011. DOI: http://doi.org/10.1089/brain.2011.0008.

[86] J. Hlinka, M. Palus, M. Vejmelka, D. Mantini, and M. Corbetta, "Functional connectivity in resting-state fmri: Is linear correlation sufficient?" *NeuroImage*, vol. 54, pp. 2218–25, Feb. 2011. DOI: 10.1016/j.neuroimage.2010.08.042.

[87] R. Goebel, A. Roebroeck, D. Kim, and E. Formisano, "Investigating directed cortical interactions in time-resolved fmri data using vector autoregressive modeling and granger causality mapping," *Magnetic resonance imaging*, vol. 21, pp. 1251–61, Jan. 2004. DOI: 10.1016/j.mri.2003.08.026.

[88] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, "Transfer entropy—a model-free measure of effective connectivity for the neurosciences," *Journal of computational neuroscience*, vol. 30, pp. 45–67, Feb. 2011. DOI: 10.1007/s10827-010-0262-3.

[89] M. Wibral, R. Vicente, and M. Lindner, "Transfer entropy in neuroscience," *Understanding Complex Systems*, pp. 3–36, Jan. 2014. DOI: 10.1007/978-3-642-54474-3-1.

[90] G. Varoquaux and C. Craddock, "Learning and comparing functional connectomes across subjects," *NeuroImage*, vol. 80, Apr. 2013. DOI: 10.1016/j.neuroimage.2013.04.007.

[91] V. Calhoun, K. Kiehl, J. Turner, E. Allen, and G. Pearlson, "Exploring the psychosis functional connectome: Aberrant intrinsic networks in schizophrenia and bipolar disorder," *Frontiers in psychiatry / Frontiers Research Foundation*, vol. 2, p. 75, Nov. 2011. DOI: 10.3389/fpsyt.2011.00075.

[92] M. Fox and M. Greicius, "Clinical applications of resting state functional connectivity," *Frontiers in systems neuroscience*, vol. 4, p. 19, Jun. 2010. DOI: 10.3389/fnsys.2010.00019.

[93] C. E. Shannon, "A mathematical theory of communication," *Key Papers in the Development of Information Theory*, 1948. [Online]. Available: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

[94] C. Cutts and S. Eglen, "Detecting pairwise correlations in spike trains: An objective comparison of methods and application to the study of retinal waves," *Journal of Neuroscience*, vol. 34, p. 14 288, Oct. 2014. DOI: 10.1523/JNEUROSCI.2767-14.2014.

[95] R. Ince, F. Montani, E. Arabzadeh, M. Diamond, and S. Panzeri, "On the presence of high-order interactions among somatosensory neurons and their effect on information transmission," *Journal of Physics: Conference Series*, vol. 197, p. 012 013, Dec. 2009. DOI: 10.1088/1742-6596/197/1/012013.

[96] W. Li, "Mutual information functions versus correlation functions," *Journal of Statistical Physics*, vol. 60, pp. 823–837, Sep. 1990. DOI: 10.1007/BF01025996.

[97] Y.-I. Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Physical review. E, Statistical physics, plasmas, fluids, and related interdisciplinary topics*, vol. 52, pp. 2318–2321, Oct. 1995. DOI: 10.1103/PhysRevE.52.2318.

[98] H. Singh, V. Hnizdo, A. Demchuk, and N. Misra, "Nearest neighbor estimates of entropy," *American Journal of Mathematical and Management Sciences*, vol. 23, Feb. 2003. DOI: 10.1080/01966324.2003.10737616.

[99] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 69, p. 066 138, Jul. 2004. DOI: 10.1103/PhysRevE.69.066138.

[100] R. Ince, B. Giordano, C. Kayser, G. Rousselet, J. Gross, and P. Schyns, "A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula," *Human brain mapping*, vol. 38, Nov. 2016. DOI: 10.1002/hbm.23471.

[101] J. Ma and Z. Sun, "Mutual information is copula entropy," *Tsinghua Science Technology*, vol. 16, pp. 51–54, Feb. 2011. DOI: 10.1016/S1007-0214(11)70008-6.

[102] E. Eqlimi, N. riyahi alam, M. Sahraian, A. Eshaghi, and H. Rad, "Mutual information weighted graphs for resting state functional connectivity in fmri data," in *Joint Annual Meeting ISMRM-ESMRMB, Milan, Italy*, Jan. 2014.

[103] S. Gao, R. Brekelmans, G. V. Steeg, and A. Galstyan, "Auto-encoding total correlation explanation," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, K. Chaudhuri and M. Sugiyama, Eds., ser. Proceedings of Machine Learning Research, vol. 89, PMLR, 16–18 Apr 2019, pp. 1157–1166. [Online]. Available: http://proceedings.mlr.press/v89/gao19a.html.

[104] G. V. Steeg, "Unsupervised learning via total correlation explanation," in *IJCAI*, 2017.

[105] C. Chan, A. Al-Bashabsheh, and Q. Zhou, "Agglomerative info-clustering: Maximizing normalized total correlation," *IEEE Transactions on Information Theory*, vol. PP, pp. 1–1, Nov. 2020. DOI: 10.1109/TIT.2020.3040492.

[106] Z. Walczak, "Total correlations and mutual information," *Physics Letters, Section A: General, Atomic and Solid State Physics*, vol. 373, Jun. 2008. DOI: 10.1016/j.physleta.2009.03.047.

[107] T. Ferenci and L. Kovács, "Using total correlation to discover related clusters of clinical chemistry parameters," *SISY 2014 - IEEE 12th International Symposium on Intelligent Systems and Informatics, Proceedings*, pp. 49–54, Oct. 2014. DOI: 10.1109/SISY.2014.6923614.

[108] A. Gomez-Villa, M. Bertalmío, and J. Malo, "Visual information flow in wilson-cowan networks," *Journal of Neurophysiology*, vol. 123, Mar. 2020. DOI: 10.1152/jn.00487.2019.

[109] M. Gatica *et al.*, "High-order interdependencies in the aging brain," *Brain Connectivity*, vol. 11, no. 9, pp. 734–744, 2021. DOI: 10.1089/brain.2020.0982.

[110] R. Herzog *et al.*, "Genuine high-order interactions in brain networks and neurodegeneration," *Neurobiology of Disease*, vol. 175, p. 105 918, 2022, ISSN: 0969-9961. DOI: https://doi.org/10.1016/j.nbd.2022.105918.

[111] A. Hepburn, V. Laparra, R. Santos, J. Ballé, and J. Malo, "On the relation between statistical learning and perceptual distances," *CoRR*, vol. abs/2106.04427, 2021. arXiv: 2106.04427.

[112] A. Gomez-Villa, A. Martin, J. Vazquez, and M. Bertalmio, "Convolutional neural networks can be deceived by visual illusions," in *Proc. IEEE Comp. Vis. Patt. Recogn. (CVPR 19)*, 2019, pp. 12 309–12 317.

[113] A. Gomez-Villa, A. Martin, J. Vazquez, M. Bertalmío, and J. Malo, "Color illusions also deceive CNNs for low-level vision tasks: Analysis and implications," *Vision Research*, vol. 176, pp. 156–174, 11 Nov. 2020.

[114] L. Martinez-Otero, M. Molano, X. Wang, F. Sommer, and J. Hirsch, "Statistical wiring of thalamic receptive fields optimizes spatial sampling of the retinal image," *Neuron*, vol. 81, no. 4, pp. 943–956, 2014.

[115] J. J. Atick, "Could information theory provide an ecological theory of sensory processing?" *Network: Computation in Neural Systems*, vol. 22, pp. 4–44, 2011.

[116] J. Atick and A. Redlich, "What does the retina know about natural scenes?" *Neural Comp.*, vol. 4, no. 2, pp. 196–210, 1992.

[117] J. Atick, Z. Li, and A. Redlich, "Understanding retinal color coding from first principles," *Neural Comp.*, vol. 4, no. 4, pp. 559–572, 1992.

[118] D. Brainard and B. Wandell, *Isetbio: Tools for modeling the human visual system front end*, Imageval Consulting, Nov. 2020.

[119] H. Barlow, "Possible principles underlying the transformations of sensory messages," *Sensory Communication*, vol. 1, Jan. 1961. DOI: 10.7551/mitpress/9780262518420.003.0013.

[120] H. Barlow, "Vision: Coding and efficiency," in C. Blakemore, Ed. UK: Cambridge Univ. Press, 1990, ch. A theory about the functional role and synaptic mechanism of visual aftereffects.

[121] H. Barlow, "Redundancy reduction revisited," *Network: computation in neural systems*, vol. 12, no. 3, pp. 241–253, 2001.

[122] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, pp. 2660–2673, Nov. 2017. DOI: 10.1109/TNNLS.2016.2599820.

[123] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Springer International Publishing, 2014, pp. 818–833.

[124] A. M. Saxe *et al.*, "On the information bottleneck theory of deep learning," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, p. 124 020, Dec. 2019. DOI: 10.1088/1742-5468/ab3985.

[125] A. Painsky and N. Tishby, "Gaussian lower bound for the information bottleneck limit," *Journal of Machine Learning Research*, vol. 18, no. 213, pp. 1–29, 2018.

[126] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 1999, pp. 368–377. [Online]. Available: https://arxiv.org/abs/physics/0004057.

[127] S. Yu and J. C. Príncipe, "Understanding autoencoders with information theoretic concepts," *Neural Networks*, vol. 117, pp. 104–123, 2019. DOI: https://doi.org/10.1016/j.neunet.2019.05.003.

[128] S. Yu, K. Wickstrom, R. Jenssen, and J. Principe, "Understanding convolutional neural networks with information theory: An initial exploration," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 435–442, Jan. 2021, ISSN: 2162-237X. DOI: 10.1109/tnnls.2020.2968509.

[129] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *ArXiv*, vol. abs/1703.00810, 2017.

[130] V. Laparra, J. E. Johnson, G. Camps-Valls, R. Santos-Rodríguez, and J. Malo, "Information theory measures via multidimensional gaussianization," *ArXiv*, vol. abs/2010.03807, 2020.

[131] R. M. Cichy *et al.*, "The algonauts project 2021 challenge: How the human brain makes sense of a world in motion," *CoRR*, vol. abs/2104.13714, 2021. arXiv: 2104.13714. [Online]. Available: https://arxiv.org/abs/2104.13714.

[132] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[133] A. Mordvintsev, C. Olah, and M. Tyka, *Inceptionism: Going deeper into neural networks*, 2015. [Online]. Available: `https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html`.

[134] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9. DOI: `10.1109/CVPR.2015.7298594`.

[135] D. L. Yamins, H. Hong, C. Cadieu, and J. J. DiCarlo, "Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26, Curran Associates, Inc., 2013.

[136] C. Huang and D. Kruger, *Proc. icml workshop on invertible neural nets and normalizing flows*, Int. Conf. Mach. Learn. `https://invertibleworkshop.github.io/INNF_2019/accepted_papers/`, Jun. 2019.

[137] D. I. Inouye and P. Ravikumar, "Deep density destructors," in *ICML*, 2018.

[138] J. Johnson, V. Laparra, R. Santos, G. Camps, and J. Malo, "Information theory in density destructors," in *7th Int. Conf. Mach. Learn., ICML 2019, Workshop on Invertible Normalization Flows*, 2019.

Part II

Appendix with Scientific Publications

Paper I
Qiang Li, Functional connectivity inference from fMRI data using multivariate information measures, Neural Networks, Volume 146, 2022, Pages 85-97. doi:10.1016/j.neunet.2021.11.016.

Paper II
Qiang Li, Greg Ver Steeg, Shujian Yu and Jesús Malo. "Functional Connectome of the Human Brain with Total Correlation" Entropy 24, no. 12: 1725, 2022.

Paper III
Qiang Li, Greg Ver Steeg, and Jesús Malo. "Functional Connectivity in Visual Areas from Total Correlation." arXiv preprint arXiv:2208.05770 (2022). [Under Review of Neural Networks Journal]

Paper IV
Qiang Li, Alex Gomez-Villa, Marcelo Bertalmío, Jesús Malo; Contrast sensitivity functions in autoencoders. Journal of Vision 2022;22(6):8. doi: https://doi.org/10.1167/jov.22.6.8.

.

Paper I

Neural Networks 146 (2022) 85–97

Contents lists available at ScienceDirect

Neural Networks

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

Functional connectivity inference from fMRI data using multivariate information measures

Qiang Li

Image Processing Laboratory, Parc Cientific, University of Valencia, Valencia, Spain

ARTICLE INFO

Article history:
Received 9 July 2021
Received in revised form 1 October 2021
Accepted 11 November 2021
Available online 17 November 2021

Keywords:
Information transmission
Mutual information
Interaction information
Total correlation
Multivariate distribution
Functional connectivity

ABSTRACT

Shannon's entropy or an extension of Shannon's entropy can be used to quantify information transmission between or among variables. Mutual information is the pair-wise information that captures nonlinear relationships between variables. It is more robust than linear correlation methods. Beyond mutual information, two generalizations are defined for multivariate distributions: interaction information or co-information and total correlation or multi-mutual information. In comparison to mutual information, interaction information and total correlation are underutilized and poorly studied in applied neuroscience research. Quantifying information flow between brain regions is not explicitly explained in neuroscience by interaction information and total correlation. This article aims to clarify the distinctions between the neuroscience concepts of mutual information, interaction information, and total correlation. Additionally, we proposed a novel method for determining the interaction information between three variables using total correlation and conditional mutual information. On the other hand, how to apply it properly in practical situations. We supplied both simulation experiments and real neural studies to estimate functional connectivity in the brain with the above three higher-order information-theoretic approaches. In order to capture redundancy information for multivariate variables, we discovered that interaction information and total correlation were both robust, and it could be able to capture both well-known and yet-to-be-discovered functional brain connections.

1. Introduction

Shannon launched information theory in 1948 (Shannon, 1948), and it is firstly proposed for solving information transmission and compaction in the communicate noise channel between source and receiver. Shannon's entropy quantities uncertainty of system and has been widely used in many branches of science. Understanding information flowing or coupled among neurons in the micro-level or macro-level is always a critical research problem. Entropy can be used to quantify information representation and capacity in the brain. There is a long history of applying entropy to interpret brain function (Garcia, Fernández-Caballero, & Martinez Rodrigo, 2021; Keshmiri, 2020; Saxe, Calderone, & Morales, 2018; Timme & Lapish, 2018). On the one hand, entropy has been widely used to analyze neural data and interpret information capacity in the brain. The *Principle of maximum entropy* is one of the classical theories in the statistics and neuroscience (Jaynes, 1957; Rosenkrantz, 1989). On the other hand, entropy-based brain models or hypotheses also became one of the critical topics in neuroscience. The *Entropic Brain Hypothesis* is one of the classic representatives to model brain mechanisms that posits a relationship between total entropy in the brain and long-term psychological states based on recent brain imaging research (Carhart-Harris et al., 2014). Both aforementioned theories have been explained specific neuroscience problems. However, cognitive function is derived by population neural activity or interaction with other neurons in the brain rather than single neuron. Therefore, exploring information interaction between brain regions and how the neuron activity coupled to drive cognitive behaviors is necessary to understand brain function. One of the common information-theoretic tools is mutual information which mainly statistically described two random variables dependencies relationship (Cover & Thomas, 1991; Shannon, 1948). The mutual information has been used to study neural activity (Cutts & Eglen, 2014; Mijatović et al., 2021), characterize different neurons relationship (Ince, Montani, Arabzadeh, Diamond, & Panzeri, 2009), and apply to neuroimaging (Functional Magnetic Resonance Imaging (fMRI) is an imaging technique applied to study human brain function and neurological diseases) studies for a long time (Bandettini, 2012; Gao et al., 2021; Li, 1990; Stozer et al., 2013; Tedeschi et al., 2003). There are different practical approaches for estimating mutual information from observations. The standard estimators are binned method, continuous methods, e.g., Kernel Density Estimation (KDE) (Moon, Rajagopalan, & Lall, 1995), naive non-parametric *k*-nearest neighbor (KNN)

E-mail address: qiang.li@uv.es.

https://doi.org/10.1016/j.neunet.2021.11.016
0893-6080/

(Singh, Hnizdo, Demchuk, & Misra, 2003) and other variants KNN estimators (Kraskov, Stögbauer, & Grassberger, 2004), Gaussian copula-based estimators (Ince et al., 2016; Ma & Sun, 2011) and so on. Mutual information already became a popular way to estimate pair-wise functional connectivity then based on it to understand information interaction across brain regions (Eqlimi, riyahi alam, Sahraian, Eshaghi, & Rad, 2014; Hlinka, Palus, Vejmelka, Mantini, & Corbetta, 2011; Ince et al., 2016). However, we faced two challenges when we used above mentioned mutual information estimators. First, in a typical scenario, we estimate that mutual information from a limited number of samples is often infeasible in many practical situations. Second, like the amount and variety of big data growth, mutual information is not performing very well for identifying or picking strong relationships from dependence data, but in some cases, we are interested in understanding these strong dependences rather than independence (Gao, Ver Steeg, & Galstyan, 2015). Last but not least, we want to investigate the relationship of neurons beyond pair-wise, but general mutual information cannot solve the above problems.

Considered aforementioned general mutual information limitations, two generalizations of mutual information theory proposed for solving multivariate variables relationship and they are interaction information (McGill, 1954) (also called co-information Bell, 2003) and total correlation (Watanabe, 1960) (also noted multi-mutual information Studený & Vejnarová, 1999), respectively. The mathematics expression of interaction information and total correlation will be introduced in Section 2. Therefore, we only describe both applications in the neuroscience field here. First, interaction information quantifies the amount of information (redundancy or synergy) shared among variables, and it can be used to measure redundancy information in the set of neurons. However, there are fewer studies in empirical neuroscience done with interaction information. One reason is hard interpretation from both information theory and neuroscience perspectives because it can be positive or negative. Second, it will be hard to estimate from real neural dataset. However, since not many studies were done in neuroscience with interaction information, but it still has been used in the feature selection and machine learning fields (Jakulin & Bratko, 2004; Jakulin, Bratko, Smrke, Demsar, & Zupan, 2003). Furthermore, understanding the role of redundancy information is a significant problem, and it can help us further understanding brain function. As we mentioned above, another generalization of mutual information is total correlation. Total correlation can be used to measured redundancy information within a set of variables. The difference of total correlation and interaction information can be directly visualized with Venn diagrams in Fig. 1. The total correlation-related studies are mostly done in the image statistics and machine learning fields (Chan, Al-Bashabsheh, & Zhou, 2020; Gao, Brekelmans, Steeg, & Galstyan, 2019; Ver Steeg, 2017; Walczak, 2008) instead of neuroscience. However, it gradually gets attention in understanding information dependence of multivariate variables in biological studies (Ferenci & Kovács, 2014; Gomez-Villa, Bertalmío, & Malo, 2020; Li, 2021; Li, Johnson, Esteve-Taboada, Laparra, & Malo, 2021).

I will quickly review some clustering algorithms used for clustering information or statistics relationships among multivariate variables in the following contents. Clustering analysis is a good way for us to group similar variables into a set (Saxena et al., 2017). There has varieties type of clustering algorithms for solving different practical problems. I will only review some standard clustering models which are usually used in brain neuroimaging researches. The first one is agglomerative hierarchical clustering which is based on variables distance, then merges nearest clusters from beginning in which variables are considered a single cluster then until group all similarity objects into one group (Everitt, Landau, Leese, & Stahl, 2011; Fischer, 1995). The way to visualize the clustering is called dendrogram, and it already has been widely used in biostatistics (Everitt et al., 2011; Nowak & Tibshirani, 2007). In brain imaging, it also can be used for measure of functional connectivity in the brain Akiki and Abdallah (2019) and Wang, Msghina, and Li (2016). However, hierarchical clustering always uses only pair-wise distances, hence, it is unable to utilize interaction information and total correlation's possibility to describe to correlation of $n > 2$ variables. To address the above problem, one way in which we can replace cluster criterion with interaction information or total correlation. This may be called greedy clustering as it retains the greedy nature of hierarchical clustering (Cormen, Leiserson, Rivest, & Stein, 2001). In another way, we can solve it with graph theory-based models. In graph theory and complex networks, the network structure is encoded in the edges between nodes, and weights represent the strength of the connection between nodes. The graph can supply between or beyond pair-wise relationships for variables, and it has been a growing interest in brain neuroimaging studies over the recent years (Mansoory, Oghabian, Jafari, & Shahbabaie, 2017; Uehara, Tobimatsu, Kan, & Miyauchi, 2012; Wang, Zuo, & He, 2010). The graph-theoretic analyses play a significant role in exploring the brain's intrinsic, task-related activity and help us to describe and predict dysfunction using a network perspective (Farahani, Karwowski, & Lighthall, 2019; Uehara et al., 2012).

This paper explored mutual information, interaction information, and total correlation information-theoretic methods from a neuroimaging perspective. Section 2, we review the definition of entropy, mutual information, interaction information, and total correlation and their application mainly in neuroimaging. Moreover, we present a novel interaction information estimator through total correlation and condition mutual information. Section 3, we supplied simulation studies and explored redundancy with mutual information, interaction information, and total correlation. Section 4 then showed some applications in brain development. Section 5 gives a general discussion and concludes with some directions for future research.

In this study, our main contributions are threefold,

- Considered mutual information can only capture pairwise dependency (n = 2) as opposed to beyond two variables (n>2). We estimate statistics associated among variables (n>2) via interaction information and total correlation. We proposed new estimators for interaction information based on total correlation and conditional mutual information based on previous works.
- In computational neuroimaging analysis, such as Functional Connectivity (FC), one of the immense challenges is high-dimensional fMRI data. We showed that interaction information and total correlation could capture more rich dependency than mutual information. Meanwhile, it also opens a door for our study in image statistics, deep learning tasks, such as disentangled representation learning, ensemble learning, and model distillation.
- As we stated before, it is hard to structure statistics relationships among variables. Therefore, we used greedy clustering and graph theory to represent functional connectivity among (Region of Interests) ROIs with resting-state and task-related fMRI data. We proved that high-order information-theoretical could capture both some known functional connectivity and some unknown functional connectivity in the brain.

2. Theory and method

This section will first introduce the definition of entropy, mutual information, and multivariate mutual information under different cases. As we know, mutual information cannot satisfy

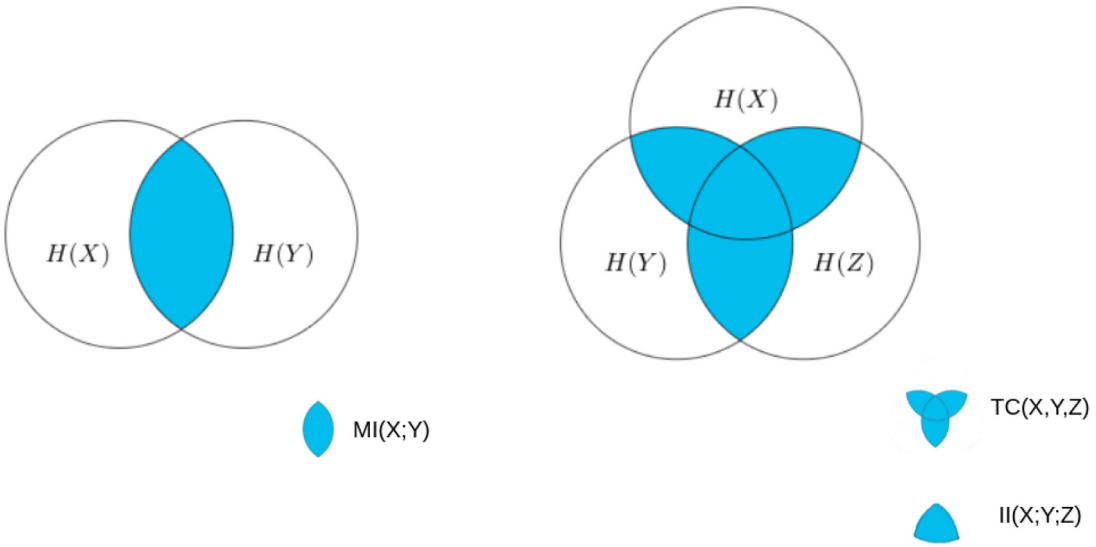


Fig. 1. Venn diagrams illustrating mutual information, interaction information, and total correlation. (a) Information content of 2 variables and their mutual information, MI(X; Y). (b) Information content of 3 variables and their common information, which are interaction information, II(X; Y; Z) and total correlation, TC(X, Y, Z). The $H(X)$, $H(Y)$, and $H(Z)$ indicate entropy of X, Y and Z, respectively.

needs in specific conditions (N>2). Therefore, we will introduce interaction information and total correlation and differences between them in the following contents. (Fig. 1).

2.1. Entropy

Entropy $H(X)$ indicates how surprising it is, on average, to get a symbol x from a random variable X that can take the possible symbols $x_1, x_2, \ldots, x_n$ each with probability $p(x_i)$:

$$H(X) = \mathbb{E}[-\log p(X)] = -\sum_{i=1}^{k} p(x_i) \log p(x_i)$$

entropy can be analogously defined for multivariable case (Cover & Thomas, 1991), i.e., for a n-dimensional random vector X, called joint entropy:

$$H(X_1, X_2, \ldots, X_n) = \mathbb{E}[-\log p(X_1, X_2, \ldots, X_n)]$$
$$= -\sum_{i_1=1}^{k_1} \cdots \sum_{i_p=1}^{k_p} p\left(x_{i_1}, \ldots, x_{i_p}\right) \log p\left(x_{i_1}, \ldots, x_{i_n}\right)$$

The corresponding units of entropy are bits for the base is 2. The entropy of Gaussian variables depends on the variables dimensional d and determinant of the covariance matrix Σ. The entropy of Gaussian variables can be defined as,

$$H(X) = \frac{1}{2\ln 2} \ln\left[(2\pi e)^d |\Sigma|\right] \tag{1}$$

The above-limited approach to estimate entropy exists bias because limited data used, and this effect can be avoided via analytic correction (Goodman, 1963; Misra, Singh, & Demchuk, 2005). The bias-corrected entropy estimate is given by,

$$H(X) = \frac{1}{2\ln 2}\left(\ln\left[(2\pi e)^d |\Sigma|\right] - d\ln\frac{2}{N-1} - \sum_{i=1}^{d} \Psi\left(\frac{N-i}{2}\right)\right)$$

where N is the total samples and d is the dimentionality of X with covariance matrix Σ.

Entropy measured the stochastic system uncertain state. Shannon's entropy is widely used in the communication system and quantifies lossless information compression and transmission. The entropy concept is also already widely applied across various research fields. In this research, we are interested in explaining neural information flow among brain regions and how each region functional coupled with each other.

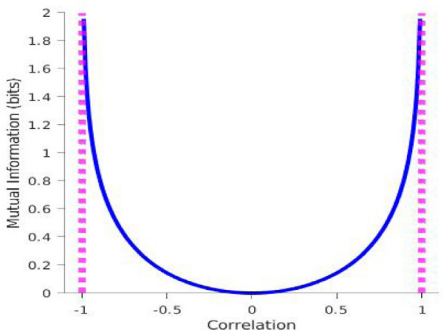


Fig. 2. The relationship between correlation and mutual information. The linear correlation could be negative but mutual information is always positive.

2.2. Mutual information (2-variate case)

The mutual information of pair-wise variables mainly described amount of information that one variable contained about another random variables (Cover & Thomas, 1991; Shannon, 1948). It quantifies how much information is shared between variables. The amount of information that is contained in two variables X, Y is given by the joint entropy,

$$I(X; Y) = \sum_{x\in X}\sum_{y\in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

If $\mathbf{x} = [X_1, X_2]$ satisfies a bivariate Gaussian distribution $x \sim \mathcal{N}(0,\ 1)$, then mutual information between X_1 and X_2 is given by,

$$I(X_1; X_2) = -\frac{1}{2}\log\left((1 - \rho^2_{X_1X_2})\right) \tag{2}$$

where ρ is the correlation coefficient between X_1 and X_2. Therefore, mutual information can be thought of as a nonlinear function of correlation (see Fig. 2). Compared to correlation, mutual information captures not only nonlinear dependencies but also handles discrete random variables. From Eq. (2). For the multivariate Gaussian case, the mutual information between X and Y is given by,

$$I(X; Y) = \frac{1}{2\ln 2} \ln\left[\frac{|\Sigma_X|\,|\Sigma_Y|}{|\Sigma_{XY}|}\right] \tag{3}$$

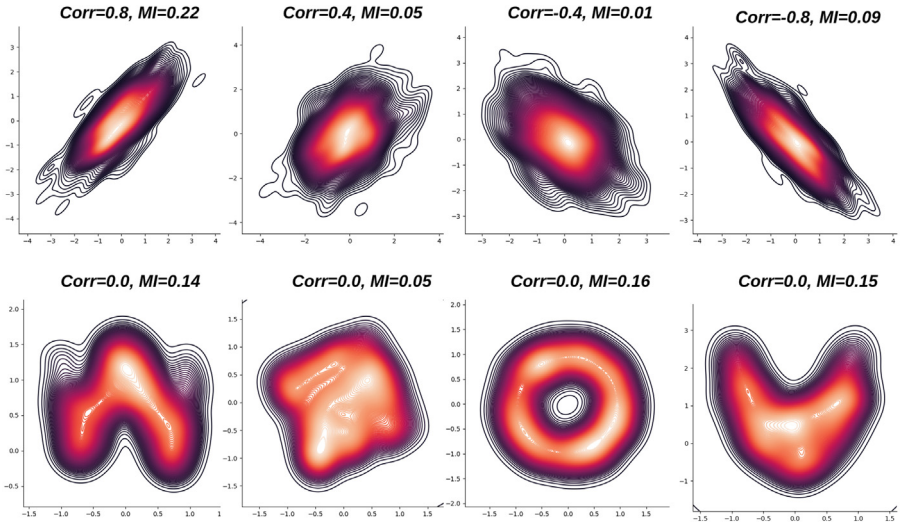


Fig. 3. Estimated dependencies on linear and nonlinear variables with correlation and mutual information. The correlation between the two variables is shown on the left, and the MI is shown on the right side; (discrete method, 16 bins, no bias correction). The top row shows linear dependencies, for which MI and correlation both detect a relationship. The distributions in the bottom row have no correlation because they are nonlinear.

where Σ_X and Σ_Y are the covariance matrices of variable X and Y, respectively. Σ_{XY} is the covariance matrix for joint variable (X, Y). Eq. (3) can be further expressed (Scharf & Mullis, 2000) in terms of correlation coefficients ρ_i as follows,

$$\mathrm{MI}(X;Y) = -\frac{1}{2}\sum_i \ln\left(1-\rho_i^2\right) \tag{4}$$

Comparison with linear correlation, mutual information capture not only pair-wise linear relationships but also nonlinear dependencies. Following the simulation experiments in Ince et al. (2016), we simulated varying bivariate distribution with 10^4 samples, and measured their dependency with linear correlation and mutual information. In Fig. 3, The correlation and mutual information both capture dependencies in the bivariate distribution variables. However, mutual information still captures information from nonlinear distribution signal, but correlation of all nonlinear signal get zeros. Therefore, as we mentioned above, mutual information only estimates pair-wise relationships in the system. However, we usually faced more than two variables in real situations. e.g., neurons in the brain. Multivariate mutual information is an extension of mutual information that can estimate relationship beyond two variables.

2.3. Multivariate mutual information (N-variate case, $N > 2$)

Two high-order information-theoretic generalizations of mutual information are presented in this section. They are interaction information and total correlation, respectively. The difference among mutual information, interaction information and total correlation were illustrated in Fig. 1. The detail of the concept about interaction information and total correlation will be introduced in the following sections.

2.3.1. Interaction information or co-information (3-variate case)

Interaction information (McGill, 1954), also named co-information (Bell, 2003), is used to describe the information shared among three variables. The definition of interaction information relies on conditional mutual information that mutual information of two random variables (X and Y) condition on a third one (Z) (Wyner, 1978). It can be expressed as,

$$II(X;Y\mid Z) = \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)\log\frac{p(x,y\mid z)}{p(x\mid z)p(y\mid z)} \tag{5}$$

it can be negative compared to mutual information, and Eq. (5) could be rewritten as,

$$II(X;Y\mid Z) = \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)\log\frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} \tag{6}$$

In the 3-variate case, the interaction information is defined by conditional mutual information among three variables subtracted standard mutual information between variables.

$$\begin{aligned} II(X;Y;Z) &= I(X;Y\mid Z) - I(X;Y) \\ &= I(X;Z\mid Y) - I(X;Z) \\ &= I(Y;Z\mid X) - I(Y;Z) \end{aligned} \tag{7}$$

According to Eqs. (2) and (6), Eq. (7) can be rewritten as,

$$\begin{aligned} II(X;Y;Z) = &\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(x,y,z)\log\frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} \\ &-\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \end{aligned} \tag{8}$$

Therefore, from Eq. (8), we can measure interaction information through estimate variables of probability and joint probability.

Estimating interaction information for Gaussian variables We estimate interaction information for Gaussian variables through total correlation and conditional mutual information in the simulation study. From Fig. 1, it can be mathematically expressed as,

$$II(X;Y;Z) = TC(X,Y,Z) - (I(X;Y|Z) + I(X;Z|Y) + I(Y;Z|X)) \tag{9}$$

The conditional mutual information between two Gaussian variables conditioned on a third, $I(X;Y|Z)$, $I(X;Z|Y)$, and $I(Y;Z|X)$ can be measured with function *cmi_ggg*[1] which supplied by Ince (Ince et al., 2016). Total correlation among X, Y, Z can be estimated with Rotation-Based Iterative Gaussianization (RBIG) (Laparra, Camps-Valls, & Malo, 2011; Laparra, Johnson, Camps-Valls, Santos-Rodríguez, & Malo, 2020) that introduced in the later section. Now from Eq. (9), we can estimate interaction information for three variables via total correlation and conditional mutual information.

[1] https://github.com/robince/gcmi/blob/master/matlab/cmi_ggg.m.

2.3.2. Total correlation or multi-mutual information (N-variate case ($N \geq 3$))

Total correlation is one of several generalizations of mutual information, and the multivariate constraint, or multi-information constraint, is another name for it. Total correlation firstly proposed by Wantanable in 1996 (Watanabe, 1960), and also noted as multi-mutual information by Studený (Studený & Vejnarová, 1999). The concept of total correlation, $TC(X_1, X_2, \ldots, X_N)$, provides a direct and effective way of assessing the dependency among multiple variables (N > 3): It determines how redundant or dependent a set of n random variables is.

$$TC\left(X_1, X_2, \ldots, X_N\right) = \sum_{i=1}^{N} H\left(X_i\right) - H\left(X_1, X_2, \ldots, X_N\right) \tag{10}$$

Based on Eq. 2.1, it can be rewritten as,

$$TC\left(X_1, X_2, \ldots, X_N\right) = \sum_{x_1 \in X_1, x_2 \in X_2, \ldots, x_n \in X_n} p\left(x_1, x_2, \ldots, x_n\right) \log \frac{p(x_1, x_2, \ldots, x_n)}{\prod_{i=1}^{n} p(x_i)} \tag{11}$$

It can be noted that for the case of $N = 2$ (Bivariate Case), total correlation is equivalent to the well-known mutual information. However, we are interested in investigating variables normally $N > 2$ (N-variate Case) in the biological neural network. In the N-variate Case situation, we used total correlation to measured redundancy among variables in the system, and only under all variables are independent then total correlation will get zeros. In this study, total correlation is estimated through the definition of total correlation and two advanced information-theoretical estimators: Rotation-Based Iterative Gaussianization (RBIG) (Laparra et al., 2011, 2020) and the unsupervised learning model, CorEX. Ver Steeg and Galstyan (2014), respectively.

Estimating total correlation based on definition

The total correlation can be estimated directly from its definition. From Eq. (11), in order to get total correlation, we only need to estimate margin and joint entropies from variables and practically possible. If X be a random variable with a probability density function f whose support is a set $\mathcal{X}$. The continuous variable of differential entropy (Cover & Thomas, 1991) can be given as,

$$H(X) = -\int_{\mathcal{X}} f(x) \log f(x) dx \tag{12}$$

The primary step is first to discretize the continuous variable and then apply a James–Stein-type shrinkage which is an effective way in many scenarios with $p = 1000$ variables even with less than 100 to estimate entropy (Hausser & Strimmer, 2008). The discretization was performed by equal width binning, with 4 bins for each variable. The higher number of bins will give more accurate representation of original data distribution, but it will face dimensional and computing complexity problems because dimensional is increasing exponentially with the number of bins. After we estimate entropy and joint entropies, we can directly get total correlation from the definition of total correlation. Based on the definition of total correlation in the Eq. (11), the monotonicity property of total correlation could be used in the greedy clustering. The monotonicity property of total correlation could be written as,

$$TC\left(X_1, \ldots, X_p\right) = \left[\sum_{i=1}^{p} H\left(X_i\right)\right] - H\left(X_1, \ldots, X_p\right) \geq TC\left(X_1, \ldots, X_{p-1}\right) \tag{13}$$

RBIG[2] The RBIG approach was firstly proposed in 2011 by Laparra et al. (2011, 2020). The nature of RBIG is to convert any non-Gaussian distribution data into Gaussian distribution format (see Fig. 4). The RBIG is a cascade nonlinear plus linear transform for the input PDF of data.

$$\begin{aligned} &X_0 \rightarrow X_1 \rightarrow X_2 \rightarrow \cdots X^{n-1} \rightarrow X^n \\ &\boldsymbol{X}^{(l+1)} = R^{(l)} \cdot \Psi^{(l)}\left(\boldsymbol{X}_{\mathbf{0}}^{(l)}\right) \end{aligned} \tag{14}$$

where l indicates l-layer in the RBIG transform. R represents rotation after marginal Gaussization, Ψ of input data. Furthermore, variables of total correlation can calculate in any differentials transform the situation with:

$$\begin{aligned} \Delta TC\left(\mathrm{x}, \mathrm{x}'\right) &= TC(\mathrm{x}) - TC\left(\mathrm{x}'\right) \\ &= \sum_{i=1}^{D_x} H\left(x_i\right) - \sum_{i=1}^{D_{x'}} H\left(x_i'\right) + \frac{1}{2}\mathbb{E}_x \\ &\quad \times \left(\log\left|\nabla G_x(\mathrm{x})^{\top} \cdot \nabla G_x(\mathrm{x})\right|\right) \end{aligned} \tag{15}$$

Based on Eq. (15), the TC depends on the marginal entropy of variables and the Jacobin of transform function ($\mathbf{G}$) with each variables. After RBIG with non-Gaussian distribution data, $TC(\mathrm{x}' = 0)$ and TC can be defined as:

$$TC(\mathrm{x}) = \sum_{i=1}^{D_x} H\left(x_i\right) - \frac{D_x}{2}\log(2\pi e) + \mathbb{E}_x\left(\log\left|\nabla G_x(\mathrm{x})\right|\right) \tag{16}$$

The TC measure will be easy calculate if you know $\mathbb{E}_x(\log \nabla G_x(\mathrm{x}))$. However, the Jacobin of transform function with variable not easy to calculate because of its a multi-variable object, but RBIG algorithm through marginal Gaussian and rotation can easily solve the above problem, and TC will be:

$$TC(\mathrm{x}) = \sum_{n=0}^{N-1} \Delta TC^{(n)} = \frac{(N-1)D_x}{2}\log(2\pi e) - \sum_{n=1}^{N}\sum_{i=1}^{D_x} H\left(x_i^{(n)}\right) \tag{17}$$

where N is total samples and D indicates dimensionality of data.

CorEx[3] The unsupervised learning model named with Total Correlation Ex-planation, CorEx, which can be used for clustering variables via TC and it initially proposed by Greg in 2014 (Steeg & Galstyan, 2014, 2015). The main idea is to capture input data information hierarchical representations maximally in each hidden layer. The anatomy of the model is a bottom-up optimization procedure to capture maximum information of input data (see Fig. 5). The layer structure defined depends on input data distribution, and each hidden layer will maximum give a representation of the input data information. CorEx are given both upper and lower bounds to characterize the informativeness of the representation. In other words, the more profound layers can tighten bounds on the information in the data. In the following content, we will shortly introduce core mathematical equations of CorEx.

Basic decomposition of information: Assuming Y is a representation of X then we have,

$$TC_l(X; Y) = \sum_{i=1}^{n} I(Y : X_i) - \sum_{j=1}^{m} I(Y_j : X) \tag{18}$$

where TC_l indicate hidden layers, then the bound and decomposition hold.

$$TC(X) \geq TC(X_i; Y) = TC(Y) + TC_l(X; Y) \tag{19}$$

Hierarchical Lower Bound on TC(X): Assuming $Y^{1:l}$ refers a latent layer which hierarchical representation of X then we have,

$$TC(X) \geq \sum_{l=1}^{l} TC_l(Y^{l-1}, Y^l) \tag{20}$$

[2] https://www.uv.es/vista/vistavalencia/RBIG.htm.

[3] https://github.com/gregversteeg/CorEx.

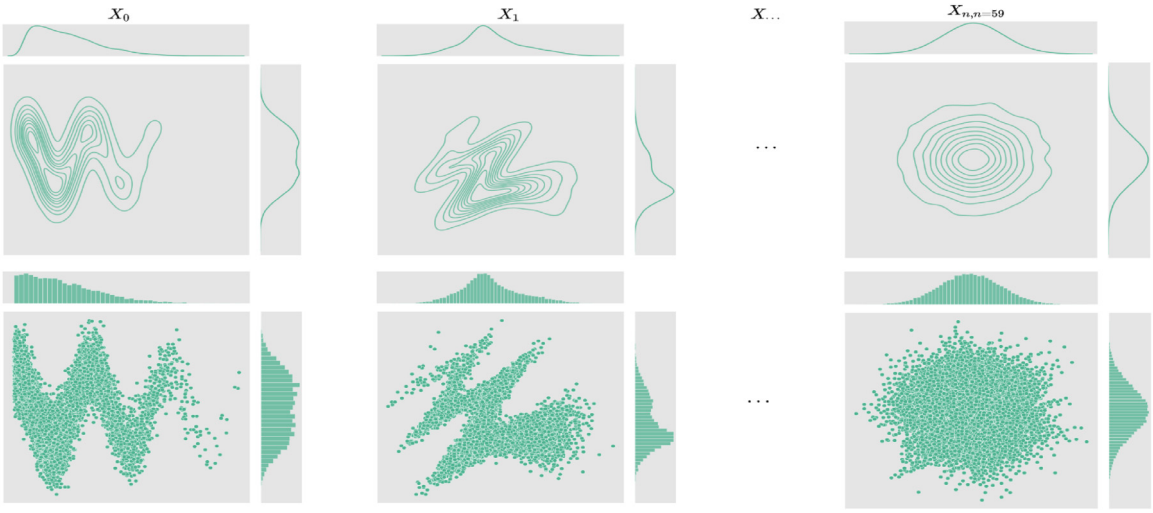


Fig. 4. The work processes of RBIG. RBIG transforms any non-Gaussian data distribution into a Gaussian distribution. The left side graph shows initial probability density distribution (PDF) and each dimensional distribution is non-Gaussian distribution. The middle figure shows data distribution after a transform (marginal Gaussization plus PCA rotation) through RBIG. The last figure shows data density distribution became to Gaussian distribution in each dimensional.

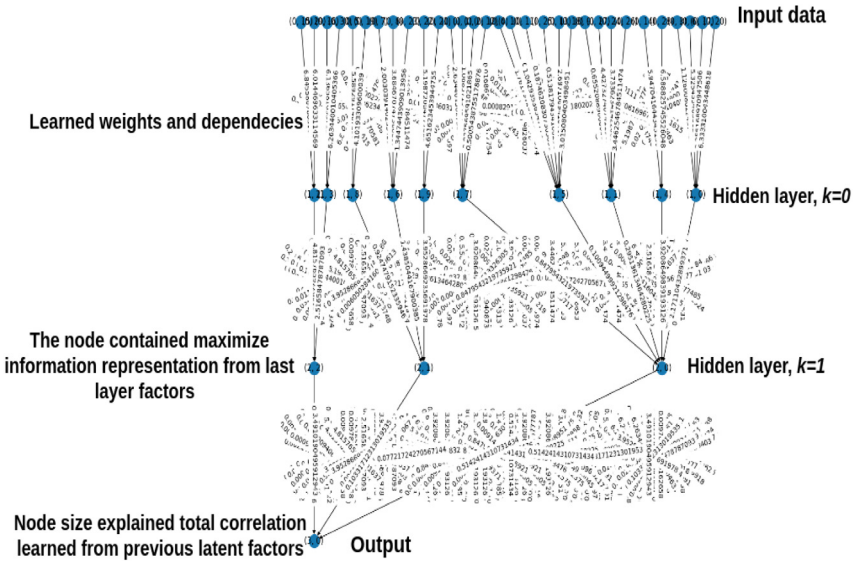


Fig. 5. The work processes of CorEx. The information flow in the inner of CorEx and core mechanism maximizes representation information in the CorEx. Edge thickness indicates the strength of dependence between factors, and node thickness indicates how each latent factor explains much total correlation.

Hierarchical Upper Bound on TC(X): Assuming $Y^{1:l}$ refers a latent layer which hierarchical representation of X, and additionally $m_l = 1$ and all variables are discrete, then we define,

$$TC(X) \leq \sum_{k=1}^{l} \left(TC_l\left(Y^{k-1}; Y^k\right) + \sum_{i=1}^{m_{k-1}} H\left(Y_i^{k-1} \mid Y^k\right) \right) \tag{21}$$

The CorEx uses iterative solutions to optimize and tight bounds until each latent layer maximum represents previous data information.

3. Application in the neuroscience

3.1. Synthetic neural data

We have generated BOLD signals with 2000 trails, and 10 000 times points for simulating neural signals from brain's A, B and C areas. In order to investigate how each component affects functional connectivity among brain regions, we used the same previous studies simulation approach (Cole, Yang, Murray, Repovs, & Anticevic, 2015), but we considered nonlinear case in our simulation. The time series for A, B and C mainly consist with shared neural signal (S_{abc}), non-shared neural signal (NS_a, NS_b, NS_c), and noise (n_a, n_b, n_c) across brain regions (Fig. 6). The basic neural time series can be formula as:

$$\begin{aligned} T_A &= SS_{abc} + NS_a + n_a \\ T_B &= |SS_{abc} + NS_b + n_b|^e \\ T_C &= |SS_{abc} + NS_c + n_c|^e \end{aligned} \tag{22}$$

where S indicates shared weights, N indicates non-shared weights and $n_a \sim \mathcal{N}(0, \sigma^2)$, $n_b \sim \mathcal{N}(0, \sigma^2)$, and $n_c \sim \mathcal{N}(0, \sigma^2)$, symbols $\|$ indicates absolute value computing, here e is constant value, 2.5. We generated 10 000 normally distributed time points. In order to investigate sensitivity of information-theoretical under different functional connectivity states, we through adjusted the

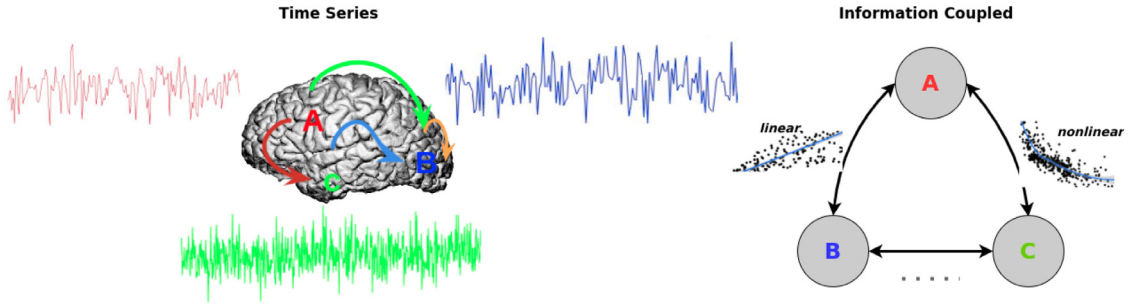


Fig. 6. Simulated neural signal of three inter-brain regions (e.g., three nodes). (Left figure) formed a complex system such as the brain, we try to estimate the underlying information dependencies (Right figure), accounting for linear and nonlinear dependencies. Coupled discovery aims to unveil such dependencies, leading to estimated information coupled networks..

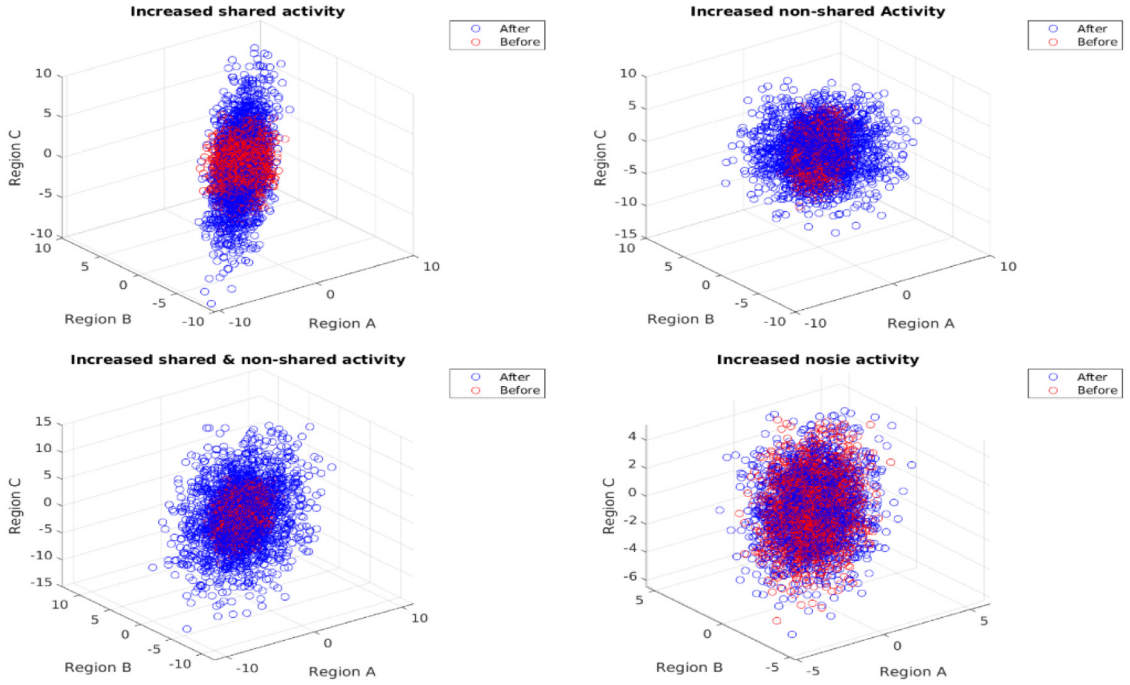


Fig. 7. The simulated functional connectivity states in the inter-brain regions under different situations. The functional connectivity affects by changing noise, shared and non-shared neuron pool activity in the brain. The blue dot infers functional connectivity states before, and the red dot indicates functional connectivity states changed via control parameters in the models. Here constant value $e = 1$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Coupled information among three brain regions under various situations. We reported pair-wised mutual information, such as $I(A;B), I(A;C), I(B;C)$, interaction information, $II(A;B;C)$, total correlation, $TC(A,B,C)$ and their percentage of information changed comparison with baseline under different situations. Where $\uparrow SS_{abc}(0.8\times)$ refers to increase shared activity $0.8\times$, $\uparrow (NS_a, NS_b, NS_c)(2.5\times)$ indicates increase non-shared activity $2.5\times$, $\uparrow SS_{abc}$& $\uparrow (NS_a, NS_b, NS_c)(2.5\times)$ indicates increase both shared and non-shared activity $0.8\times$, and $\uparrow (n_a, n_b, n_c)(1.8\times)$ indicates increase noise activity $1.8\times$, respectively. When we increased shared activity $(0.8\times)$, we found that mutual information was not very sensitivity of connectivity as opposed to interaction information and total correlation. The same situation also happened in the other three cases. All values with 2 decimal places were presented in the table.

States	Information/bits				
	$I(A;B)$	$I(A;C)$	$I(B;C)$	$II(A;B;C)$	$TC(A,B,C)$
Baseline	0.76	0.78	1.68	4.72	6.54
$\uparrow SS_{abc}$ $(0.8\times)$	0.76 [0.00%]	0.78[0.00%]	1.08[35.71%]	8.51[80.30%]	9.12[39.45%]
$\uparrow (NS_a, NS_b, NS_c)$ $(2.5\times)$	0.75[1.32%]	0.75[3.85%]	0.90[46.43%]	1.98[58.05%]	2.28[65.14%]
$\uparrow SS_{abc}$& $\uparrow (NS_a, NS_b, NS_c)$ $(0.8\times)$	0.78[2.56%]	0.77[1.28%]	1.55[7.74%]	7.99[69.28%]	9.77[49.39%]
$\uparrow (n_a, n_b, n_c)$ $(1.8\times)$	0.76[0.00%]	0.76[0.00%]	1.21[27.98%]	2.48[47.46%]	3.32[49.24%]

weights of shared neural activity, non-shared neural activity and noise amplitude to study it, and mutual information, interaction information and total correlation measured under below four conditions (see Fig. 7).

From Table 1, we can get mutual information, interaction information, and total correlation under various functional connectivity changed among triplet neurons compared to original signals. We found that multivariate mutual information capture

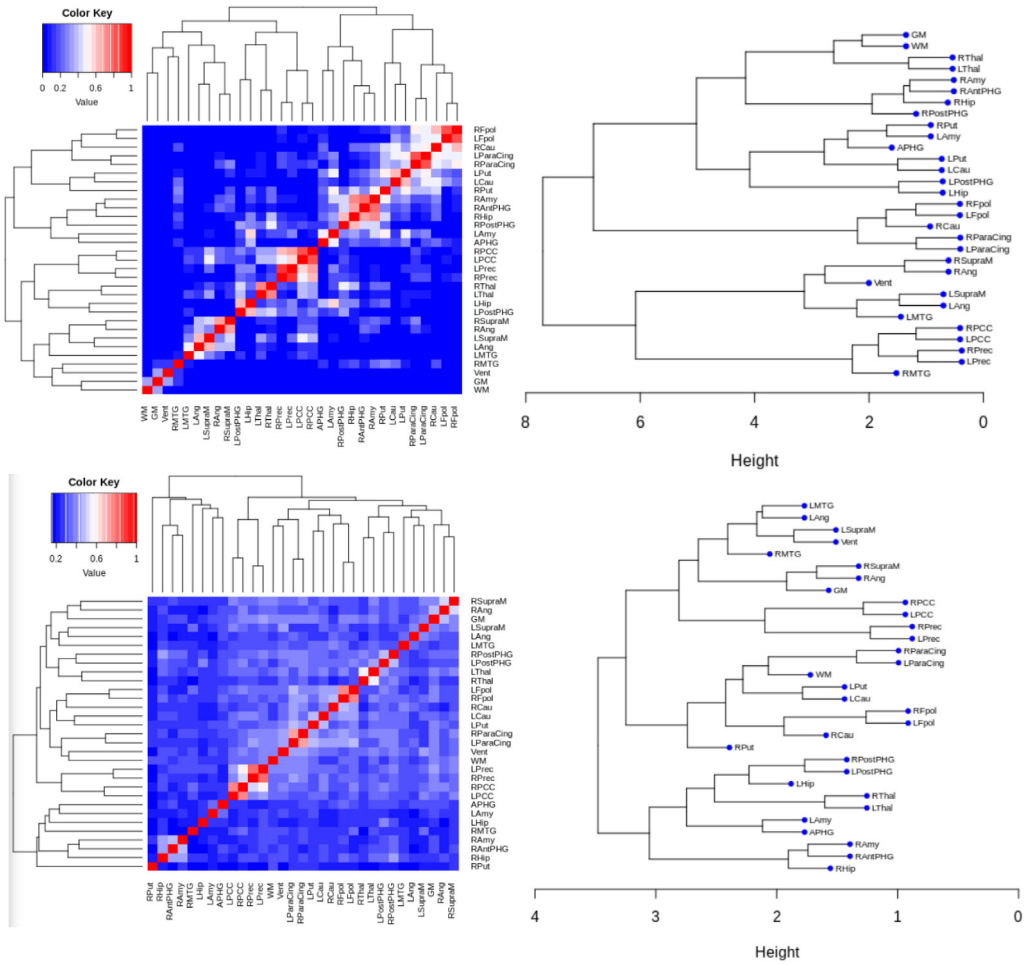


Fig. 8. Hierarchical clustering dendrograms. The graph was depicted the hierarchical clustering dendrograms for linear correlation coefficient and mutual information with a threshold of 0.02.

more information compared to mutual information in the high dimensional and nonlinear distribution neural signals. It also can be confirmed from Venn Diagram Fig. 1. If we increase the information coupled among triplet neurons, there will be an increase in mutual information, interaction information, and total correlation. The opposite will be true if we increase non-shared activity. Briefly stated, we have demonstrated that multivariate information measures are more robust and sensitive than mutual information measures. However, interaction information is more difficult to interpret when it is negative as compared to total correlation.

3.2. Real neural datasets

3.2.1. Resting-state fMRI

The data came from a resting-state fMRI experiment in which a single participant kept alert wakefulness while watching across but did not perform any other behavioral activity. The data was preprocessed, and time series from various brain regions of interest (ROIs) were collected. The ROIs are listed as follows,

Cau, **Caudate**; Pau, **Paudate**; Thal, **Thalamus**; Fpol, **Frontal pole**; Ang, **Angular gyrus**; SupraM,**Supramarginal Gyrus**; MTG, **Middle Temporal Gyrus**; Hip, **Hippocampus**; PostPHG, **Posterior Parahippocamapl gyrus**; APHG, **Anterior parahippocamapl gyrus**; Amy, **Amygdala**; ParaCing, **Paracingulate gyrus**; PCC, **Posterior cingulatecortex**; Prec, **Precuneus**.

The dendrograms of Pearson correlation and mutual information were computed with hierarchical clustering method using linkage methods ward.D2. In Fig. 8, the connectivity matrix was measured with Pearson correlation and mutual information, respectively. The mutual information capture more dependence compared to Pearson correlation (see Fig. 8). Meanwhile, on the right side of the hierarchical clustering tree, revealing more clear dependency among ROIs in the brain, e.g., The LPCC-RPCC, LPrec-RPrec, are clustered together in correlation and mutual information. The left and right hemispheres have symmetric connectivity except for a few subcortical ROIs, e.g., Put, Cau, Amy, Hip, et al.

Next, we applied the method that takes the multivariate dependences measure with total correlation from definition into account. However, the greedy clustering results in a single cluster merger variables and being grown larger and larger. Namely, the merges in the first eight steps cluster are the following:

{LParaCing, RParaCing} → {LParaCing, RParaCing, LPut} →
{LParaCing, RParaCing, LPut, LCau} →
{LParaCing, RParaCing, LPut, LCau, RCau} →
{LParaCing, RParaCing, LPut, LCau, RCau, LAmy} →
{LParaCing, RParaCing, LPut, LCau, RCau, LAmy, APHG} →
{LParaCing, RParaCing, LPut, LCau, RCau, LAmy, APHG, RMTG}

The merges in the second eight steps cluster are the following (see Table 2):

{RPrec, LPrec} → {RPrec, LPrec, RPCC} →
{RPrec, LPrec, RPCC, LPCC} →

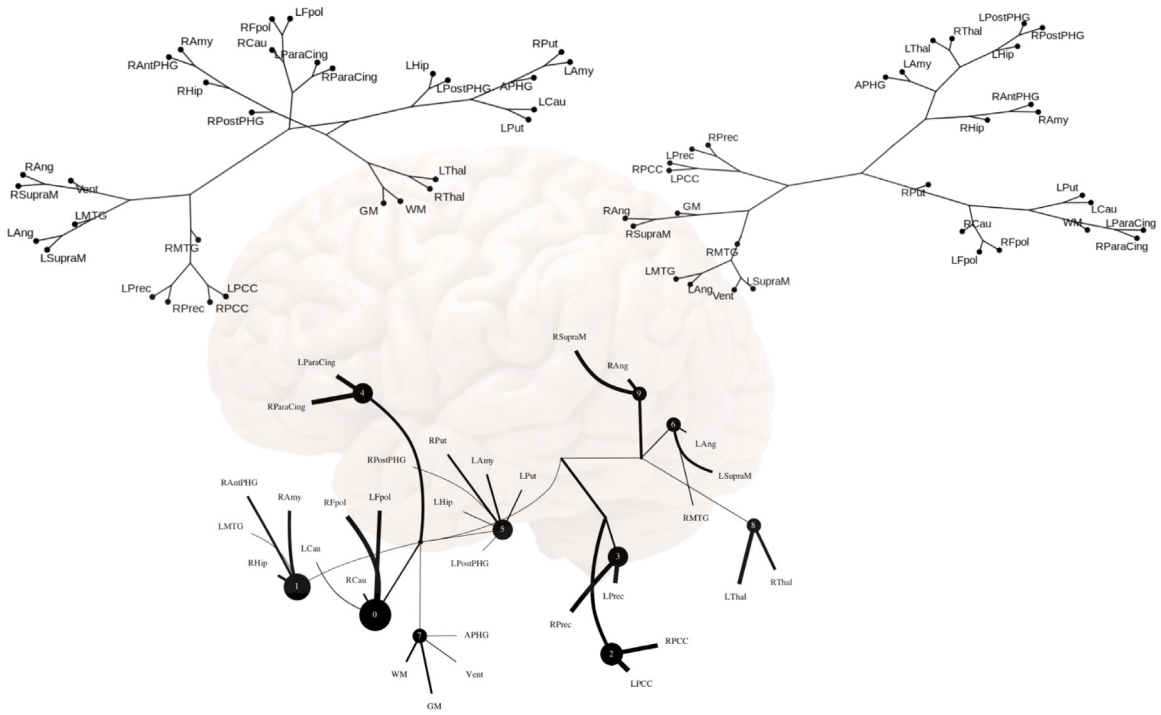


Fig. 9. Graph of latent factors for functional connectivity measures constructed by correlation, mutual information and total correlation. The tree graphs illustrated correlation (left tree), mutual information (right tree), and total correlation (bottom tree) derived functional connectivity in the brain in which corresponding to connectivity matrix in Fig. 8. The diagram presented bottom graphs with thresholded weights at 0.19 for visualizing. The edge thickness is proportional to mutual information and node size, reflecting the total correlation among child ROIs.

Table 2
Multivariate mutual information measured with minimally connected clusters.

Cluster size	Cluster members	Information theoretic measures/bits
2	{LThal, RThal}	0.1064
2	{LPCC, RPCC}	0.1813
2	{LPrec, RPrec}	0.2170
2	{RAng, RSupraM}	0.1057
2	{LFpol, RFpol}	0.1714
3	{LCau, RCau, RAmy}	0.1007
3	{RHip, RAntPHG, RAmy}	0.1976
3	{RHip, RPostPHG, RAmy}	0.1369

{RPrec, LPrec, RPCC, LPCC, LSupraM} →
{RPrec, LPrec, RPCC, LPCC, LSupraM, LAng} →
{RPrec, LPrec, RPCC, LPCC, LSupraM, LAng, RAmy} →
{RPrec, LPrec, RPCC, LPCC, LSupraM, LAng, RAmy, RHip}

This can be readily explained by the monotonicity of the total correlation already discussed: clusters with a high number of variables have an advantage when their total correlation is compared to smaller clusters because of inflation of total correlation. For the CorEx approach, we fit the CorEx model consisting of three layers, 10, 3, and 1 unit, to measure multivariate mutual information among neurons. We found that total correlation capture more relationship among ROIs.

In Fig. 9, while the dendrogram trees are globally different, many similar structures can be observed: LPCC-RPCC, LPCC-RPCC-LPrec-RPrec, LFpol-RFpol, LThal-RThal are examples for structures that remain unchanged, indicating no strong presence of non-linear relationships. Meanwhile, some unknown relationships were found under total correlation measures, e.g., LCau-RCau-RFpol-LFpol, RHip-LMTG-RAmy-RAntPHG. In summary, theoretical information approaches can quantify information coupled states in the brain, but interpreting results with information-theoretic results is always challenging.

3.2.2. Brain development fMRI

Datasets

The data was taken from a task-related fMRI experiment in which participants watched a silent version of "Partly Cloudy", a 5.6-min animated movie (Jacoby, Bruneau, Koster-Hale, & Saxe, 2015) (see Fig. 10). A short description of the plot can be found online.[4] Meanwhile, the BOLD signal was recorded. The data was pre-processed, and time series were extracted from different regions of interest (ROIs) with pre-defined atlas (Richardson, Lisandrelli, Riobueno-Naylor, & Saxe, 2018). The data is down-sampled to 4 mm resolution for convenience with a repetition time (TR) of 2 secs. The origin of the data is coming from OpenNeuro.[5] In the experiments, the prepossessed data included 122 children with ages 3–12 and can be directly downloaded with nilearn.datasets function.[6]

BOLD signal construction

A spatially constrained parcellation, MSDL, was used for extracting BOLD signals, and it can be download from here[7] (Varoquaux & Craddock, 2013; Varoquaux, Gramfort, Pedregosa, Michel, & Thirion, 2011) (see Fig. 11). We extract time-series for each ROI in each subject, then weighted average fMRI BOLD signals over all voxels within that specific region. Furthermore, each region's BOLD signal is normalized and detrended for the following information-theoretical measures.

Functional connectivity of brain development with information-theoretical measures

Understanding functional connectivity altered with brain development is a crucial researches field in which not only help us

[4] https://www.pixar.com/partly-cloudy#partly-cloudy-1.
[5] https://openneuro.org/datasets/ds000228/versions/1.0.0.
[6] https://nilearn.github.io/modules/generated/nilearn.datasets.fetch_development_fmri.html.
[7] https://team.inria.fr/parietal/files/2015/01/MSDL_rois.zip.

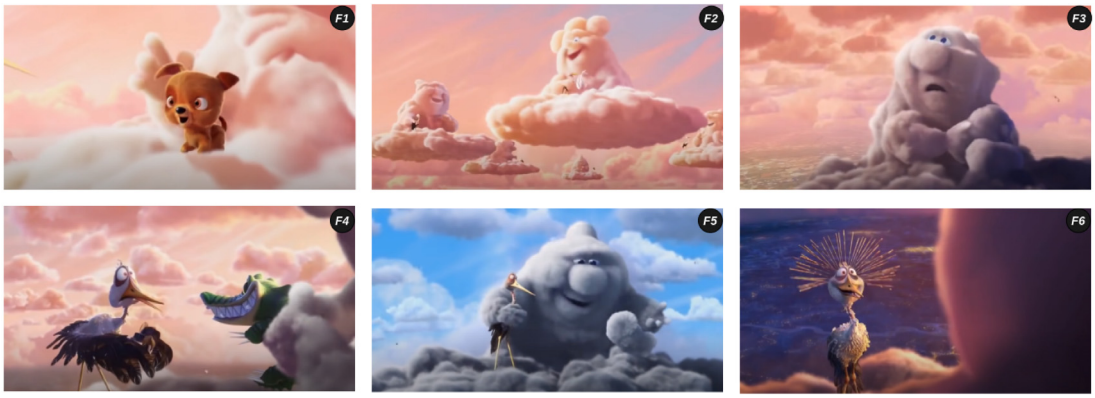


Fig. 10. Visual stimuli frames. Example frames for the six events from Partly Cloudy. Images ©2009 Pixar, reused with permission. These images are not covered under the CC BY license for this article.

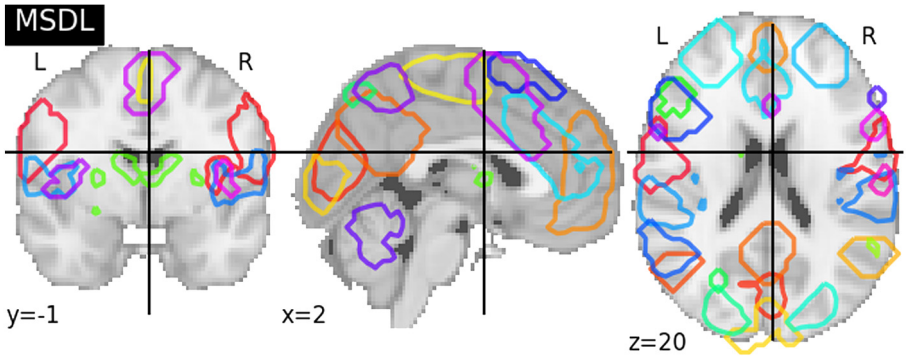


Fig. 11. Pre-defined ROI atlas. MSDL atlas which including 39 ROIs in total.

further study brain function and help us diagnose neurodevelopmental disorders and learning disabilities.

In Fig. 12, while the dendrogram trees are globally different, many similar structures can be observed: LSTS-RSTS, LDMN-RDMN, LLOC-RLOC, LTPJ-RTPJ, LAud-RAud, et al. are examples for structures that remain unchanged, indicating no strong presence of nonlinear relationships. The result of correlation, mutual information, and total correlation is most consistent because they all capture the pair-wise relationship. However, total correlation gives more intuitive results than correlation and mutual information cluster trees, e.g., LIns-RIns-V ACC LSTS-RSTS-Cing. Multivariate information-theoretic approaches do not find more relationships than correlation and mutual information with brain development fMRI because function connectivity or information coupled among brain regions is not fully developed in the children.

4. Discussion and conclusion

Information theory provides a principled methodology for studying and quantifying functional connectivity based on statistical relationships between variables. As we have reviewed, the foundational quantities of information theory are entropy and mutual information. Here we have tried to use multivariate mutual information approach to the synthetic and practical estimation of these quantities. We proved that multivariate mutual information results are consistent with traditional methods and found some unknown relationships when estimated functional connectivity. However, there are also have some limitations in this study are presented in the following contents.

Firstly, In the simulation study, we simulated neural signals under Gaussian distribution. The reason for that is that comparison under different functional connectivity states how the information coupled between or among neurons changed. While we found multivariate information theory could capture more relationships among neurons, the neural signal could not fit Gaussian distribution or even more complex distribution under real cases. Secondly, we applied interaction information and total correlation in the simulation studies, which proved that both high-order information-theoretic methods could be used to quantify information flow derived functional connectivity among brain regions. However, we have not applied interaction information in practical studies because it is hard to interpret it when it becomes negative properly. As we aforementioned, theoretical information approaches are easy to apply to analyze neural signals, but it is very hard to properly interpret it from statistics and neuroscience perspectives. Thirdly, we used a pre-defined atlas, MSDL, to extract neural signals, and it only includes 39 ROIs, so we only explored the limited relationships with these ROIs. Moreover, considering the dense overlap and distribution properties of neural signals, multivariate mutual information could not find more prosperous relationships because of the dataset itself problems, specific fMRI datasets because it supplied more spatial information than time-order information. Fourthly, estimating total correlation from its definition, we will have a memory problem when we have a larger-scale neural signal with a larger bin value, and the accuracy of total correlation estimated were not considered in this research. It could be as extension work for the future studies.

This paper presented two multivariate generalizations of mutual information, interaction information, and total correlation. Firstly, quantitative measures intra-cortex regions dependent or independent from others from the information-theoretic views. Secondly, Total correlation is an efficient way to assess the functional connectivity of human brain, according to the findings.

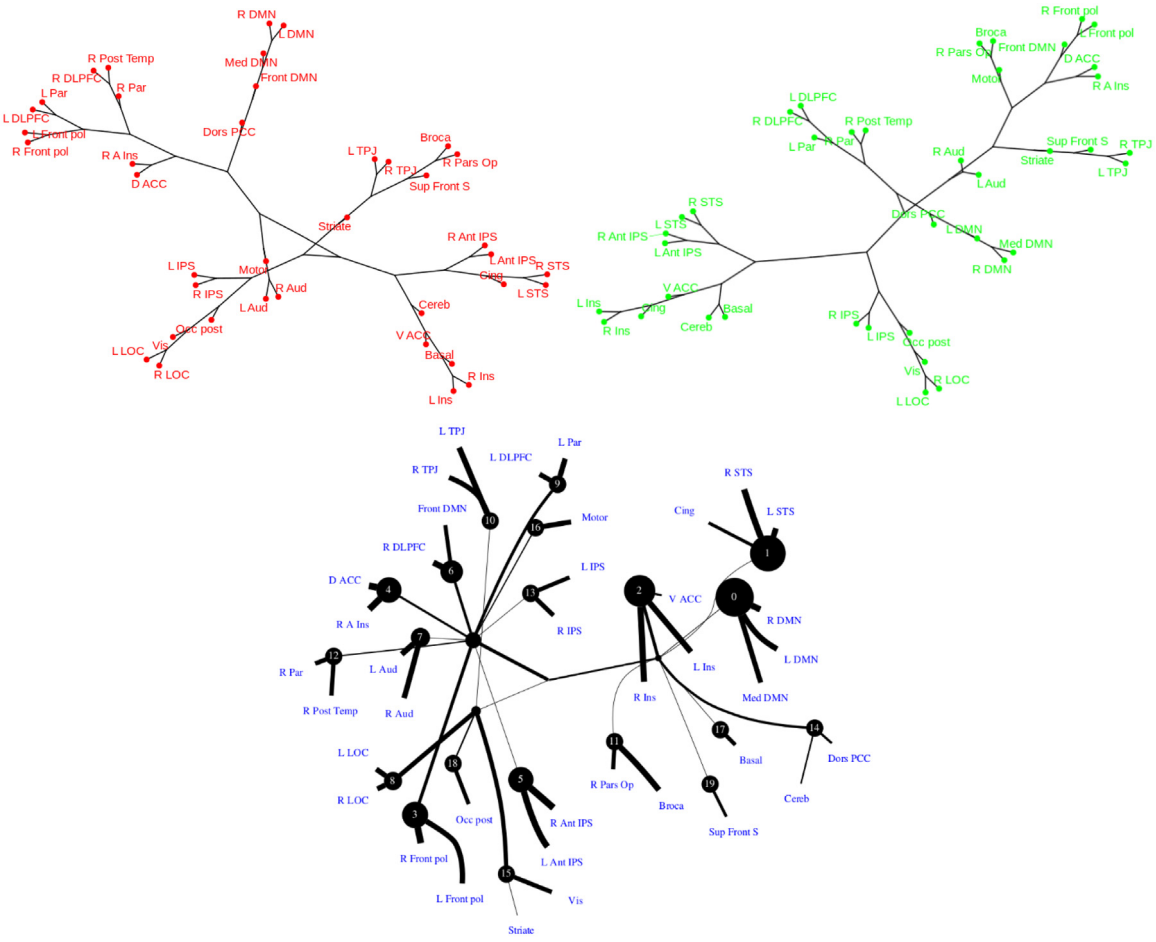


Fig. 12. Correlation, mutual information, and total correlation are used to generate a graph of latent variables for functional connectivity metrics. The tree graphs illustrated correlation (left tree with red color), mutual information (right tree with green color), and total correlation (bottom tree with blue color) derived functional connectivity in the development brain. The diagram presented bottom graphs with threshold weights at 0.16 for visualizing, and edge thickness is proportional to mutual information and node size, reflecting the total correlation among child ROIs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Thirdly, total correlation is a powerful clustering method with a graph to find multivariate independence, and it has the potential to apply to other field investigations in the future. Moreover, the main goal of this paper is to show that there is not just one straightforward generalization of pair-wise mutual information for the multivariate cases, and neuroscience researchers that want to exploit high-order information measures in an information-theoretic framework should take this fact into account. Moreover, we have applied the two different measures to the functional connectivity in the brain, and demonstrated that the results might differ significantly. These are just experimental and rudimentary observations; more research into the exact nature of both generalizations and their repercussions for neuroscience – as well as a proper quantitative evaluation – is imperative. This brings us to some avenues for future work. More research needs to be carried about the exact nature of the dependencies that both measures capture. Preliminary results show that they extract different information, but it is unclear what the exact nature of that information is. Secondly, we want to conduct a proper quantitative evaluation on different cognitive behavior tasks to indicate which measure works best and which measure might be more suitable for a particular task.

5. Code

In this study, there is hybrid software used to study multivariate mutual information. The Matlab libraries are used to estimate conditional mutual information is listed as follows, GCMI.[8] Python libraries are used to download dataset and estimate multivariate mutual information are listed as follows, Nilearn,[9] RBIG,[10] CorEx.[11] R packages used to visualize tree graph in this studies are listed as follows: factoextra,[12] igraph,[13] entropy,[14] cluster,[15] and gplots.[16] Furthermore, the source code for all experiments can be found at: https://github.com/sinodanishspain/MI_II_TC_2021.

8 https://github.com/robince/gcmi/blob/master/matlab/.
9 https://nilearn.github.io/.
10 https://isp.uv.es/RBIG4IT.htm.
11 https://github.com/gregversteeg/CorEx.
12 https://cran.r-project.org/web/packages/factoextra/index.html.
13 https://igraph.org/r/.
14 http://www.strimmerlab.org/software/entropy/.
15 https://svn.r-project.org/R-packages/trunk/cluster/.
16 https://cran.r-project.org/web/packages/gplots/index.html.

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

A number of anonymous reviewers provided fruitful remarks and comments on an earlier draft of this paper, from which the current version has significantly benefited. I also thank to all of the open-source contributors in the open-source community. This work was partially funded by the grants: Grisolía-P/2019/035, MICINN DPI2017-89867-C2-2-R and MICINN PID2020-118071GB-I00.

Appendix

This paper also a extension research for my poster[17] which was presented at entropy2021 conference.[18] The preliminary findings were summarized in a poster at the entropy2021 conference. Furthermore, we only compared Pearson correlation and Total correlation ($n > 3$)/Mutual information ($n = 2$) application on the measure of information flow in the human intra-cortex. However, In this manuscript, We mainly tried to do three things. First, we mainly explored the difference between mutual information, interaction information, and total correlation. Second, we explored how to estimate it correctly. Third, we investigated this high-order information-theoretical approach to measuring human functional connectivity with fMRI data and confirmed that above mentioned high-order information theory could capture rich dependency.

References

Akiki, Teddy, & Abdallah, Chadi (2019). Determining the hierarchical architecture of the human brain using subject-level clustering of functional networks. *Scientific Reports*, *9*, 19290.

Bandettini, Peter (2012). Twenty years of functional MRI: The science and the stories. *NeuroImage*, *62*, 575–588.

Bell, Anthony J. (2003). The co-information lattice. In *Proc. 4th int. symp. independent component analysis and blind source separation* (pp. 921–926).

Carhart-Harris, Robin, Leech, Robert, Hellyer, Peter, Shanahan, Murray, Feilding, Amanda, Tagliazucchi, Enzo, et al. (2014). The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience*, *8*, 20.

Chan, Chung, Al-Bashabsheh, Ali, & Zhou, Qiaoqiao (2020). Agglomerative info-clustering: Maximizing normalized total correlation. *IEEE Transactions on Information Theory*, *PP*, 1.

Cole, Michael, Yang, Genevieve, Murray, John, Repovs, Grega, & Anticevic, Alan (2015). Functional connectivity change as shared signal dynamics. *Journal of Neuroscience Methods*, *259*.

Cormen, Thomas H., Leiserson, Charles E., Rivest, Ronald L., & Stein, Clifford (2001). *Introduction to algorithms* (2nd ed.). The MIT Press.

Cover, Thomas M., & Thomas, Joy A. (1991). *Elements of information theory*. New York: Wiley.

Cutts, Catherine, & Eglen, Stephen (2014). Detecting pairwise correlations in spike trains: An objective comparison of methods and application to the study of retinal waves. *Journal of Neuroscience*, *34*, 14288.

Eqlimi, Ehsan, riyahi alam, Nader, Sahraian, Mohammadali, Eshaghi, Arman, & Rad, Hamidreza (2014). Mutual information weighted graphs for resting state functional connectivity in fMRI data. *Joint Annual Meeting ISMRM-ESMRMB, Milan, Italy*.

Everitt, Brian, Landau, Sabine, Leese, Morven, & Stahl, Daniel (2011). *Cluster analysis* (5th ed.). Wiley.

Farahani, Farzad, Karwowski, Waldemar, & Lighthall, Nichole (2019). Application of graph theory for identifying connectivity patterns in human brain networks: A systematic review. *Frontiers in Neuroscience*, *13*, 585.

Ferenci, Tamas, & Kovács, Levente (2014). Using total correlation to discover related clusters of clinical chemistry parameters. *SISY 2014 - IEEE 12th International Symposium on Intelligent Systems and Informatics, Proceedings*, 49–54.

Fischer, J. (1995). Hierarchical cluster analysis. *Computational Statistics*, *10*.

Gao, Shuyang, Brekelmans, Rob, Steeg, Greg Ver, & Galstyan, Aram (2019). Auto-encoding total correlation explanation. In Kamalika Chaudhuri, & Masashi Sugiyama (Eds.), *Proceedings of machine learning research*: *Vol. 89*, *Proceedings of the twenty-second international conference on artificial intelligence and statistics* (pp. 1157–1166). PMLR.

Gao, Shuyang, Ver Steeg, Greg, & Galstyan, Aram (2015). Efficient estimation of mutual information for strongly dependent variables. ArXiv abs/1411.2003.

Gao, Zhongke, Weidong, Dang, Wang, Xinmin, Hong, Xiaolin, Hou, Linhua, Ma, Kai, et al. (2021). Complex networks and deep learning for EEG signal analysis. *Cognitive Neurodynamics*, *15*.

Garcia, Beatriz, Fernández-Caballero, Antonio, & Martinez Rodrigo, Arturo (2021). Entropy and the emotional brain: Overview of a research field. In *Brain-Computer Interface*. intechopen.98342.

Gomez-Villa, Alex, Bertalmío, Marcelo, & Malo, Jesús (2020). Visual information flow in wilson-cowan networks. *Journal of Neurophysiology*, *123*.

Goodman, N. (1963). The distribution of the determinant of a complex wishart distributed matrix. *Annals of Mathematical Statistics*, *34*.

Hausser, Jean, & Strimmer, Korbinian (2008). Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, *10*.

Hlinka, Jaroslav, Palus, Milan, Vejmelka, Martin, Mantini, Dante, & Corbetta, Maurizio (2011). Functional connectivity in resting-state fMRI: Is linear correlation sufficient? *NeuroImage*, *54*, 2218–2225.

Ince, Robin, Giordano, Bruno, Kayser, Christoph, Rousselet, Guillaume, Gross, Joachim, & Schyns, Philippe (2016). A statistical framework for neuroimaging data analysis based on mutual information estimated via a Gaussian copula. *Human Brain Mapping*, *38*.

Ince, Robin, Montani, Fernando, Arabzadeh, Ehsan, Diamond, Mathew, & Panzeri, Stefano (2009). On the presence of high-order interactions among somatosensory neurons and their effect on information transmission. *Journal of Physics: Conference Series*, *197*, Article 012013.

Jacoby, Nir, Bruneau, Emile, Koster-Hale, Jorie, & Saxe, Rebecca (2015). Localizing pain matrix and theory of mind networks with both verbal and non-verbal stimuli. *NeuroImage*, *126*.

Jakulin, Aleks, & Bratko, Ivan (2004). Testing the significance of attribute interactions. In *Proceedings of the twenty-first international conference on machine learning (ICML-2004): 2004*, Banff, Canada.

Jakulin, Aleks, Bratko, Ivan, Smrke, Dragica, Demsar, Janez, & Zupan, Blaz (2003). *Attribute interactions in medical data analysis, Vol. 2780*. Protarus: Cyprus.

Jaynes, Edwin (1957). Information theory and statistical mechanics I. *Physical Review*, *106*, 620–630.

Keshmiri, Soheil (2020). Entropy and the brain: An overview. *Entropy*, *22*, 917.

Kraskov, Alexander, Stögbauer, Harald, & Grassberger, Peter (2004). Estimating mutual information. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, *69*, Article 066138.

Laparra, Valero, Camps-Valls, Gustau, & Malo, Jesus (2011). Iterative gaussianization: from ICA to random rotations. *IEEE Transactions on Neural Networks*, *22*(4), 537–549.

Laparra, Valero, Johnson, J. Emmanuel, Camps-Valls, Gustau, Santos-Rodríguez, Raul, & Malo, Jesus (2020). ArXiv abs/2010.03807.

Li, Wentian (1990). Mutual information functions versus correlation functions. *Journal of Statistical Physics*, *60*, 823–837.

Li, Qiang (2021). Measuring functional connectivity of human intra-cortex regions with total correlation. *Proceedings of the Entropy 2021: The Scientific Tool of the 21st Century* (p. 9797). Entropy2021-09797.

Li, Qiang, Johnson, Emmanuel, Esteve-Taboada, Jose, Laparra, Valero, & Malo, Jesús (2021). Computing variations of entropy and redundancy under nonlinear mappings not preserving the signal dimension: quantifying the efficiency of V1 cortex. In *Proceedings of the Entropy 2021: The Scientific Tool of the 21st Century* (p. 9813). Entropy2021-09813.

Ma, Jian, & Sun, Zengqi (2011). Mutual information is copula entropy. *Tsinghua Science & Technology*, *16*, 51–54.

Mansoory, Meysam, Oghabian, Mohammad, Jafari, Amir, & Shahbabaie, Alireza (2017). Analysis of resting-state fMRI topological graph theory properties in methamphetamine drug users applying box-counting fractal dimension. *Basic and Clinical Neuroscience Journal*, *8*, 371–386.

McGill, William (1954). Multivariate information transmission. *Psychometrika*, *19*, 97–116.

Mijatović, Gorana, Loncar-Turukalo, Tatjana, Bozanic, Nebojsa, Milosavljevic, Nina, Storchi, Riccardo, & Faes, Luca (2021). A measure of concurrent neural firing activity based on mutual information. *Neuroinformatics*, 1–17.

Misra, Neeraj, Singh, Harshinder, & Demchuk, Eugene (2005). Estimation of the entropy of a multivariate normal distribution. *Journal of Multivariate Analysis*, *92*, 324–342.

[17] https://sciforum.net/manuscripts/9797/slides.pdf.

[18] https://entropy2021.sciforum.net/.

Moon, Young-Il, Rajagopalan, Balaji, & Lall, Upmanu (1995). Estimation of mutual information using kernel density estimators. *Physical Review. E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, *52*, 2318–2321.

Nowak, Gen, & Tibshirani, Robert (2007). Complementary hierarchical clustering. *Biostatistics*, *9*(3), 467–483.

Richardson, Hilary, Lisandrelli, Grace, Riobueno-Naylor, Alexa, & Saxe, Rebecca (2018). Development of the social brain from age three to twelve years. *Nature Communications*, *9*.

Rosenkrantz, R. (1989). Information theory and statistical mechanics II (1957). (pp. 17–38).

Saxe, Glenn, Calderone, Daniel, & Morales, Leah (2018). Brain entropy and human intelligence: A resting-state fMRI study. *PLoS One*, *13*, Article e0191582.

Saxena, Amit, Prasad, Mukesh, Gupta, Akshansh, Bharill, Neha, Patel, op, Tiwari, Aruna, et al. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*.

Scharf, Louis, & Mullis, C. T. (2000). Canonical coordinates and the geometry of inference, rate, and capacity. *IEEE Transactions on Signal Processing*, *48*, 824–831.

Shannon, Claude E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Singh, Harshinder, Hnizdo, V., Demchuk, Adam, & Misra, Neeraj (2003). Nearest neighbor estimates of entropy. *American Journal of Mathematical and Management Sciences*, *23*.

Steeg, Greg Ver, & Galstyan, Aram (2014). Discovering structure in high-dimensional data through correlation explanation. In *Advances in neural information processing systems, NIPS'14*.

Steeg, Greg Ver, & Galstyan, Aram (2015). Maximally informative hierarchical representations of high-dimensional data. In *AISTATS'15*.

Stozer, Andraz, Gosak, Marko, Dolenšek, Jurij, Perc, Matjaž, Marhl, Marko, Slak Rupnik, Marjan, et al. (2013). Functional connectivity in islets of langerhans from mouse pancreas tissue slices. *PLoS Computational Biology*, *9*, Article e1002923.

Studený, M., & Vejnarová, J. (1999). The multiinformation function as a tool for measuring stachastic dependence. In *Learning in graphical models* (pp. 261–297). Cambridge, MA, USA: MIT Press, ISBN: 0262600323.

Tedeschi, W., Müller, Hans-Peter, Ludolph, A., Kraft, H., de Araujo, Draulio, Neves, Ubiraci, et al. (2003). A new method for the analysis of functional magnetic resonance imaging data: Mutual information tests. *Biomedizinische Technik - BIOMED TECH*, *48*, 102–103.

Timme, Nicholas, & Lapish, Christopher (2018). A tutorial for information theory in neuroscience. *Eneuro*, *5*, ENEURO.0052–18.2018.

Uehara, Taira, Tobimatsu, Shozo, Kan, Shigeyuki, & Miyauchi, Satoru (2012). Modular organization of intrinsic brain networks: A graph theoretical analysis of resting-state fMRI. In *ICME International Conference on Complex Medical Engineering (CME)*. ISBN: 9781467316187, ICCME.2012.6275597.

Varoquaux, Gael, & Craddock, Cameron (2013). Learning and comparing functional connectomes across subjects. *NeuroImage*, *80*.

Varoquaux, Gael, Gramfort, Alexandre, Pedregosa, Fabian, Michel, Vincent, & Thirion, Bertrand (2011). Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Information processing in medical imaging: proceedings of the conference, Vol. 22* (pp. 562–573).

Ver Steeg, Greg (2017). Unsupervised learning via total correlation explanation. In *International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 5151–5155).

Ver Steeg, Greg, & Galstyan, Aram (2014). Discovering structure in high-dimensional data through correlation explanation. *Advances in Neural Information Processing Systems*, *1*.

Walczak, Zbigniew (2008). Total correlations and mutual information. *Physics Letters, Section A: General, Atomic and Solid State Physics*, *373*.

Wang, Yanlu, Msghina, Mussie, & Li, Tie-Qiang (2016). Studying sub-dendrograms of resting-state functional networks with voxel-wise hierarchical clustering. *Frontiers in Human Neuroscience*, *10*.

Wang, Jinhui, Zuo, Xi-Nian, & He, Yong (2010). Graph-based network analysis of resting-state functional MRI. *Frontiers in Systems Neuroscience*, *4*, 16.

Watanabe, Satosi (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of Research Developement*, *4*(1), 66–82.

Wyner, A. D. (1978). A definition of conditional mutual information for arbitrary ensembles. *Information and Control*, *38*, 51–59.

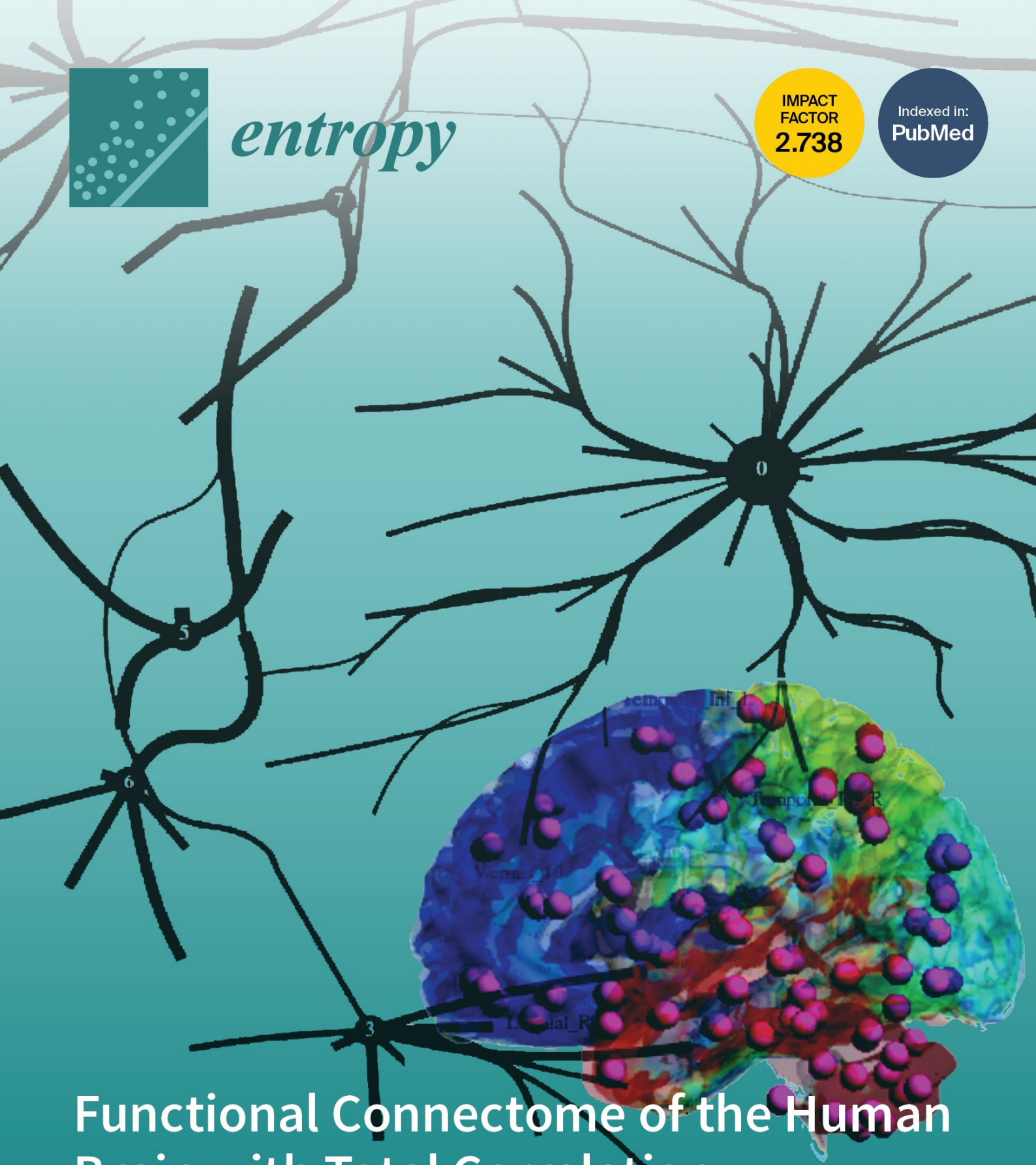
.

Paper II

entropy

IMPACT FACTOR 2.738

Indexed in: PubMed



Functional Connectome of the Human Brain with Total Correlation

Volume 24 · Issue 12 | December 2022

MDPI

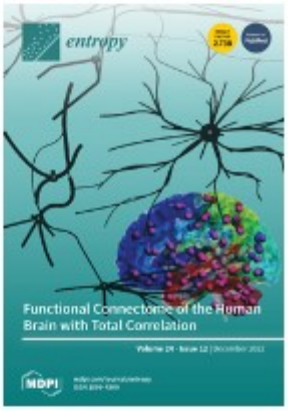
mdpi.com/journal/entropy
ISSN 1099-4300

entropy
Submit to *Entropy*
Review for *Entropy*

Journal Menu

- *Entropy* Home
- Aims & Scope
- Editorial Board
- Reviewer Board
- Topical Advisory Panel
- Video Exhibition
- Instructions for Authors
- Special Issues
- Topics
- Sections & Collections
- Article Processing Charge
- Indexing & Archiving
- Editor's Choice Articles
- Most Cited & Viewed
- Journal Statistics
- Journal History
- Journal Awards
- Society Collaborations
- Conferences
- Editorial Office

Entropy, Volume 24, Issue 12 (December 2022) – 152 articles



Cover Story (view full-size image): The studies used total correlation to describe functional connectivity among brain regions as a multivariate alternative to conventional pairwise measures. In particular, this work uses Correlation Explanation (CorEx) to estimate total correlation. To begin, we demonstrate that CorEx estimates of total correlation and clustering results are reliable when compared to ground truth. Second, the inferred large-scale connectivity network extracted from the more extensive open fMRI datasets is consistent with existing neuroscience studies but, interestingly, can estimate additional relations beyond pairwise regions. Finally, we show how the connectivity graphs based on total correlation can also be an effective tool to aid in the discovery of brain diseases. **View this paper**

- Issues are regarded as officially published after their release is announced to the **table of contents alert mailing list**.
- You may **sign up for e-mail alerts** to receive table of contents of newly released issues.
- PDF is the official format for papers published in both, html and pdf forms. To view the papers in pdf format, click on the "PDF Full-text" link, and use the free **Adobe Reader** to open them.

Order results: Publicatio... | Result details: Normal | Section: All Sections

entropy

MDPI

Article

Functional Connectome of the Human Brain with Total Correlation

Qiang Li [1,*], Greg Ver Steeg [2], Shujian Yu [3] and Jesus Malo [1]

[1] Image Processing Laboratory, University of Valencia, 46980 Valencia, Spain
[2] Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292, USA
[3] Machine Learning Group, UiT—The Arctic University of Norway, 9037 Tromsø, Norway
* Correspondence: qiang.li@uv.es

Abstract: Recent studies proposed the use of Total Correlation to describe functional connectivity among brain regions as a multivariate alternative to conventional pairwise measures such as correlation or mutual information. In this work, we build on this idea to infer a large-scale (whole-brain) connectivity network based on Total Correlation and show the possibility of using this kind of network as biomarkers of brain alterations. In particular, this work uses Correlation Explanation (CorEx) to estimate Total Correlation. First, we prove that CorEx estimates of Total Correlation and clustering results are trustable compared to ground truth values. Second, the inferred large-scale connectivity network extracted from the more extensive open fMRI datasets is consistent with existing neuroscience studies, but, interestingly, can estimate additional relations beyond pairwise regions. And finally, we show how the connectivity graphs based on Total Correlation can also be an effective tool to aid in the discovery of brain diseases.

Keywords: Total Correlation; CorEx; fMRI; functional connectivity; large-scale connectome; biomarkers

Citation: Li, Q.; Steeg, G.V.; Yu, S.; Malo, J. Functional Connectome of the Human Brain with Total Correlation. *Entropy* **2022**, *24*, 1725. https://doi.org/10.3390/e24121725

Academic Editor: Joanna Tyrcha

Received: 12 October 2022
Accepted: 21 November 2022
Published: 25 November 2022

1. Introduction

The human brain is a complex system comprised of interconnected functional units. Millions of neurons in the brain interact with each other at both a structural and functional level to drive efficient inference and processing in the brain. Furthermore, the functional connectivity among these regions also reveals how they interact with each other in specific cognitive tasks. Functional connectivity refers to the statistical dependency of activation patterns between various brain regions that emerges as a result of direct and indirect interactions [1,2]. It is usually measured by how similar neural time series are to each other, and it shows how the time series statistically interact with each other.

A variety of ways to analyze functional connectivity exist. A seedwise analysis can be performed by selecting a seed-driven hypothesis and analyzing its statistical dependencies with all other voxels outside its limits. It is a common tool for studying how different parts of the brain are connected to one another. Connectivity is determined by calculating the correlation between the time series of each voxel in the brain and the time series of a single seed voxel. Another option is to perform a wide analysis of the voxel or Region Of Interest (ROI), where statistical dependencies on all voxels or ROIs are studied [3]. Structural connectivity refers to the anatomical organization of the brain by means of fiber tracts [4]. The sharing of communication between neurons in multiple regions is coordinated dynamically via changes in neural oscillation synchronizations [5]. When it comes to the brain connectome, functional connectivity refers to how different areas of the brain communicate with one another during task-related or resting-state activities [6]. The use of information-theoretic metrics can efficiently detect their interaction in dynamical brain networks, and it is widely used in the field of neuroscience [7], for instance to quantify information encoding and decoding in the neural system [8–11], measure visual information flow in the biological neural networks [12,13] and color information processing

in the neural cortex [14], and so on. However, although functional connectivity has already become a hot research topic in neuroscience [15,16], systematic studies on the information flow or the redundancy and synergy amongst brain regions remain limited. One extreme type of redundancy is full synchronization, where the state of one neural signal may be used to predict the status of any other neural signal, and this concept of redundancy is thus viewed as an extension of the standard notion of correlation to more than two variables [17]. Synergy, on the other hand, is analogous to those statistical correlations that govern the whole, but not its constituent components [18]. High-order brain functions are assumed to require synergies, which give simultaneous local independence and global cohesion, but are less suitable for them under high synchronization situations, such as epileptic seizures [19]. Most functional connectivity approaches until now have mainly concentrated on pairwise relationships between two regions. The conventional approach used to estimate indirect functional connectivity among brain regions is the Pearson Correlation (CC) [20] and Mutual Information (I) [8,21–23]. However, real brain network relationships are often complex, involving more than two regions, and the pairwise dependencies measured by correlation or mutual information cannot reflect these multivariate dependencies. Therefore, recent studies in neuroscience focus on the development of information-theoretic measures that can handle more than two regions simultaneously such as Total Correlation [24,25].

Total Correlation (TC) [26] (also known as multi-information [27–29]) mainly describes the amount of dependence observed in the data and, by definition, can be applied to multiple multivariate variables. Its use to describe functional connectivity in the brain was first proposed as an empirical measure in [24], but in [25], the superiority of TC over mutual information was proven analytically. The consideration of low-level vision models allows deriving analytical expressions for TC as a function of the connectivity. These analytical results show that pairwise I cannot capture the effect of different intra-cortical inhibitory connections, while TC can. Similarly, in analytical models with feedback, synergy can be shown using TC, while it is not so obvious using mutual information [25]. Moreover, these analytical results allow calibrating computational estimators of TC.

In this work, we build on these empirical and theoretical results [24,25] to infer a larger-scale (whole-brain) network based on TC for the first time. As opposed to [24,25], where the number of considered nodes was limited to the range of tens and focused on specialized subsystems, here, we consider wider recordings [30,31], so we use signals coming from hundreds of nodes across the whole brain. Additionally, we apply our analysis to data of the same scale for regular and altered brains (http://fcon_1000.projects.nitrc.org/indi/ACPI/html/ accessed on 12 March 2021). We also show the possibility of using this kind of wide-range networks as biomarkers. From the technical point of view, here, we use Correlation Explanation (CorEx) [32,33] to estimate TC in these high-dimensional scenarios. Furthermore, graph theory and clustering [15,16] are used here to represent the relationships between the considered regions.

The rest of this paper is organized as follows: Section 2 introduces the necessary information-theoretic concepts and explains CorEx. Sections 3 and 4 show two synthetic experiments that prove that the CorEx results are trustable. Section 5 estimates the large-scale connectomes with fMRI datasets that involve more than 100 regions across the whole brain. Moreover, we show how the analysis of these large-scale networks based on TC may indicate brain alterations. Sections 6 and 7 give a general discussion and the conclusion of the paper, respectively.

2. Total Correlation as Neural Connectivity Descriptor

2.1. Definitions and Preliminaries

Mutual Information: Given two multivariate random variables X_1 and X_2, the mutual information between them, $\mathrm{I}(X_1; X_2)$, can be calculated as the difference between the sum

of individual entropies, $H(X_i)$, and the entropy of the variables considered jointly as a single system, $H(X_1, X_2)$ [34]:

$$\mathrm{I}(X_1; X_2) = \mathrm{H}(X_1) + \mathrm{H}(X_2) - \mathrm{H}(X_1, X_2) \quad (1)$$

where, for each (multivariate) random variable v, the entropy is $\mathrm{H(v)} = \langle -\log_2 \mathrm{p(v)} \rangle$ and the brackets represent expectation values spanning random variables. The mutual information also can be seen as the information shared by the two variables or the reduction of uncertainty in one variable given the information about the other [35].

Mutual information is better than linear correlation: For Gaussian sources, mutual information reduces to linear correlation because the entropy factors in Equation (1) just depend on $|\langle X_1 \cdot X_2^\top \rangle|$. However, for more general (non-Gaussian) sources, mutual information cannot be reduced to covariance and cross-covariance matrices. In these (more realistic) situations, I is better than the linear correlation because I captures nonlinear relations that are ruled out by $|\langle X_1 \cdot X_2^\top \rangle|$. For an illustration of the qualitative differences between I and linear correlation, see the examples in Section 2.2 of [24].

As a result, mutual information has been proposed as a good alternative to linear correlation for estimating functional connectivity [8,21]. However, mutual information cannot capture dependencies beyond pairs of nodes. This may be a limitation in complex networks [36].

Total Correlation: This magnitude describes the dependence among n variables, and it is a generalization of the mutual information concept from two parties to n parties. The Venn diagram in Figure 1 qualitatively illustrates this for three variables. The definition of Total Correlation from Watanabe [26] can be denoted as:

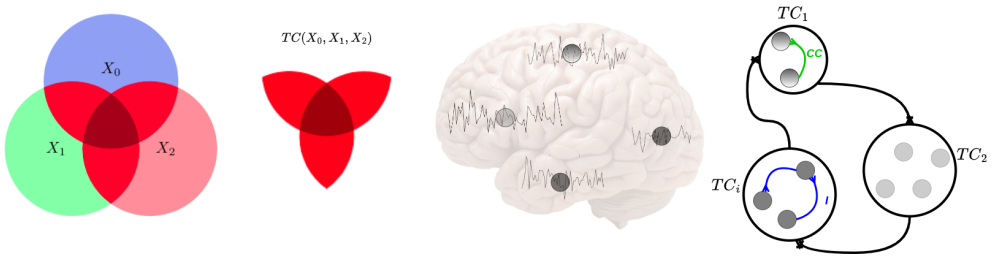


Figure 1. Conceptual scheme of information-theoretic measures of neural information flow. The left circle areas represent the amounts of information, and intersections represent shared information among the corresponding variables, X_0, X_1, X_2. Examples of entropy, $H(X_0), H(X_1), H(X_2)$, Total Correlation (red color), and $TC[X_0, X_1, X_2]$ are given. The middle figures show some neural time series extracted from brain regions, which correspond to the nodes in the right figure. The right figures illustrate large-scale time series in the brain and how the coupled information is transmitted among the brain regions. The blue and green lines show Linear Correlation (CC) and Mutual Information (I), respectively, between different parts of the brain. The modules represent the lobes of the human brain. Each module has specific brain regions, and each module works with the others.

$$\mathrm{TC(X_1, \ldots, X_n)} \equiv \sum_{i=1}^{n} \mathrm{H(X_i)} - \mathrm{H(X_1, \ldots, X_n)} = \mathrm{D_{KL}} \left(\mathrm{p(X_1, \ldots, X_n)} \| \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{p(X_i)} \right) \quad (2)$$

where $\mathrm{X} \equiv (\mathrm{X_1}, \ldots, \mathrm{X_n})$ and TC can also be expressed as the Kullback–Leibler divergence, D_{KL}, between the joint probability density and the product of the marginal densities. From these definitions, if all variables are independent, then TC will be zero.

For the conditional Total Correlation, which is similar to the definition of Total Correlation, but with a condition appended to each term, the Kullback–Leibler divergence of the two conditional probability distributions can also be used to define the conditional Total Correlation. The estimation method used in this work (CorEx presented in the next

subsection) uses TC after conditioning on some other variable Y, which can be defined as [34]:

$$TC(X|Y) = \sum_i H(X_i|Y) - H(X|Y) = D_{KL}(p(\mathrm{x}|\mathrm{y})\|\prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}(\mathrm{x_i}|\mathrm{y})) \tag{3}$$

Total Correlation is better than mutual information: This superiority is not only due to the obvious n-wise versus pairwise definitions in Equations (1) and (2). It also has to do with the different properties of these magnitudes. To illustrate this point, let us recall one of the analytical examples in [25]. Consider the following feedforward network:

$$X_1 \longrightarrow X_2 \longrightarrow \mathbf{e} \xrightarrow{f} X_3 \tag{4}$$

where the nodes X_1, X_2, e, and X_3 can have any number of neurons, the first two transforms, $X_1 \longrightarrow X_2 \longrightarrow \mathbf{e}$, are linear and affected by additive noise, and the last transform, $f(\cdot)$, is nonlinear, but deterministic. Imagine that, in this network, one is interested in the connectivity between the neurons in the hidden layer, $\mathbf{e}$; however, the nonlinear function $f(\cdot)$ is unknown, and one only has experimental access to the signal in the regions X_1, X_2, and X_3. In this situation, one could think of measuring $I(X_1, X_3) = I(X_1, f(\mathbf{e}))$ or $I(X_2, X_3) = I(X_1, f(\mathbf{e}))$. However, the invariance of I under arbitrary nonlinear reparametrization of the variables [35] implies that these measures are insensitive to f and the connectivity therein. On the contrary, as pointed out in [25], using the expression for the variation of TC under nonlinear transforms [13,37], the variation of H under nonlinear transforms [34], and the definition in Equation (2), one obtains $TC(X_1, X_2, X_3) = [TC(X_1, X_2, \mathbf{e}) - TC(\mathbf{e})] + TC(X_3)$, where the term in the bracket does not depend on $f(\cdot)$, but the last term definitely does, which proves the superiority of TC over I in describing connectivity.

In [25], the network in Equation (4) specifically refers to the flow from the retina, X_1, to the LGN, X_2, and finally, to the visual cortex, $\boldsymbol{e}$ and X_3. However, the result of the superiority of TC over I to describe the connectivity in the hidden layer is totally general for every network with the generic properties listed after Equation (4).

2.2. Total Correlation Estimated from CorEx

Straightforward application of the direct definition of TC is not feasible in high-dimensional scenarios, and alternatives are required [28,29]. A practical approach to estimate Total Correlation is via *latent factor modeling*. A latent factor model is a statistical model that relates a set of observable variables to a set of latent variables. The idea is to explicitly construct latent factors, Y, that somehow capture the dependencies in the data. If we measure dependencies via Total Correlation, $TC(X)$, then we say that the latent factors *explain* the dependencies if $TC(X|Y) = 0$. We can measure the extent to which Y explains the correlations in X by looking at how much Total Correlation is reduced:

$$TC(X) - TC(X|Y) = \sum_{i=1}^{n} I(X_i; Y) - I(X; Y) \tag{5}$$

Total Correlation is always non-negative, and the decomposition on the right in terms of mutual information can be verified directly from the definitions. Any latent factor model can be used to lower-bound Total Correlation, and the terms on the right-hand side of Equation (5) can be further lower-bounded with tractable estimators using variational methods; Variational Autoencoders (VAEs) are a popular example [38].

Although latent factor models do not give a direct Total Correlation estimation as the Rotation-based Iterative Gaussianization (RBIG) [28,29] and the matrix-based Rényi entropy [39] did, the approach can be complementary because the construction of latent factors can help in dealing with the curse of dimensionality and for interpreting the dependencies in the data. Compared to CorEx, the main goal of (RBIG https://isp.uv.es/RBIG4IT.htm (ac-

cessed on 12 October 2022)) is to convert any non-Gaussian-distributed data into a Gaussian distribution through marginal Gaussianization and rotation to obtain TC. The matrix-based Rényi entropy (http://www.cnel.ufl.edu/people/people.php?name=shujian (accessed on 12 October 2022)) is mainly used for estimating multivariate information based on Shannon's entropy, which is Rényi's α-order entropy [40]. With these goals in mind, we now describe a particular latent factor approach known as Total Correlation Explanation (CorEx (https://github.com/gregversteeg/CorEx) (accessed on 12 October 2022)) [32].

CorEx constructs a factor model by reconstructing latent factors using a factorized probabilistic function of the input data, $p(y|x) = \prod_{j=1}^{m} p(y_j|x)$, with m discrete latent factors, Y_j. This function is optimized to give the tightest lower bound possible for Equation (5).

$$TC(X) \geq \max_{p(Y_j|x)} \sum_{i=1}^{n} I(X_i;Y) - I(X;Y) = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} \alpha_{i,j} I(X_i;Y_j) - I(Y_j;X) \right) \tag{6}$$

The factorization of the latent factors leads to the terms $I(X;Y) = \sum_j I(Y_j;X)$, which can be directly calculated. The term $I(X_i;Y)$ is still intractable and is decomposed using the chain rule into $I(X_i;Y) \approx \sum \alpha_{i,j} I(X_i;Y_j)$. Each $I(X_i;Y_j)$ can then be tractably estimated [32,33]. There are free parameters $\alpha_{i,j}$ that must be updated while searching for latent factors and achieving objective functions. When $t = 0$, the $\alpha_{i,j}$ initializes and then updates according to:

$$\alpha_{i,j}^{t+1} = (1-\lambda)\alpha_{i,j}^{t} + \lambda \alpha_{i,j}^{**} \tag{7}$$

The second term $\alpha_{i,j}^{**} = \exp\left(\gamma\left(I\left(X_i:Y_j\right) - \max_j I\left(X_i:Y_j\right)\right)\right)$, and λ and γ are constant parameters. This decomposition allows us to quantify the contribution to the Total Correlation bound from each latent factor, which can aid interpretability.

CorEx can be further extended into a hierarchy of latent factors [33], helping to reveal the hierarchical structure that we expect to play an important role in the brain. The latent factors at layer k explain the dependence of the variables in the layer below.

$$TC(X) \geq \sum_{k=1}^{r} \left(\sum_{j=1}^{m} \left(\sum_{i=1}^{n} \alpha_{i,j}^{k} I(Y_i^{k-1};Y_j^{k}) - \sum_{j=1}^{m} I(Y_j^{k};Y^{k-1}) \right) \right) \tag{8}$$

Here, k gives the layer and $Y^0 \equiv X$ denotes the observed variables. Ultimately, we have a bound on TC that becomes tighter as we add more latent factors and layers and for which we can quantify the contribution for each factor to the bound. We exploit this decomposition for interpretability [41], as illustrated in Figure 2. CorEx prefers to find modular or tree-like latent factor models, which are beneficial for dealing with the curse of dimensionality [42]. For neuroimaging, we expect this modular decomposition to be effective because functional specialization in the brain is often associated with spatially localized regions. We explore this hypothesis in the experiments.

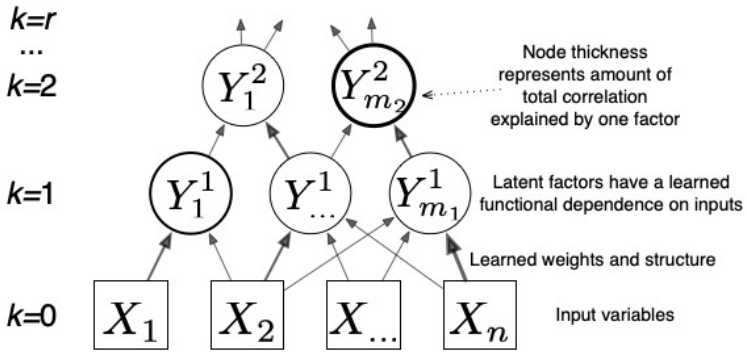


Figure 2. CorEx learns a hierarchical latent factor as illustrated above. Edge thickness indicates strength of the relationship between factors, and node thickness indicates how much Total Correlation is explained by each latent factor.

3. Experiment 1: Total Correlation for Independent Mixtures

In this experiment, we estimated the Total Correlation of three independent variables X, Y, and Z, and each follows a Gaussian distribution. For this setup, the ground truth of TC should satisfy $TC(X, Y, Z) = 0$, and we generated various samples with different lengths. Then, the estimated Total Correlation values are shown in Figure 3. Here, we compared CorEx with other different Total Correlation estimators, such as RBIG [28,29], matrix-based Rényi entropy [39], Shannon discrete entropy (https://github.com/nmtimme/Neuroscience-Information-Theory-Toolbox accessed on 12 October 2022), and the ground truth. The left figure (2-dimensional) is mutual information, and the middle (3-dimensional) and right figure (4-dimensional) are Total Correlation. As we mentioned above, the simulation data are totally Gaussian-distributed. Therefore, their dependency should be zero. We find that CorEx and RBIG both perform very well and are very stable, and matrix-based Rényi entropy's performance becomes more and more nice with increased dimensions, while Shannon discrete entropy becomes more and more accurate with an increase of the samples. All these make sense, and it also explains the accuracy of Total Correlation estimation with CorEx. Here, compared to other estimators, the main functionality goal of CorEx is to cluster statistical dependency variables based on Total Correlation. However, other estimators mainly focus on directly obtaining the Total Correlation value and do not supply very nice visualization results. The CorEx gives us a nice connection with graph theory to visualize and show their functional relationship.

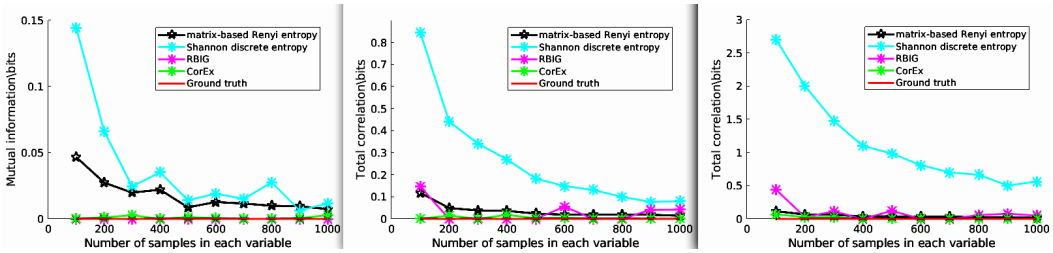


Figure 3. The estimated Total Correlation values for three independent variables. The various Total Correlation estimators are compared with the ground truth value (red line), for example matrix-based Rényi entropy (black line), Shannon discrete entropy (cyan line), RBIG (magenta line), and CorEx (green line). See the main text for more information.

4. Experiment 2: Clustering by Total Correlation for Dependent and Independent Mixtures

To evaluate the performance of CorEx in clustering tasks. The elements in group X include $X1$, $X2$, and $X3$, which satisfy Gaussian distributions and are completely independent of each other and of group Y, and the variables in group Y include $Y1$, $Y2$ from $Y1$,

and $Y3$ from $Y2$, which are connected to each other. Then, we compared the CorEx cluster results with the pairwise Pearson correlation, pairwise mutual information, and partial correlation, which consider confounding effects to find the groups.

In Figure 4, we find that CorEx based on Total Correlation has high accuracy in estimating their dependencies (Figure 4e) compared to pairwise Pearson correlation (Figure 4b), pairwise mutual information (Figure 4c), and partial correlation (Figure 4d). As we established in this experiment, the elements in group Y should be clustered together, and the elements in group X should be completely independent of each other and of group Y. The ground truth is presented in Figure 4a. Then, we estimated the cluster result with the pairwise Pearson correlation with a threshold of 0.1, pairwise mutual information with a threshold of 0.4, and partial correlation without a threshold. Obviously, we found that pairwise approaches have high errors in accurately estimating their statistical dependencies, and pairwise mutual information is better than pairwise Pearson correlation, but still has high errors in correctly clustering tasks. When we considered the confounding effect of the third variables, we still did not obtain a better clustering result compared to TC. Therefore, the clustering results with CorEx by Total Correlation obtain the best performance compared to pairwise approaches. Moreover, we used *purity* as a criterion of clustering quality to qualify the performance of clustering because it is a straightforward and transparent evaluation metric [43]. To calculate purity, each cluster is allocated to the class that occurs most frequently within it, and the accuracy of this assignment is determined by counting the number of correctly assigned elements and dividing by $N(N = 6)$. Formally:

$$\text{Purity}(X, Y) = \frac{1}{N} \sum_i \max_j |X_i \cap Y_j| \tag{9}$$

where $X = \{X1, X2, X3\}$ is the set of clusters and $Y = \{Y1, Y2, Y3\}$ is the set of classes. Figure 4f presents the clustering performance of pairwise approaches and CorEx with purity as a criterion. Poor clusters have near-zero purity ratings (lower bound). A perfect cluster possesses a purity of one (maximum value). Based on Equation (9), we obtain purity values of 0.17 and 0.33 for pairwise approaches and partial correlation, and the purity value for CorEx is 0.83. All in all, we show that CorEx based on Total Correlation has the best performance compared to pairwise approaches.

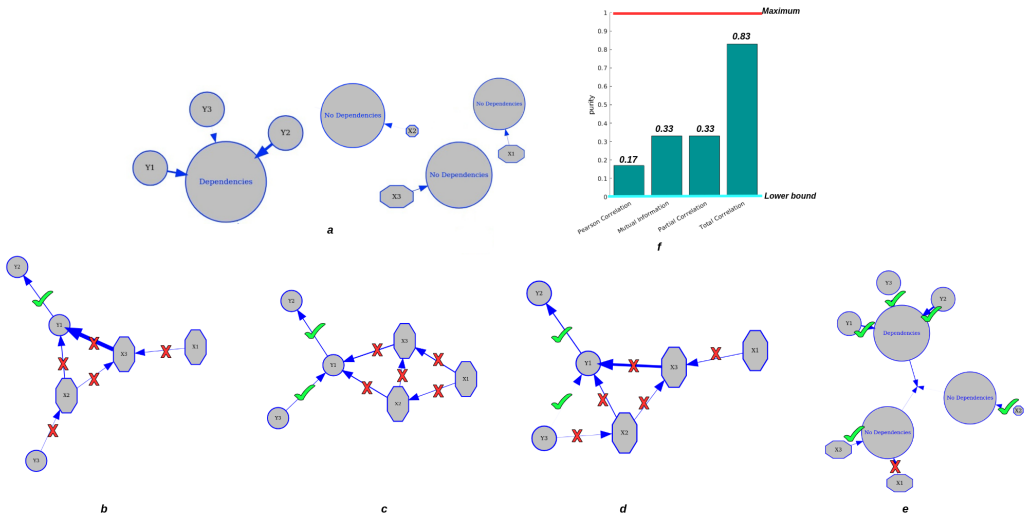


Figure 4. Clustering performance for dependent and independent mixtures. The top row: (**a**) displays the ground truth of variable clustering in two groups. (**f**) shows the purity value of each approach. The second row: (**b**) shows the clustering result based on Pearson correlation. (**c**) shows the clustering result by pairwise mutual information. (**d**) shows the clustering result by partial correlation. (**e**) shows clustering results by CorEx based on Total Correlation.

5. Experiment 3: Brain Functional Connectivity Analysis Using Total Correlation

A network is a collection of nodes and edges, where nodes represent fundamental elements (e.g., brain regions) within the system of interest (e.g., the brain) and edges represent the dependencies that exist between those fundamental elements with the considered weights. Typically, the threshold is chosen based on the visual effect on functional connectivity, and here, we set the optimal threshold for community detection in brain connectivity networks. We used it to identify a threshold that maximizes information on the network modular structure, removes the weakest edges, and keeps the largest connected component. Figure 5 illustrates the schematic representation of network construction using fMRI. Firstly, the time series were extracted from fMRI data based on a selected structural atlas, and then, functional connectivity was estimated with CC, I, and CorEx, respectively. The results are presented with a graph that includes both brain nodes and their functional connectivity with weight edges.

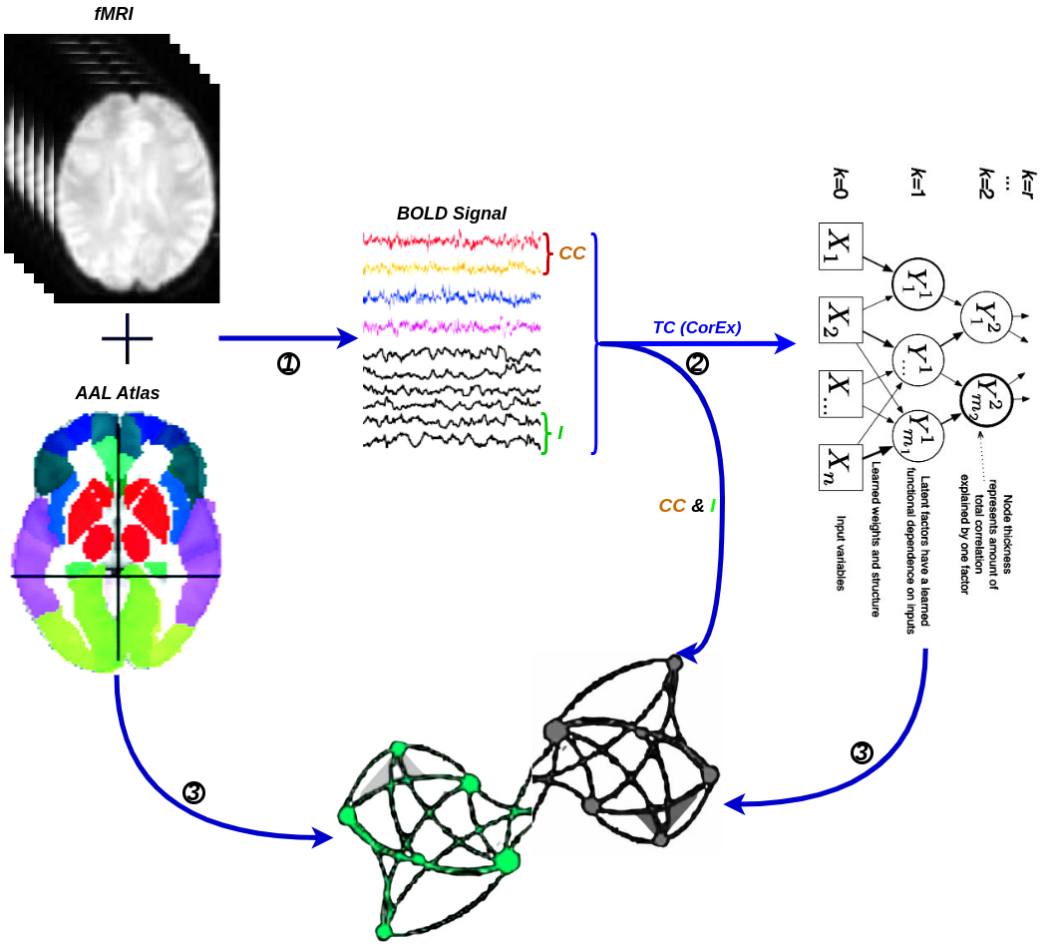


Figure 5. A flowchart for the construction of a functional brain network by fMRI. ① Time series extraction from fMRI data within each anatomical unit (i.e., network node). ② Estimation of functional connectivity with CC, I, and TC (CorEx), respectively. ③ Visualization of functional connectivity as tree and circle graphs (i.e., network edges and network nodes).

5.1. First Total-Correlation-Based Clustering Example from fMRI Data

The data were taken from a resting-state fMRI experiment in which a subject was watching and maintaining alert wakefulness, but not performing any other behavioral task. Meanwhile, the BOLD signal was recorded. These data were downloaded from Nitime (https://nipy.org/nitime/index.html accessed on 12 October 2022). The data were preprocessed, and time series were extracted from different Regions Of Interest (ROIs)

in the brain. The ROIs' abbreviations and related full names are listed as follows: Cau, Caudate; Pau, Paudate; Thal, Thalamus; Fpol, Frontal pole; Ang, Angular gyrus; SupraM, Supramarginal gyrus; MTG, Middle Temporal Gyrus; Hip, Hippocampus; PostPHG, Posterior Parahippocamapl Gyrus; APHG, Anterior Parahippocamapl Gyrus; Amy, Amygdala; ParaCing, Paracingulate gyrus; PCC, Posterior Cingulate Cortex; Prec, Precuneus; R, Right hemisphere; L, Left hemisphere. First, we estimated the pairwise functional connectivity metrics with Pearson correlation, mutual information, and the corresponding functional connectivity, a circle-weighted graph used to visualize the outcome of pairwise functional connectivity. In Figure 6, top row (left and right), Pearson correlation and mutual information estimate the same pairwise dependencies, but later approaches capture stronger weights between ROIs, such as LPCC and RPCC, LThal and RThal, and LAmy and RAmy.

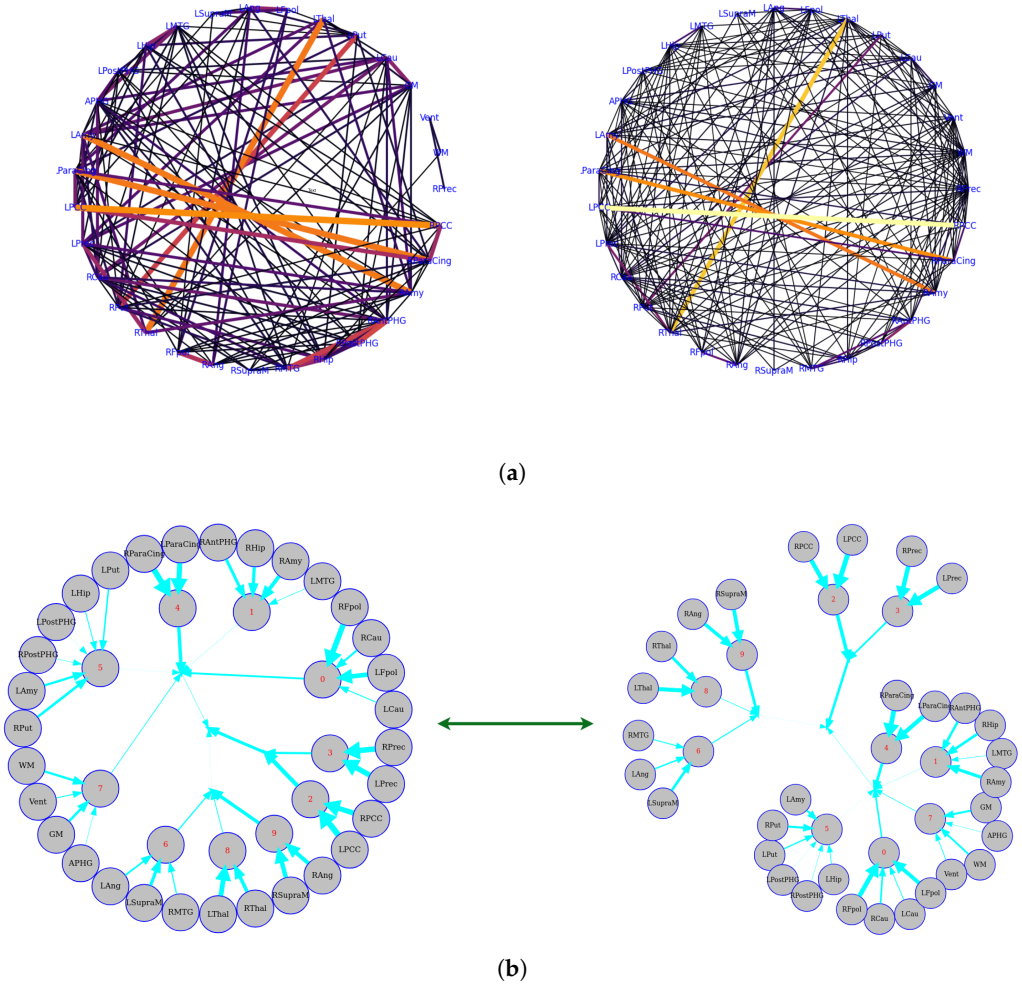


Figure 6. Functional connectivity representation with graph-based networks. The functional connectivity is represented in the cycle (**a**) and tree (**b**) graphs. Top row: the left and right figures correspond to Pearson correlation with a threshold of 0.14 and mutual information with a threshold of 0.02, respectively. Bottom row: the figures show the Total Correlation with a threshold of 0.16 that was estimated by CorEx. To more directly display the statistical dependencies of brain regions, we here converted the circle graph to a tree graph. The weights are shown by the thickness of the edges, which shows how strongly information is coupled between or among brain regions.

Meanwhile, we also used weighted graph theory to cluster dependence among ROIs, and we thresholded edges with a weight of less than 0.16 for legibility with the CorEx approach. As we mentioned above, mutual information only estimates a more robust relationship between ROIs compared to correlation. However, when we go beyond pairwise

ROIs, CorEx captures richer information among all ROIs (see Figure 6 (bottom row)). Here, we selected $m_1 = 10$, $m_2 = 3$, $m_3 = 1$ as the latent dimension for each layer in our estimate of TC with CorEx, and their corresponding convergent curves are plotted in Figure 7; it shows the Total Correlation lower bound stops increasing. Figure 6 (bottom row) shows the overall structure of the learned hierarchical model. Edge thickness is determined by $\alpha_{i,j} I\left(X_i : Y_j\right)$. The size of each node is proportional to the Total Correlation that a latent factor explains about its children. The discovered structure captures several significant relationships among ROIs that are consistent with correlation and mutual information results, e.g., LPCC and RPCC, LThal and RThal, LParaCing and RParaCing, and LPut and RPut. Furthermore, TC discovered some beyond pairwise unknown relationships; for example, LCau, RCau, LFpol, and RFpol are clustered under Node 0, which explains why they have dense dependency during this cognitive task compared to other ROIs in the brain.

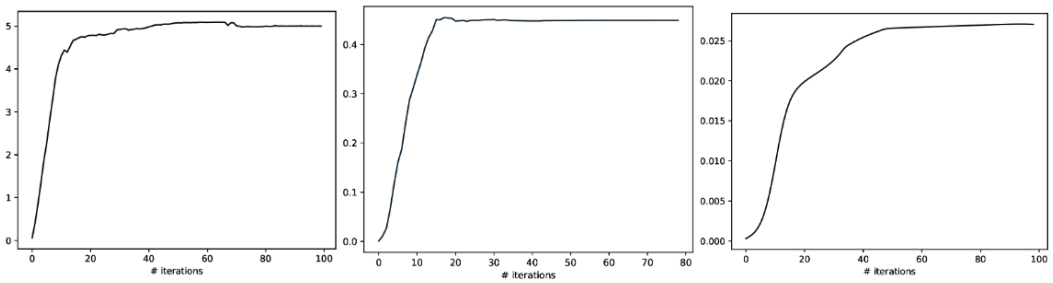


Figure 7. The Total Correlation convergence curve of CorEx in Layers 1, 2, and 3 is shown above. From left to right, their corresponding Layer 1, Layer2, and Layer3 parameters are selected in event-related experiments, and it shows that the Total Correlation lower bound stops increasing and tends to converge.

5.2. Large-Scale Connectome with Resting-State fMRI

5.2.1. A Selection of Pre-Defined Atlas

We used the Automated Anatomical Labeling (AAL) atlas [44], a structural atlas with 116 ROIs identified from the anatomy of a reference subject (see Figure 8).

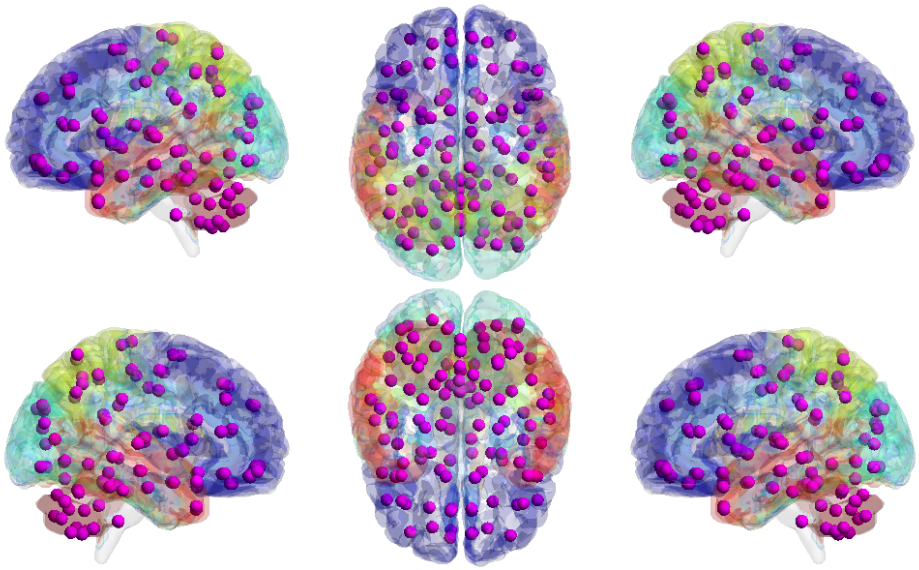


Figure 8. Automated Anatomical Labeling (AAL) atlas. The graph shows the volume of AAL (116 regions) mapped to the smoothed Colin27 brain surface template. The different brain areas are labeled on the brain surface with different colors, and detailed ROI/purple node information can be found in the Appendix A with Table A1.

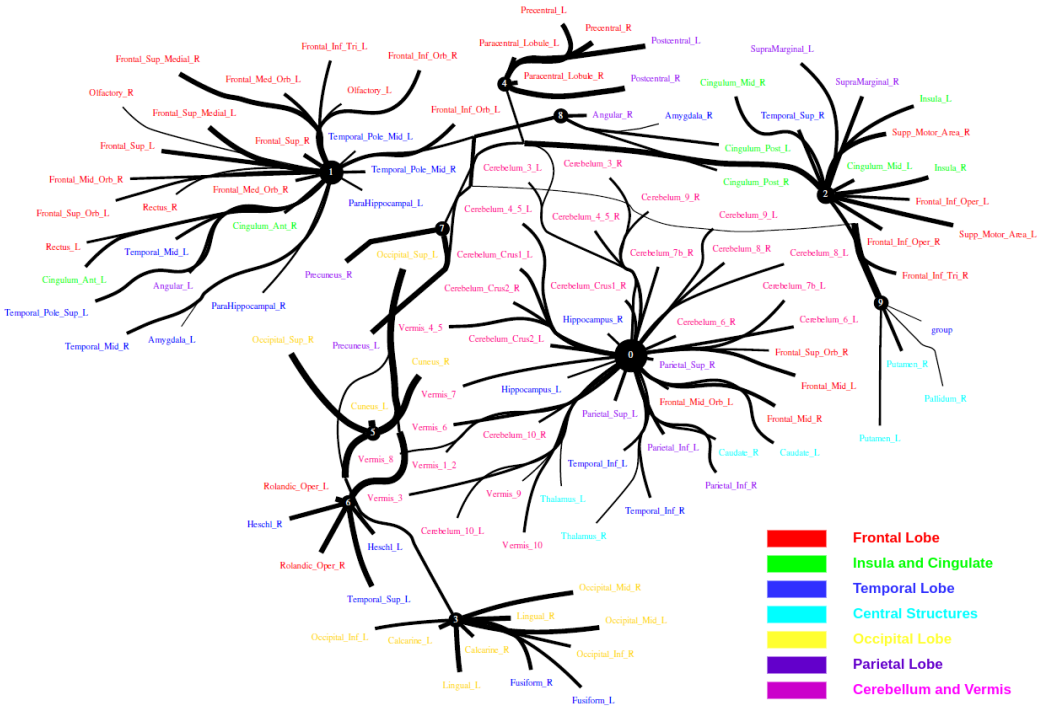
5.2.2. Time Series Signals Extraction

The HCP and ACPI can access raw and preprocessed data, as well as phenotypic information about data samples. The raw rs-fMRI data were preprocessed using the Configurable Pipeline for the Analysis of Connectomes, an open-source software pipeline that allows for automated rs-fMRI data preprocessing and analysis. We extracted time series for each ROI in each subject after defining anatomical brain ROIs with the AAL atlas. We calculated the weighted average of the fMRI BOLD signals across all voxels in each region. Furthermore, the BOLD signal in each region was normalized and subsampled by the repetition time. Finally, we averaged all of the subjects' time series signals in each ROI.

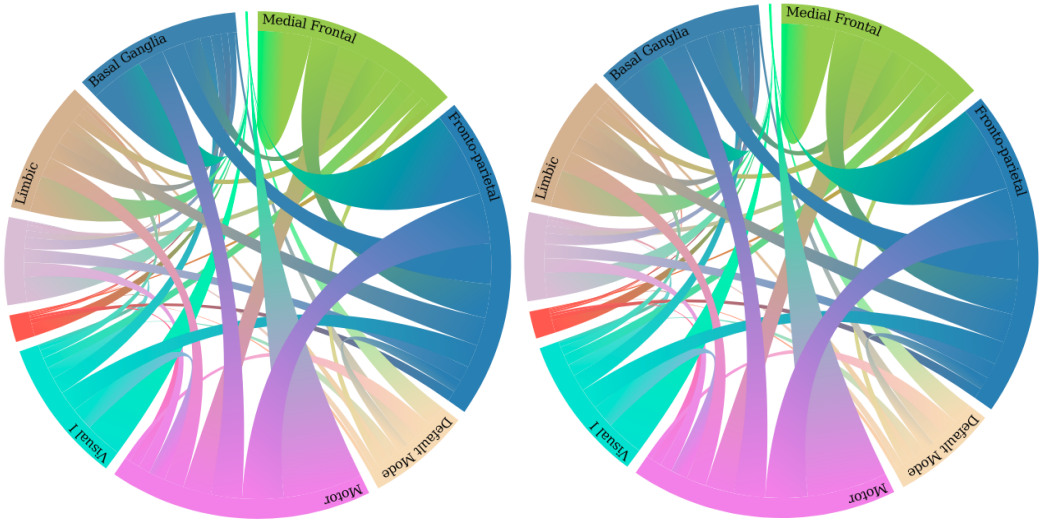
5.2.3. HCP900

The Human Connectome Project contains imaging and behavioral data from healthy people [30]. To investigate resting-state functional connectivity, we used preprocessed rest-fMRI data from the HCP900 (https://www.humanconnectome.org/ (accessed on 12 March 2021)) release [31]. Here, we selected $m_1 = 10$, $m_2 = 5$, $m_3 = 1$ as the latent dimension for each layer in our estimate of TC with CorEx. We thresholded edges with a weight of less than 0.16 for legibility. Figure 9 shows that whole-brain resting-state functional connectivity is estimated with CorEx compared to Pearson correlation and mutual information. It mostly captures relationships among brain regions, and neighboring brain regions cluster together and communicate with other areas, e.g., Node 0 has a bigger node size than other nodes.

From Figure 9, we found that brain regions are functionally clustered together, which is also consistent with structure connectivity based on their physical connectivity distance. For example, under Node 0, the cerebellum and vermis regions densely cluster together, while under Node 1, the frontal lobes cluster together and are also densely functionally connected with the temporal lobe, and so on. The different colors indicate different brain regions, which are based on Table A1. In addition, we can see that functional integration and separation exist in our brain from Figure 9.



(**a**)



(**b**) (**c**)

Figure 9. Large-scale functional connectivity with the HCP900. The functional connectivity is represented in the tree (**a**) and cycle (**b**,**c**) graphs. Top row: A weighted threshold graph with a max of 86 edges showing the overall structure of the representation learned from AAL ROIs (a high-resolution figure is represented in the appendix with Figure 10). Edge thickness is proportional to mutual information, and node size represents Total Correlation among children. In the node with red color, the frontal lobe is represented, while green color represents the insula and cingulate regions, blue color the temporal lobe, cyan color the central areas, gold color the occipital lobe, purple color the parietal lobe, and deep pink color the cerebellum and vermis. Bottom row: Two representative connectomes are presented in the form of a circular chord that shows the connections of all 116 nodes with (**b**) correlation and (**c**) mutual information of the HCP dataset. Each lobe was labeled with a different color.

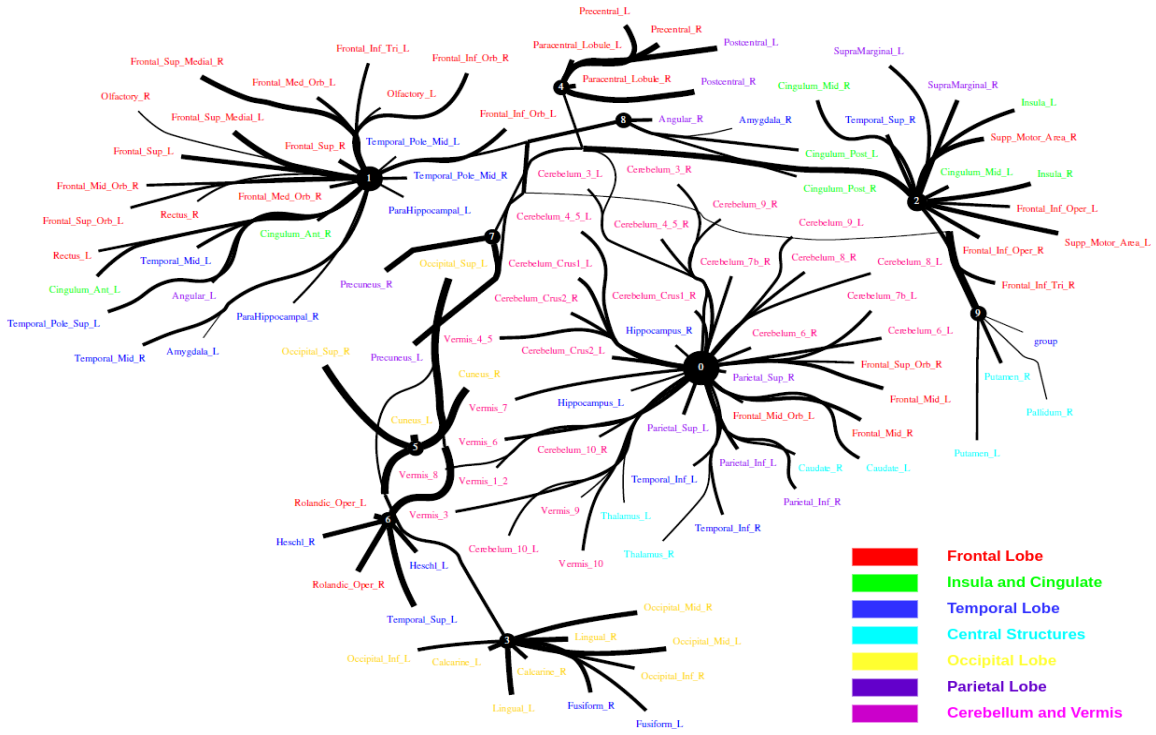
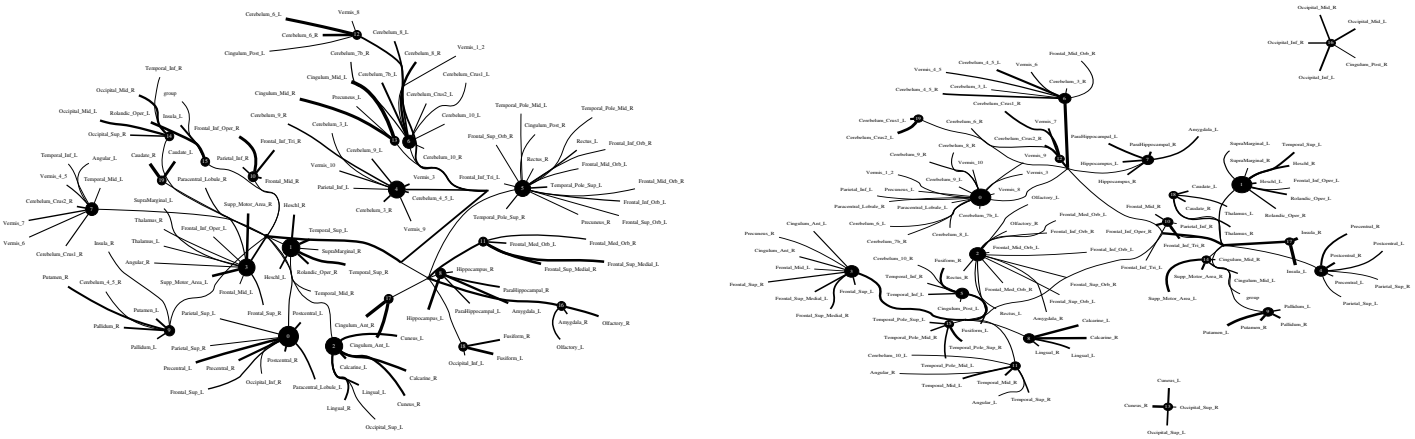


Figure 10. Functional connectivity of HCP900.

5.2.4. Computational Psychiatry Applications with ACPI

The Addiction Connectome Preprocessed Initiative is a longitudinal study to investigate the effects of cannabis use among adults with a childhood diagnosis of ADHD. In particular, we used readily preprocessed rest-fMRI data from the Multimodal Treatment Study of Attention Deficit Hyperactivity Disorder (MTA).We attempted to use functional connectivity as a bio-marker to discriminate whether individuals have consumed marijuana or not (62 in the marijuana group vs 64 in the control group). In a comparison of whole-brain functional connectivity between the control and patient groups, we found altered functional connectivity in the patient group compared to the healthy group (see Figure 11). We quantified the difference between the patient group and the healthy group, and the purity of the patient group compared to the control group was 0.85 ± 0.23. The significant altered functional connectivity happened between the frontoparietal and motor regions. Meanwhile, we found sparse functional connectivity in the patient group compared to the control group in general. Meanwhile, we also discovered that marijuana users had more interaction between neural time series in particular ROIs such as the cerebellum, frontoparietal, and default model regions than controls, e.g., cerebellum regions mainly densely cluster around Node 0 compared to the control group. It also may explain differences in behavior in marijuana users because the frontoparietal network controls cognitive behavior execution and decision-making, cerebellum-related action, and default model network dysfunction in addicted users. All the above results are consistent with previous related research [45–47]. Moreover, we found some unknown disconnect between some visual regions and other brain areas. Based on related research [48,49], we suggest that marijuana patients may have altered visual perception as well.



(**a**) (**b**)

Figure 11. Functional connectivity between healthy group and patient group. A weighted threshold graph showing the overall structure of the representation learned from ALL ROIs. Edge thickness is proportional to mutual information, and node size represents Total Correlation among children. Here, we selected $m_1 = 20$, $m_2 = 3$, $m_3 = 1$ as the latent dimension for each layer in our estimate of TC with CorEx. (**a**) refers to normal people's functional connectivity, and (**b**) shows the marijuana group's functional connectivity in the brain. Both groups were measured with a TC that used the same parameters in the model. In comparison with the healthy group, we found less functional connectivity happened in the patient group, e.g., frontoparietal lobe and default model regions. (A high-resolution figure is represented in the appendix with Figures 12 and 13).

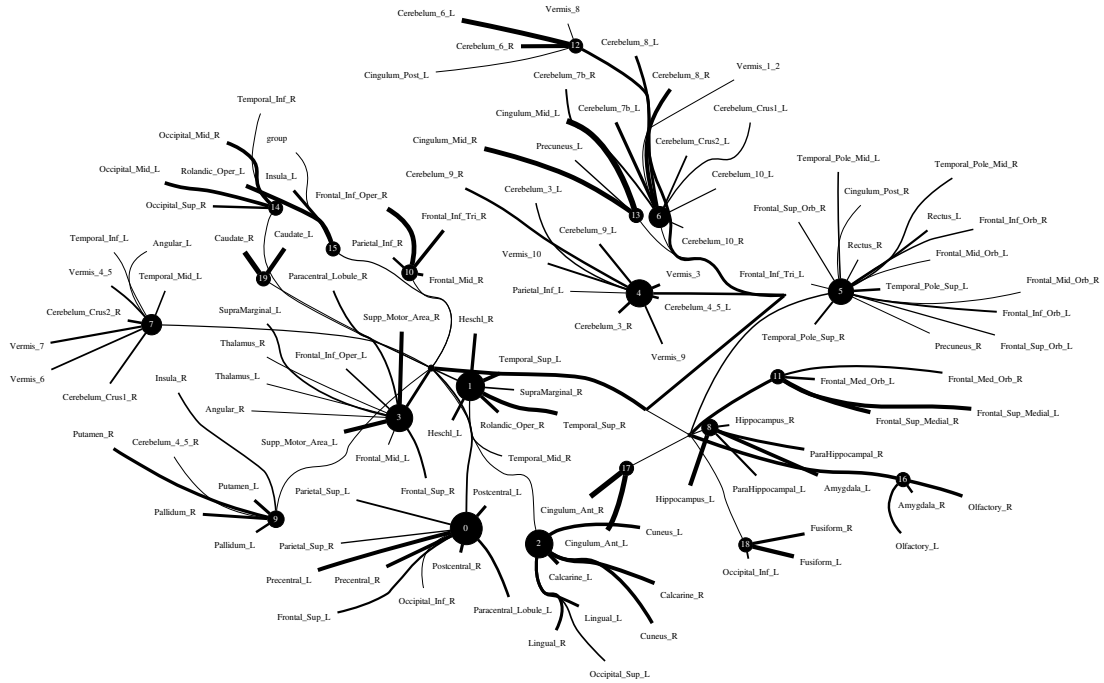


Figure 12. Functional connectivity of healthy group.

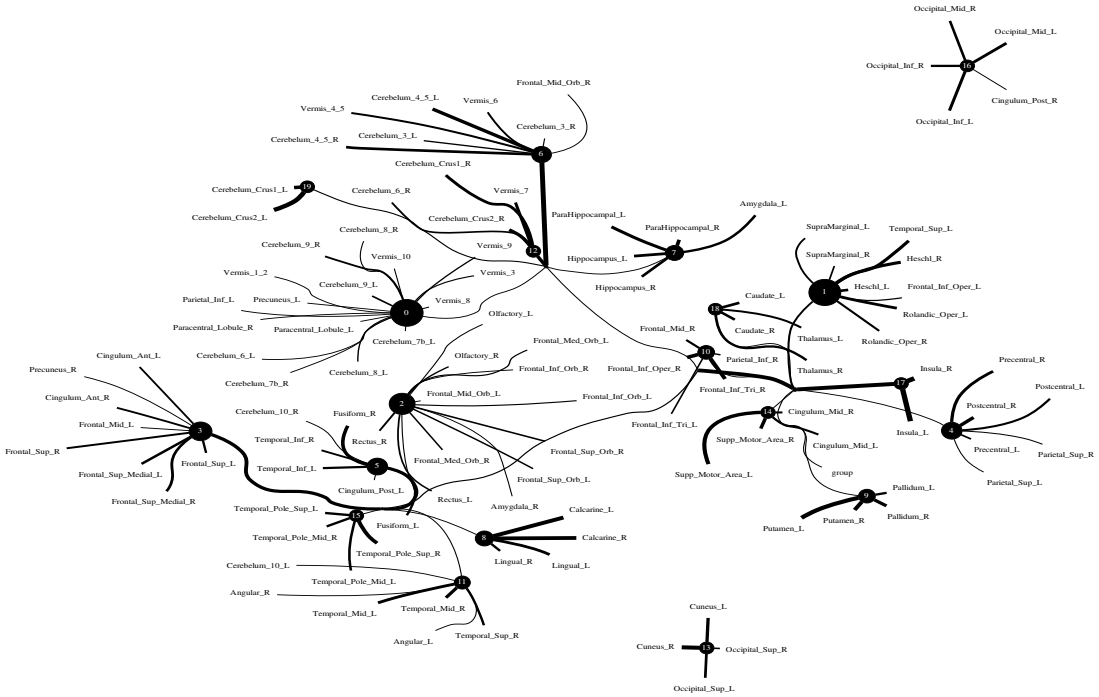


Figure 13. Functional connectivity of patient group.

6. Discussion

This manuscript presents a higher-order information-theoretic measure to estimate functional connectivity. We estimated Total Correlation with CorEx under different situations. However, the approach has its own pros and cons, which we will discuss later. Furthermore, we found that Total Correlation can be a metric to estimate functional connectivity in the human brain. It can identify some well-known functional connectivities and capture a few unknown nonlinear relationships among brain regions as well. To the best of our knowledge, this is the first time that Total Correlation has been used to estimate larger-scale functional connectivity for a whole-brain AAL atlas with 116 structural ROIs. Total Correlation can also be a tool to find biomarkers to help us diagnose brain-related diseases.

Here, we discuss some advantages and limitations of this research now. Firstly, given the curse of dimensionality of fMRI, we need to find a low-dimensional representation that helps us characterize the connectivity. Traditional General Linear Models (GLMs), such as expert-defined ROIs or the ALL atlas, are frequently used to find ROIs in resting-state experiments. However, we should be able to do better with a data-driven approach. Sample sizes and statistical thresholds are known to have a major impact on the statistical power and accuracy of GLM-based ROI selection. Previous research has revealed that the GLM has limited statistical power when inferring from fMRI data [50,51]. However, we used GLM-based ROI selection in the real fMRI datasets, which may affect the final result when we estimate functional connectivity.

Second, CorEx is model-independent, which means no anatomical or functional prior knowledge is required to estimate the ROIs. The method is entirely data-driven; this way, it is possible to analyze networks that have not been investigated and could be a future extension of work. It is also possible to use Total Correlation as a pre-analysis for other techniques such as dynamic causal modeling, which need constraints about the underlying network [52]. What differentiates the CorEx algorithm is that it tries to break the variables into clusters with high TC. In other words, CoRex finds a tree of latent factors that explain Total Correlation, so this tree of clusters based on TC is a more data-driven way to define regions and then connectivity than ROIs predefined by hand. This prioritization of "modular" solutions in CorEx was not realized or emphasized in the original research. The second reason why we used CorEx to estimate functional connectivity on larger-scale fMRI datasets is that it is a clustering approach via TC. Furthermore, CorEx estimates Total Correlation via hierarchical maximization correlation between previous layer and current layer variables with a tight information bound that estimates a more accurate relationship among variables in real neural signals.

Third, TC is an indirect information quantitative tool that cannot determine the direction of information flow between brain regions. Meanwhile, we discovered some unknown functional connectivity in the real fMRI dataset before.

Fourth, given the irregularity of neural time series and the difficulties in quantifying graph signals when brain networks are represented by graphs, we should avoid quantifying too many graph signals. However, there is a metric called permutation entropy that gives us the possibility to quantify the graph signal in complex systems [36]. It could be very interesting to apply this metric to brain networks to check how much information could be obtained from the complex graph signals, which could then help us more deeply understand brain networks in the future. Moreover, as we mentioned the complexity of neural time series, one of the important potential problems is the length of time series, except for the additional dimensional problem. It is a significant challenge when you are processing long lengths of time series, but it could be solved by transforming the time series into embedding space or segmenting the long time series into specific time windows [53].

Finally, we applied TC to estimate large-scale functional connectivity with the real fMRI dataset across the HCP and ACPI. The functional connectivity with the HCP900 gives us the potential to estimate a full brain atlas with TC in the future, and our result shows that TC can capture the right functional connectivity; beyond this, it could also give us some unknown functional connectivity. Therefore, it could be a future extension

project. Furthermore, we used TC as a possible method to find biomarkers of brain disease with the ACPI dataset. We compared whole-brain functional connectivity between control and patient groups. We found altered functional connectivity in the patient group compared to the healthy group, and we quantified this difference with purity metrics because it is a simple and transparent evaluation measure. The purity in the patient group compared to the control group is not too large, and it shows that there is some altered functional connectivity in the patient group; for instance, we mentioned brain networks in the cerebellum, frontoparietal, and default model regions. However, it was just examined with one dataset with a small number of subjects and does not consider within-subject variability, and it could be extended with more large datasets in the future.

7. Conclusions

We introduced Total Correlation to capture multivariate large-scale interactions within brain regions. They were experimentally verified as effective steps for reconstructing multivariate relationships in the brain. In this study, CorEx was adopted to estimate Total Correlation. The CorEx approach can capture functional connectivity characteristics when going beyond pairwise brain regions. On the other hand, we evaluated the method with resting-state fMRI datasets. We found that multivariable relationships cannot be detected if we use pairwise correlation and mutual information quantities only. More generally, multivariable relationships can be clustered only if we use Total Correlation. Therefore, Total Correlation measures are significant to find complicated functional connectivity among brain regions. Furthermore, we showed that Total Correlation can estimate functional connectivity in the real neural dataset and find biomarkers for diagnosing brain diseases.

In the future, we plan to use the functional connectivity relationships discovered by Total Correlation as an input to existing Graph Neural Networks (GNNs) [54] for the purpose of interpretable brain disease diagnosis, such that practitioners or doctors can identify the most informative subgraphs (or modules) to the decision (e.g., autism patients or healthy control groups). In this regard, quantitative measures to define differences between graphs [55] and the extension of analytical results in [25] to a larger number of nodes will be critical to assess and improve the qualitative results presented here. The recently proposed approaches (e.g., [56,57]) all rely on pairwise relationships estimated by the linear correlation coefficient as the input, which ignores high-order dependence essentially. In this sense, we believe our approach has the potential to improve the explanation performances of existing GNNs on brains.

Author Contributions: Conceptualization, methodology, software, and validation, Q.L., writing—original draft preparation, writing—review, and editing, Q.L., G.V.S., S.Y. and J.M. The contribution of J.M. was focused on the definition of the paper scope about large-scale connectomes, the relation with analytical results in [25], and the criticism of performance measures. All authors have read and agreed to the published version of the manuscript.

Funding: Q-L- and J-M- were partially funded by the Spanish/European grants from GVA/AEI/FEDER/EU: MICINN PID2020-118071GB-I00, MICINN PDC2021-121522-C21, and GVA Grisolía-P/2019/035. G.V.S. acknowledges support from the Defense Advanced Research Projects Agency (DARPA) under Award FA8750-17-C-0106. S.Y. was funded by the Research Council of Norway under Grant No. 309439. Finally, we thank the organizers of the HCP and ACPI for providing these interesting dataset used in these studies.

Institutional Review Board Statement: All the human data used in this research is an open-source dataset, and therefore, this study does not relate to any ethics.

Data Availability Statement: The data and code needed to reproduce the results presented here are available at https://forms.gle/1DXDpEpi7AodQ77q7 accessed on 6 November 2022.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

TC	Total Correlation
CorEx	Correlation Explanation
CC	Linear Correlation
I	Mutual Information
VAEs	Variational Autoencoders
fMRI	functional Magnetic Resonance Imaging
BOLD	Blood-Oxygen-Level-Dependent Imaging
DCM	Dynamic Causal Modeling
GLM	General Linear Model
ROI	Region Of Interest
HCP	Human Connectome Project
MTA	Multimodal Treatment of Attention Deficit Hyperactivity Disorder
GNNs	Graph Neural Networks

Appendix A

Table A1. Information of 116 brain regions that comprises the AAL atlas.

Brain Area	AAL Regions	AAL Index No.
Frontal Lobe	Precentral gyrus	1, 2
	Superior frontal gyrus, dorsolateral	3, 4
	Superior frontal gyrus, orbital part	5, 6
	Middle frontal gyrus	7, 8
	Middle frontal gyrus, orbital part	9, 10
	Inferior frontal gyrus, opercular part	11, 12
	Inferior frontal gyrus, triangular part	13, 14
	Inferior frontal gyrus, orbital part	15, 16
	Rolandic operculum	17, 18
	Supplementary motor area	19, 20
	Olfactory cortex	21, 22
	Superior frontal gyrus, medial	23, 24
	Superior frontal gyrus, medial orbital	25, 26
	Gyrus rectus	27, 28
	Paracentral lobule	69, 70
Insula and Cingulate	Insula	29, 30
	Anterior cingulate and paracingulate gyri	31, 32
	Median cingulate and paracingulate gyri	33, 34
	Posterior cingulate gyrus	35, 36
Temporal Lobe	Hippocampus	37, 38
	Parahippocampal gyrus	39, 40
	Amygdala	41, 42
	Fusiform gyrus	55, 56
	Heschl gyrus	79, 80
	Superior temporal gyrus	81, 82
	Temporal pole: superior temporal gyrus	83, 84
	Middle temporal gyrus	85, 86
	Temporal pole: middle temporal gyrus	87, 88
	Inferior temporal gyrus	89, 90

Table A1. *Cont.*

Brain Area	AAL Regions	AAL Index No.
Central Structures	Caudate nucleus	71, 72
	Lenticular nucleus, putamen	73, 74
	Lenticular nucleus, pallidum	75, 76
	Thalamus	77, 78
Occipital Lobe	Calcarine fissure and surrounding cortex	43, 44
	Cuneus	45, 46
	Lingual gyrus	47, 48
	Superior occipital gyrus	49, 50
	Middle occipital gyrus	51, 52
	Inferior occipital gyrus	53, 54
Parietal Lobe	Postcentral gyrus	57, 58
	Superior parietal gyrus	59, 60
	Inferior parietal, but supramarginal and angular gyri	61, 62
	Supramarginal gyrus	63, 64
	Angular gyrus	65, 66
	Precuneus	67, 68
Cerebellum and Vermis	Cerebellum Crus 1	91, 92
	Cerebellum Crus 2	93, 94
	Cerebellum 3	95, 96
	Cerebellum 4, 5	97, 98
	Cerebellum 6	99, 100
	Cerebellum 7b	101, 102
	Cerebellum 8	103, 104
	Cerebellum 9	105, 106
	Cerebellum 10	107, 108
	Vermis 1, 2	109
	Vermis 3	110
	Vermis 4, 5	111
	Vermis 6	112
	Vermis 7	113
	Vermis 8	114
	Vermis 9	115
	Vermis 10	116

References

1. Friston, K. Functional and effective connectivity: A review. *Brain Connect.* **2011**, *1*, 13–36. [CrossRef] [PubMed]
2. Porta, A.; Faes, L.; Bari, V.; Marchi, A.; Bassani, T.; Nollo, G.; Perseguini, N.M.; Milan, J.; Minatel, V.; Borghi-Silva, A.; Takahashi, A.C.M.; Catai, A.M. Effect of age on complexity and causality of the cardiovascular control: Comparison between model-based and model-free approaches. *PLoS ONE* **2014**, *9*, e89463. [CrossRef]
3. Heuvel, M.; Pol, H. Exploring the brain network: A review on resting-state fmri functional connectivity. *Eur. Neuropsychopharmacol. J. Eur. Coll. Neuropsychopharmacol.* **2010**, *20*, 519–534. [CrossRef] [PubMed]
4. Sporns, O.; Tononi, G.; Kötter, R. The human connectome: A structural description of the human brain. *PLoS Comput. Biol.* **2005**, *1*, e42. [CrossRef] [PubMed]
5. Bastos, A.; Schoffelen, J.-M. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Front. Syst. Neurosci.* **2016**, *9*, 1. [CrossRef]
6. Lizier, J.T.; Heinzle, J.; Horstmann, A.; Haynes, J.; Prokopenko, M. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. *J. Comput. Neurosci.* **2011**, *30*, 85–107. [CrossRef]
7. Piasini, E.; Panzeri, S. Information theory in neuroscience. *Entropy* **2019**, *21*, 62. [CrossRef]
8. Ince, R.; Giordano, B.; Kayser, C.; Rousselet, G.; Gross, J.; Schyns, P. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Hum. Brain Mapp.* **2016**, *38*, 11. [CrossRef]
9. Dimitrov, A.; Lazar, A.; Victor, J. Information theory in neuroscience. *J. Comput. Neurosci.* **2011**, *30*, 1–5. [CrossRef]
10. Borst, A.; Theunissen, F. Information theory and neural coding. *Nat. Neurosci.* **1999**, *2*, 947–957. [CrossRef]

11. Tkacik, G.; Marre, O.; Mora, T.; Amodei, D.; Berry, M., II; Bialek, W. The simplest maximum entropy model for collective behavior in a neural network. *J. Stat. Mech. Theory Exp.* **2012**, *2013*, 7.
12. Gomez-Villa, A.; Bertalmio, M.; Malo, J. Visual information flow in Wilson-Cowan networks. *J. Neurophysiol.* **2020**. [CrossRef] [PubMed]
13. Malo, J. Spatio-chromatic information available from different neural layers via gaussianization. *J. Math. Neurosci.* **2020**, *10*, 18. [CrossRef] [PubMed]
14. Malo, J. Information flow in biological networks for color vision. *Entropy* **2022**, *24*, 1442. [CrossRef]
15. Farahani, F.; Karwowski, W.; Lighthall, N. Application of graph theory for identifying connectivity patterns in human brain networks: A systematic review. *Front. Neurosci.* **2019**, *13*, 585. [CrossRef]
16. Sporns, O. Graph theory methods: Applications in brain networks. *Dialogues Clin. Neurosci.* **2018**, *20*, 111–121. [CrossRef]
17. Rosas, F.; Mediano, P.A.M.; Ugarte, M.; Jensen, H.J. An information-theoretic approach to self-organisation: Emergence of complex interdependencies in coupled dynamical systems. *Entropy* **2018**, *20*, 793. [CrossRef]
18. Rosas, F.E.; Mediano, P.A.M.; Gastpar, M.; Jensen, H.J. Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Phys. Rev. E* **2019**, *100*, 032305. [CrossRef]
19. Tononi, G.; Edelman, G. Consciousness and complexity. *Science* **1999**, *282*, 1846–1851. [CrossRef]
20. Pereda, E.; Quian, R.; Bhattacharya, J. Nonlinear multivariate analysis of neurophysiological signals. *Prog. Neurobiol.* **2005**, *77*, 1–37. [CrossRef]
21. Chai, B.; Walther, D.B.; Beck, D.M.; Fei-Fei, L. Exploring functional connectivity of the human brain using multivariate information analysis. In Proceedings of the 22nd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009 ; Curran Associates Inc.: Red Hook, NY, USA, 2009; pp. 270–278.
22. Wang, Z.; Alahmadi, A.; Zhu, D.; li, T. Brain functional connectivity analysis using mutual information. In Proceedings of the 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Orlando, FL, USA, 14–16 December 2015; pp. 542–546.
23. Jomaa, M.E.S.H.; Colominas, M.; Jrad, N.; Bogaert, P.V.; Humeau-Heurtier, A. A new mutual information measure to estimate functional connectivity: Preliminary study. In Proceedings of the Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Berlin, Germany, 23–27 July 2019; pp. 640–643.
24. Li, Q. Functional connectivity inference from fmri data using multivariate information measures. *Neural Netw.* **2022**, *146*, 85–97. [CrossRef] [PubMed]
25. Li, Q.; Steeg, G.V.; Malo, J. Functional connectivity in visual areas from Total Correlation. *arXiv* **2022**. Available online: https://arxiv.org/abs/2208.05770 (accessed on 11 August 2022).
26. Watanabe, S. Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.* **1960**, *4*, 66–82. [CrossRef]
27. Studeny, M.; Vejnarova, J. The multi-information function as a tool for measuring stochastic dependence. In *Learning in Graphical Models*; Springer: Dordrecht, The Netherlands; pp. 261–298.
28. Laparra, V.; Camps-Valls, G.; Malo, J. Iterative gaussianization: From ICA to random rotations. *IEEE Trans. Neural Netw.* **2011**, *22*, 537–549. [CrossRef] [PubMed]
29. Laparra, V.; Johnson, E.; Camps, G.; Santos, R.; Malo, J. Information theory measures via multidimensional gaussianization. *arXiv Stats. Mach. Learn.* **2022**. Available online: https://arxiv.org/abs/2010.03807 (accessed on 25 November 2020).
30. Essen, D.V.; Smith, S.; Barch, D.; Behrens, T.; Yacoub, E.; Ugurbil, K. The wu-minn human connectome project: An overview. *NeuroImage* **2013**, *80*, 62–79. [CrossRef]
31. Essen, D.C.; Ugurbil, K.; Auerbach, E.; Barch, D.; Behrens, T.E.J.; Bucholz, R.; Chang, A.; Chen, L.; Corbetta, M.; Curtiss, S.; et al. The human connectome project: A data acquisition perspective. *NeuroImage* **2012**, *62*, 2222–2231. [CrossRef]
32. Steeg, G.V.; Galstyan, A. Discovering structure in high-dimensional data through correlation explanation. *Adv. Neural Inf. Process. Syst.* **2014**, *577*.
33. Steeg, G.V.; Galstyan, A. Maximally informative hierarchical representations of high-dimensional data. In *AISTATS'15*; PMLR; San Diego, California, USA, 2015.
34. Cover, T.M.; Thomas, J.A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*; Wiley-Interscience: Hoboken, NJ, USA, 2006.
35. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef] [PubMed]
36. Fabila-Carrasco, J.S.; Tan, C.; Escudero, J. Permutation entropy for graph signals. *IEEE Trans. Signal Inf. Process. Over Netw.* **2022**, *8*, 288–300. [CrossRef]
37. Lyu, S.; Simoncelli, E.P. Nonlinear Extraction of Independent Components of Natural Images Using Radial Gaussianization. *Neural Comput.* **2009**, *21*, 1485–1519. [CrossRef] [PubMed]
38. Gao, S.; Brekelmans, R.; Steeg, G.V.; Galstyan, A. Auto-encoding correlation explanation. In Proceedings of the 22nd International Conference on AI and Statistics (AISTATS), Naha, Japan, 16–18 April 2019.
39. Yu, S.; Giraldo, L.G.S.; Jenssen, R.; Principe, J.C. Multivariate extension of matrix-based rényi's α-order entropy functional. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2960–2966. [CrossRef] [PubMed]
40. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
41. Steeg, G.V. Unsupervised learning via Total Correlation explanation. In *IJCAI*; Melbourne, Australia, 2017.

42. Steeg, G.V.; Harutyunyan, H.; Moyer, D.; Galstyan, A. Fast structure learning with modular regularization. *Adv. Neural Inf. Process. Syst.* **2019**, 15593–15603.
43. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
44. Tzourio-Mazoyer, N.; Landeau, B.; Crivello, P.D.F.F.; Etard, O.N.D.; Delcroix, N.; Mazoyer, B.; Marc, J. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage* **2002**, *15*, 273–289. [CrossRef] [PubMed]
45. Behan, B.; Connolly, G.; Datwani, S.; Doucet, M.; Ivanovic, J.; Morioka, R.; Stone, A.; Watts, R.; Smyth, B.; Garavan, H. Response inhibition and elevated parietal-cerebellar correlations in chronic adolescent cannabis users. *Neuropharmacology* **2013**, *84*, 6. [CrossRef]
46. Bubl, E.; van Elst, L.T.; Gondan, M.; Ebert, D.; Greenlee, M. Vision in depressive disorder. *World J. Biol. Psychiatry Off. J. World Fed. Soc. Biol. Psychiatry* **2007**, *10*, 377–384. [CrossRef]
47. Zhang, R.; Volkow, N. Brain default-mode network dysfunction in addiction. *NeuroImage* **2019**, *200*, 313–331. [CrossRef]
48. Giedd, J.; Keshavan, M.; Paus, T. Why do many psychiatric disorders emerge during adolescence? *Nat. Rev. Neurosci.* **2008**, *9*, 947–957.
49. Medina, K.; Hanson, K.; Dager, A.; Cohen-Zion, M.; Nagel, B.; Tapert, S. Neuropsychological functioning in adolescent marijuana users: Subtle deficits detectable after a month of abstinence. *J. Int. Neuropsychol. Soc. JINS* **2007**, *13*, 807–820. [CrossRef]
50. Poline, J.-B.; Brett, M. The general linear model and fmri: Does love last forever? *NeuroImage* **2012**, *62*, 871–880. [CrossRef] [PubMed]
51. Dowdle, L.T.; Ghose, G.; Chen, C.C.C.; Ugurbil, K.; Yacoub, E.; Vizioli, L. Statistical power or more precise insights into neuro-temporal dynamics? assessing the benefits of rapid temporal sampling in fmri. *Prog. Neurobiol.* **2021**, *207*, 102171. [CrossRef] [PubMed]
52. Marreiros, A.; Stephan, K.; Friston, K. Dynamic causal modeling. *Scholarpedia* **2010**, *5*, 9568. [CrossRef]
53. Porta, A.; Faes, L. Wiener–granger causality in network physiology with applications to cardiovascular control and neuroscience. *Proc. IEEE* **2016**, *104*, 282–309. [CrossRef]
54. Welling, M.; Kipf, T.N. Semi-supervised classification with graph convolutional networks. In Proceedings of the (ICLR 2017), Toulon, France, 24–26 April 2017.
55. Tantardini, M.; Ieva, F.; Tajoli, L.; Piccardi, C. Comparing methods for comparing networks. *Sci. Rep.* **2019**, *9*, 17557. [CrossRef]
56. Cui, H.; Dai, W.; Zhu, Y.; Li, X.; He, L.; Yang, C. Brainnnexplainer: An interpretable graph neural network framework for brain network based disease analysis. *arXiv* **2021**, arXiv:2107.05097.
57. Zheng, K.; Yu, S.; Li, B.; Jenssen, R.; Chen, B. Brainib: Interpretable brain network-based psychiatric diagnosis with graph information bottleneck. *arXiv* **2022**, arXiv:2205.03612.

.

Paper III

FUNCTIONAL CONNECTIVITY IN VISUAL AREAS FROM TOTAL CORRELATION

Qiang Li
Image Processing Laboratory
University of Valencia
Valencia, 46980
qiang.li@uv.es

Greg Ver Steeg
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
gregv@isi.edu

Jesus Malo
Image Processing Laboratory
University of Valencia
Valencia, 46980
jesus.malo@uv.es

ABSTRACT

A recent study invoked the superiority of the *Total Correlation* concept over the conventional pairwise measures of functional connectivity in neuroscience. That seminal work was restricted to show that empirical measures of *Total Correlation* lead to connectivity patterns that differ from what is obtained using *linear correlation* and *Mutual Information*. However, beyond the obvious multivariate versus bivariate definitions, no theoretical insight on the benefits of Total Correlation was given. The accuracy of the empirical estimators could not be addressed because no controlled scenario with known analytical result was considered either.

In this work we analytically illustrate the advantages of *Total Correlation* to describe the functional connectivity in the visual pathway. Our neural model includes three layers (retina, LGN, and V1 cortex) and one can control the connectivity among the nodes, within the cortex, and the eventual top-down feedback. In this multivariate setting (three nodes with multidimensional signals), we derive analytical results for the three-way *Total Correlation* and for all possible pairwise *Mutual Information* measures. These analytical results show that pairwise *Mutual Information* cannot capture the effect of different intra-cortical inhibitory connections while the three-way *Total Correlation* can. The presented analytical setting is also useful to check empirical estimators of *Total Correlation*. Therefore, once certain estimator can be trusted, one can explore the behavior with natural signals where the analytical results (that assume Gaussian signals) are no longer valid. In this regard (a) we explore the effect of connectivity and feedback in the analytical retina-cortex network with natural images, and (b) we assess the functional connectivity in V1-V2-V3-V4 from actual fMRI recordings.

***K*eywords** Functional Connectivity, Information in Networks, Total Correlation, Mutual Information, Visual Brain, Retina-Cortex Pathway, Linear Receptive Fields, Divisive Normalization, Intra-Cortical Connections.

1 Introduction

Functional connectivity in neural networks goes beyond structural links: it is related to the way information is shared among multiple neural nodes [1, 2]. Quantifying the communication between multiple neural regions is key to understand brain function. However, most of the literature on functional connectivity just describes pairwise relationships because the conventional measures (such as correlation and mutual information) cannot cope with more than two nodes simultaneously. As a result, studies involving more than two brain regions at the same time are rare.

A recent study proposed the use of Total Correlation as a way to overcome the intrinsic pairwise limitation of the conventional measures of functional connectivity in neuroscience [3]. The multivariate nature of Total Correlation, T [4] is a *by-definition* advantage over Mutual Information, I [5]. However, the seminal work that proposed T as a measure of functional connectivity [3] had a fundamental limitation: beyond the obvious multivariate definition of T, no extra theoretical insight on its benefits was given. As a result of the lack of analytical models, the accuracy of the empirical estimators could not be addressed because no controlled scenario was considered either.

The goal of this work is addressing the limitations of [3] in the context of the visual brain. We do it through the consideration of simple but plausible analytical models of the retina-V1-cortex pathway.

The three-node model considered here (retina-LGN-V1) consists of the conventional linear receptive fields plus Divisive Normalization nonlinearities [6–9]. Following [10, 11] we keep the dimensionality relatively small so that reliable estimations of information-theoretic variables can be done. However, the biological plausibility of the considered setting is explicitly checked against human data of visual psychophysics. In this general setting every node (or layer) has noisy neurons so that part of the visual information is lost along the way. We consider two variations of this theoretical setting: *Model I* is a nonlinear network with intra-cortical connections, and *Model II* is its linear version with top-down feedback. When considering Gaussian signals both *Model I* (nonlinear) and *Model II* (recurrent) are analytically tractable.

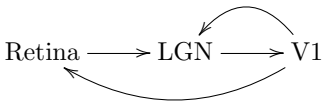
In this work we derive expressions for T and I depending on the feedforward and feedback structural connectivity. The key issue is the sensitivity of the descriptor: the bigger the variation of the descriptor in the range of explored connectivity the better. In that way, the representation of the connectivity will be more robust to errors in the estimation of the descriptor. Our analytical results show that while I is insensitive to some of the connectivity parameters, T is sensitive to the connectivity. These analytical results explicitly show the superiority of T over I as a description of the connectivity in a biologically plausible network.

On the other hand, the presented analytical results constitute a test-bed to check the accuracy of different empirical estimators for T (or I). In this way, available estimators (as for instance [12–17]) can be reliably applied to real data where theoretical results are not available (for instance because the Gaussian assumption is no longer valid [18–21]). Finally, in this paper we discuss the shared information between cortical areas using a recent fMRI dataset [22].

The structure of the paper is as follows. Section 2 describes the structural connectivity in the neural models considered throughout the work and their biological plausibility through the reproduction of visual psychophysics. In Section 3 we derive the analytical results for the functional connectivity (both I and T) in terms of the structural connectivity and the properties of the signal. In the analytical results we consider both feedforward nonlinear models and models with feedback. Section 4 presents T and I results computed with empirical estimators that can be compared to the theoretical results of Section 3. Moreover, results for real signals (natural images and and actual responses measured using fMRI) are also presented here. Finally, Section 5 summarizes the results and discusses the implications of the work.

2 Models of the retina-cortex pathway

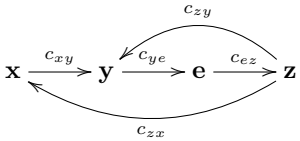
Expanding and making explicit the multi-node scenario first considered in [3], all the theoretical results of this work will be derived for the following *early vision* setting that may include feedforward and feedback connections, as for instance:



(1)

In this diagram the arrows represent structural connections between regions (or layers). Right-arrows represent feedforward flow of the visual information, and the left-arrows represent eventual feedback.

More specifically, the signal at the *retina* will be represented by the n-dimensional random vector, $\mathbf{x}$, the signal at the *LGN*, will be represented by the n-dimensional random vector, $\mathbf{y}$, and the signal at the cortex will be represented by two n-dimensional random vectors, $\mathbf{e}$ and $\mathbf{z}$. In this way, the intra-cortical connectivity is represented by the communication between $\mathbf{e}$ and $\mathbf{z}$. In the following diagram the strength of the structural connections between layers i and j is represented by the variables, c_{ij}:



(2)

In the above setting, the study of functional connectivity through information-theoretic measures (such as I or T) could be useful to describe the *unknown* strengths, c_{ij}, from recordings of the neural signal done at the different nodes or layers. In this context, proper measures of statistical relation should be sensitive to c_{ij}. And the bigger the sensitivity to the strength of the connections, the better.

2.1 *Model I*: Nonlinear and noisy model with focus on intra-cortical interactions

Our first specific example of the retina-cortex framework outlined in Eq. 2, which we will refer to as *Model I*, tries to be analytically simple yet biologically plausible. To do so, this models includes: (a) center-surround receptive fields in the LGN [23], (b) local-frequency receptive fields in the (linear) V1-cortex, approximated here as block-DCT basis functions [24, 25], (c) a Divisive Normalization transform to model cortical nonlinearities [8], and (d) noise in each of the neural layers is scaled in a way compatible with the psychophysical results in [26] and the physiological model in [27].

In Section 2.4 we will see that the above elements (the considered layers and noise levels) are critical for the correlation with human opinion in visual psychophysics. In this regard, the intra-cortical connectivity in the Divisive Normalization transform is particularly relevant. Therefore, eventual measures of the statistical relation between neural nodes should be sensitive to this intre-cortical connectivity.

The class of networks under *Model I* follows these equations:

$$\begin{aligned}
\mathbf{x}(t) &= \mathbf{s}(t) + \mathbf{n}_{\mathbf{x}}(t) + \frac{c_{zx}}{c_{xy}\, c_{ye}}\, F^{-1} \cdot \mathbf{z}(t - \Delta t) \\
\mathbf{y}(t) &= c_{xy}\, K \cdot \mathbf{x}(t) + \mathbf{n}_{\mathbf{y}}(t) \quad = \quad c_{xy}\, F^{-1} \cdot \lambda_{\text{CSF}} \cdot F \cdot \mathbf{x}(t) + \mathbf{n}_{\mathbf{y}}(t) \\
\mathbf{e}(t) &= c_{ye}\, F \cdot \mathbf{y}(t) + \mathbf{n}_{\mathbf{e}}(t) \\
\mathbf{z}(t) &= f\left(\mathbf{e}(t)\right) \quad = \quad \text{sign}(\mathbf{e}(t)) \cdot \kappa \cdot \frac{|\mathbf{e}(t)|^{\gamma}}{b \;+\; c_{ez}\, H \cdot |\mathbf{e}(t)|^{\gamma}}
\end{aligned} \tag{3}$$

where, the input to the system is the retinal image: the source vector $\mathbf{s} \in \mathbb{R}^n$, and its dimension n corresponds to the number of photoreceptors. In the models considered in this work, the networks preserve the dimension of the signal[1].

The retinal signal, the vector $\mathbf{x} \in \mathbb{R}^n$, is influenced by the input image $\mathbf{s}$, but it is also affected by the white noise $\mathbf{n}_{\mathbf{x}}$ and in this formulation, by a top-down feedback signal given by the term weighted by c_{zx}, that describes the strength of this feedback connection. Due to the eventual variations in the input and the eventual feedback, all the multivariate signals may depend on time, t. We will come back to the feedback term once we introduce the frequency meaning of vector $\mathbf{z}$.

The signal at the LGN is described by the vector $\mathbf{y} \in \mathbb{R}^n$. The matrix K contains the center-surround receptive fields of LGN [23]. According to the relation between these receptive fields and the Contrast Sensitivity Function (CSF) [30–32], we implement them using a local-frequency transform (basis in the matrix F), a diagonal matrix with CSF-related weights, λ_{CSF}, and coming back to the spatial domain using F^{-1}. The LGN signal is also affected by white noise through $\mathbf{n}_{\mathbf{y}}$.

The (intermediate) linear signal at the V1-cortex, $\mathbf{e}$, is computed from the LGN signal through a set of local-frequency receptive fields in the matrix F. This linear signal is also affected by the white noise $\mathbf{n}_{\mathbf{e}}$.

Finally, the nonlinear signal at V1, $\mathbf{z}$, results from a Divisive Normalization transform, $f(\cdot)$, of the outputs of the linear receptive fields at the previous intermediate layer, $\mathbf{e}$. Note that the division, the exponent, and the absolute values in $f(\cdot)$ are Hadamard (element-wise) operations [33], and the matrix H in the denominator represents the interaction between the neurons of the previous cortical layer $\mathbf{e}$. Specifically, the intra-cortical connectivity between the k-th and the l-th neurons is represented by $c_{ez} H_{kl}$. In this way, the k-th row of H, $H_{kl} \;\; \forall l = 1, \ldots, n$, describes how the responses of the neighbor linear neurons, e_l, affect the nonlinear response of the k-th neuron, z_k. This interaction is assumed to be local in space and frequency [9, 33, 34]. And c_{ez} controls the global strength of all these local interactions.

Finally, a comment on the top-down feedback term in the first equation. The Divisive Normalization changes the relative magnitude of the responses z_i but the rough qualitative meaning of the responses in $\mathbf{z}$ is still given by the (local-frequency) receptive fields in F. Therefore, the F^{-1} matrix in the top-down feedback term in the first equation of the system just converts the previous cortical response $\mathbf{z}(t - \Delta t)$ back into the spatial domain (where the input images $\mathbf{s}$ are). Additionally, the top-down term has been scaled by the other connectivity strengths (c_{xy} and c_{ye}) just to keep the scale of the feedback term comparable to the source independently of the (arbitrary) gains introduced along the retina-cortex path. In this way the effective weight of the feedback term only depends on c_{zx}.

The parameters that control the feedforward structural connections between retina, LGN, and the linear V1, (i.e. the strengths c_{xy} and c_{ye}) actually control the size of the signal with regard to the noise, and hence their functional role is

[1] Preservation of dimension along the pathway is convenient but it doesnt reduce the generality neither biologically, the spatial subsampling affects the extrafovea, but not the fovea [28], nor mathematically because changes of dimension could be addressed by the Jacobians of rectangular transforms [29].

quite evident: the bigger the signal compared to the noise, the stronger the information flow from one node/layer to the next. However, the role of the intra-cortical interaction $c_{ez}H$ is more interesting. There is a large body of literature that suggests that the role of the denominator in Divisive Normalization is capturing-and-removing the statistical relations between the responses of the linear local-frequency sensors [20, 21, 35–38].

The first set of analytical results derived in Section 3.1 shows how T is sensitive to this (eventually interesting) intra-cortical connectivity, while the sensitivity of I to these intra-cortical connections is equal to zero. This is an analytical example (for a biologically plausible nonlinear network) of the genuine superiority of the Total Correlation over the conventional Mutual Information.

2.2 *Model II*: Linear noisy model with focus on feedback

Model II is just a variation of *Model I* intended to simplify the analytical study of feedback. The convenience of this variation will become apparent in Section 3 when we derive the analytical results. By comparing the Eqs. 3 of *Model I* and Eqs. 4 of *Model II* it is easy to see that our second class of networks is just a linear version of the first where we disregarded the Divisive Normalization. Specifically, in the last equation of *Model II* the cortical nonlinearity $f(\cdot)$ has been substituted by a trivial identity, $\mathbb{I}$, and the input cortical signal is scaled by the strength c_{ez} with regard to the inner noise $\mathbf{n_z}$, which was not present before:

$$\begin{aligned}
\mathbf{x}(t) &= \mathbf{s}(t) + \mathbf{n_x}(t) + \frac{c_{zx}}{c_{xy}\, c_{ye}\, c_{ez}}\, F^{-1} \cdot \mathbf{z}(t - \Delta t) \\
\mathbf{y}(t) &= c_{xy}\, K \cdot \mathbf{x}(t) + \mathbf{n_y}(t) \quad = \quad c_{xy}\, F^{-1} \cdot \lambda_{\text{CSF}} \cdot F \cdot \mathbf{x}(t) + \mathbf{n_y}(t) \\
\mathbf{e}(t) &= c_{ye}\, F \cdot \mathbf{y}(t) + \mathbf{n_e}(t) \\
\mathbf{z}(t) &= c_{ez}\, \mathbb{I} \cdot \mathbf{e}(t) + \mathbf{n_z}(t)
\end{aligned} \tag{4}$$

Recurrence implied by feedback (both in *Model I* and *Model II*) implies a nontrivial evolution of the signals when the system faces dynamic inputs with fast variations compared with the updating time constant Δt. In this work we will restrict ourselves to slow-varying sources $\mathbf{s}(t)$ and we wait till the convergence of the signals to a stationary state to measure the statistical dependence between the signals at the different layers.

In the setting described by *Model II* the information about the input image (or source $\mathbf{s}$) flows through the feedforward links while being contaminated by the noise injected at each layer. However, for the slow-varying inputs described above, part of the source is injected back into the retinal signal. As a result, the scenario in *Model II* is convenient to analyze the joint effect of the strength of the feedforward links and the feedback links. For example, one may study the effect of the intra-cortical connectivity c_{ez} (that scales the signal wrt the inner noise) together with the strength of the feedback c_{zx} that reinforces the presence of the source at the retina. From a naive perspective, increasing c_{ez} and c_{zx} seems to lead to an increase of the Signal-to-Noise ratio in all the responses. Analytical results of information-theoretic descriptors can confirm or refute this intuition and provide a tool to understand a variety of situations.

The second set of analytical results derived in Section 2.2 show that while T strongly depends on the feedforward and feedback strengths c_{ez} and c_{zx}, the sensitivity of I is smaller. In this case, the sensitivity of I is just smaller (not zero) but the substantial difference in sensitivities (in a biologically plausible recurrent scenario) illustrates the conceptual superiority of T over the conventional I.

2.3 Model parameters: receptive fields, divisive normalization and responses

In this section we present and illustrate the range of parameters that we considered in Eqs. 3 and 4 of *Model I* and *Model II*.

First, note that the throughout the work we consider that the input to our system are achromatic image patches of 8×8 pixels. This means that vectors $\mathbf{s}$, $\mathbf{x}$, $\mathbf{y}$, $\mathbf{e}$, and $\mathbf{z}$ live in $\mathbb{R}^{64}$, and we consider layers (or nodes) with $n = 64$ neurons. Therefore, matrices K, F, λ_{CSF}, and H (that represent relations between neurons) are 64×64 matrices.

Figure 1 illustrates the parameters involved in the retina-to-LGN transform ($\mathbf{x} \to \mathbf{y}$) and in the LGN-to-cortex transform ($\mathbf{y} \to \mathbf{z}$), as well as in the intra-cortical nonlinearity ($\mathbf{e} \to \mathbf{z}$) of *Model I*.

First, regarding $\mathbf{x} \to \mathbf{y}$ we follow the relation between the center-surround cells in LGN and the CSF, and hence we compute K from the CSF of the Standard Spatial Observer [39] transformed from the original Fourier domain into the (more convenient) DCT domain using the procedure in [40] (second panel in Fig. 1). The result (in the spatial domain) are center-surround receptive fields which are consistent with the physiological measurements [23] (first panel in 1).

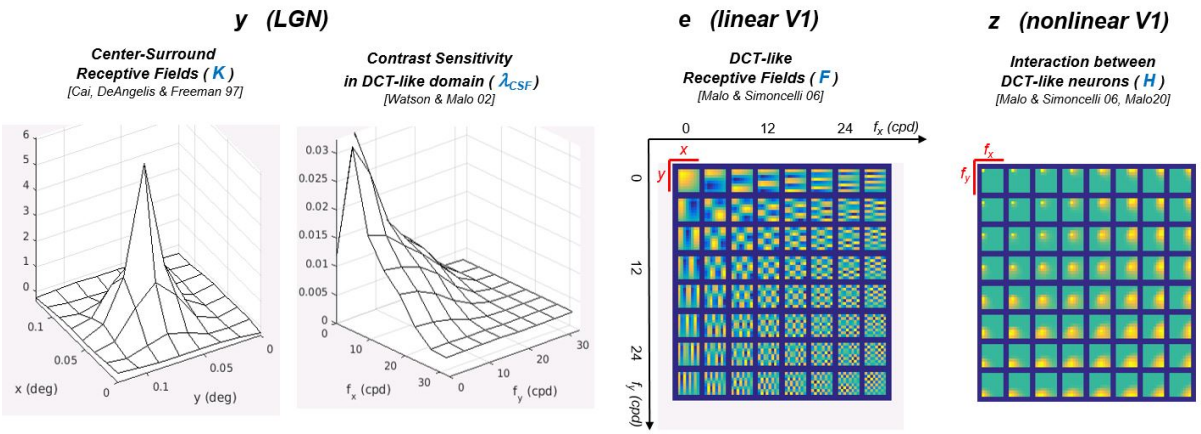


Figure 1: Center-surround receptive fields in LGN and equivalent Contrast Sensitivity Function. Local frequency filters tuned to different orientations in linear V1 and interaction kernel $H_{ff'}$ in the divisive normalization nonlinearity in V1.

Then, the linear cortical transform $\mathbf{y} \rightarrow \mathbf{e}$ uses the local-DCT representation following previous results on biologically-inspired image compression [41, 42] and subjective image quality [43, 44]. The 64×64 local-frequency receptive fields in F (DCT-like basis functions) are shown in the third panel of Fig. 1.

Finally, regarding the intra-cortical Divisive Normalization, $\mathbf{e} \rightarrow \mathbf{z}$, here we also follow models used in biologically-inspired image compression methods [37, 45]. In this case, the structural connectivity between different local-frequency sensors decays with distance in frequency according to a Gaussian [33, 34]:

$$H_{ff'} = e^{-\frac{(f-f')^2}{\sigma(f)^2}} \tag{5}$$

where the width $\sigma(f)$ increases with the frequency f, according to $\sigma(f) = \sigma_0 + \alpha_H f$, as illustrated in the example of the fourth panel of Fig. 1. In that case, the connectivity neighborhood is wider for sensors of high frequency (bottom right of the plot) than for sensors of low frequency (top left of the plot). Finally, in our experiments we set the semi-saturation constant b and the constant κ according to the method in [10] so that the Divisive Normalization is compatible with classical non-linearities such as the Wilson-Cowan recurrent model [46].

In the experiments we consider a range of intra-cortical connectivity values in *Model I* (section 3.1), and we modify the width of the kernel H by varying the constant $\alpha_H \in [0.35, 4]$, and by varying the strength $c_{ez} \in [0.01, 300]$. This has an effect in the nonlinearity of the cortical responses and, as a consequence, on the statistical effect of $f(\cdot)$.

Figure 2 illustrates the transformations of the signal along the layers of *Model I* for a representative set of parameters (those that maximize correlation with human psychophysics). The top panel shows (i) the input image $\mathbf{s}$: in this case the achromatic image of an eye in the range [0,200] cd/m^2, spatially sampled at 64 cycles/degree, (ii) how this input is distorted with the noise at the retina (leading to $\mathbf{x}$), (iii) the response of center-surround cells distorted by noise in $\mathbf{y}$, (iv) the response to 3×3 regions of local-frequency sensors in $\mathbf{e}$ (with the corresponding noise) in $\mathbf{e}$, and finally, (v) the result of the Divisive Normalization in $\mathbf{z}$. Additionally, for a qualitative understanding of the information lost along the way, the cortical signals ($\mathbf{e}$ and $\mathbf{z}$) are represented back in the spatial domain by transforming them using the linear inverse F^{-1}.

Following the argument in [26] the standard deviation of the noise injected at each layer has been selected such as it remains barely visible. This is because just-noticeable-differences are determined by this amount of noise [47]. Specifically, the standard deviation of the white noise at the different layers in *Model I* is $\sigma(n_x) = 5cd/m^2$ (for images with luminance in the range $[0, 200]cd/m^2$), $\sigma(n_y) = 0.1$, $\sigma(n_e) = 0.01$, and (on top of these values), in *Model II* we have $\sigma(n_z) = 0.01$.

Finally, the scatter plots at the bottom left of Fig. 2 illustrate the nonlinearities introduced by the considered Divisive Normalization. From the local DC components of the representation we can see the saturation of (perceived) brightness as a function of the input luminance, where we can see the Weber Law [48]. Similarly, the other plots for *low*, *medium*, and *high*, frequency coefficients, illustrate the nonlinearity of the perceived contrast as a function of the input contrasts. This sigmoidal and signal-dependent behavior is consistent with the psychophysics of contrast perception [34], and the amplitide of the responses for the different frequencies is consistent with the CSF [32, 40].

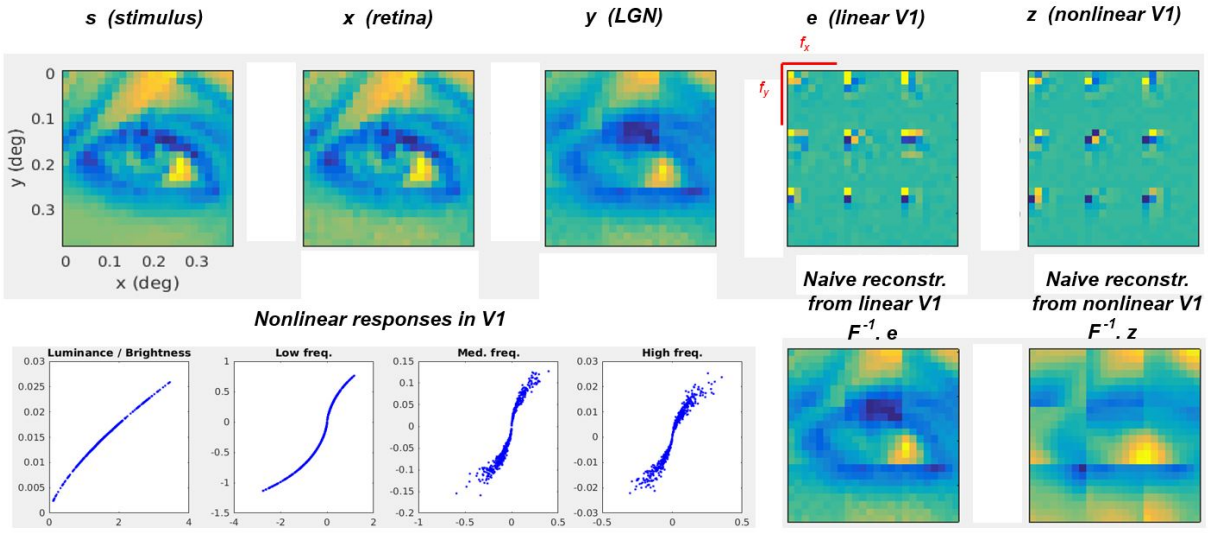


Figure 2: Responses to a sample image with the optimal set of parameters. Optimal means maximum correlation with human opinion among the considered discrete set of cortical connectivity values.

2.4 Model plausibility: image quality psychophysics

Qualitative Weber law, saturation of perceived contrast, and compatibility with the CSF displayed in Fig. 2 suggest that the parameters selected from the literature make sense. However, a more comprehensive/quantitative test is necessary particularly if a range of parameters has to be considered. To this end, in this section we assess the plausibility of the models according to their ability to predict experimental data on subjective image quality, specifically the ratings given by humans in the TID database [49]. This way of determining biologically plausible parameters is not new [21, 39, 50] and it has been subject to criticism as a single measurement of performance [9]. However, in the context presented here, prediction of subjective quality is enough to highlight the general behavior of the model and to (roughly) identify which regions of the parameter space make more biological sense.

In this regard, the scatter plots in Figure 3 show how well Euclidean distances at the different layers of *Model I* (abscisas), predict the subjective ratings (ordinates). The strong correlation obtained in the inner cortical representation $\rho = 0.84$, which is not far from the state-of-the-art in subjective image quality metrics [51] prove the plausibility of the transforms and the levels of the Gaussian noise introduced at each layer.

Specifically, the poor result for the input representation ($\mathbf{s}$ in luminance) implies that the visual brain certainly *does something else* to the input signal [52, 53]. The progressive improvement of the correlation along deeper layers means that the set of considered transforms is biologically meaningful. In fact, the consideration of the center-surround cells (or the CSF) is a major fact in explaining image quality [39, 43], and this is incorporated in both models leading to a reasonable Pearson correlation, $\rho = 0.71$, only with linear transforms. Then, we study the intra-cortical connectivity of *model I* in more detail: we consider the plausibility of a range of strengths c_{ez} and a range of widths in H.

The result shows that all the family of Divisive Normalization transforms make sense because they substantially improve the correlation with human opinion. Note that the correlation at the linear cortical layer $\mathbf{e}$ (surface in light blue at 0.71) is raised by the different $\mathbf{z}$ layers to be in the range $[0.76, 0.84]$. Moreover, the final correlation surface for the different intra-cortical connectivity values has strong curvature and a clear maximum (green dot) in the middle of the considered region. This means that it is interesting to study the behavior of the statistical descriptors of connectivity in this region of parameters.

3 Analytical results: T and I in terms of intra-layer connectivity and feedback

Here we present results for *Model I* and *Model II* which address different interesting situations that may happen in natural or artificial neural nets: (i) nonlinear intra-layer connectivity, and (ii) feedback or recurrence. In order to simplify the analytical tractability, in each case we focus on a specific feature of the models, either the nonlinearity (in *Model I*) or the feedback-recurrence (in *Model II*).

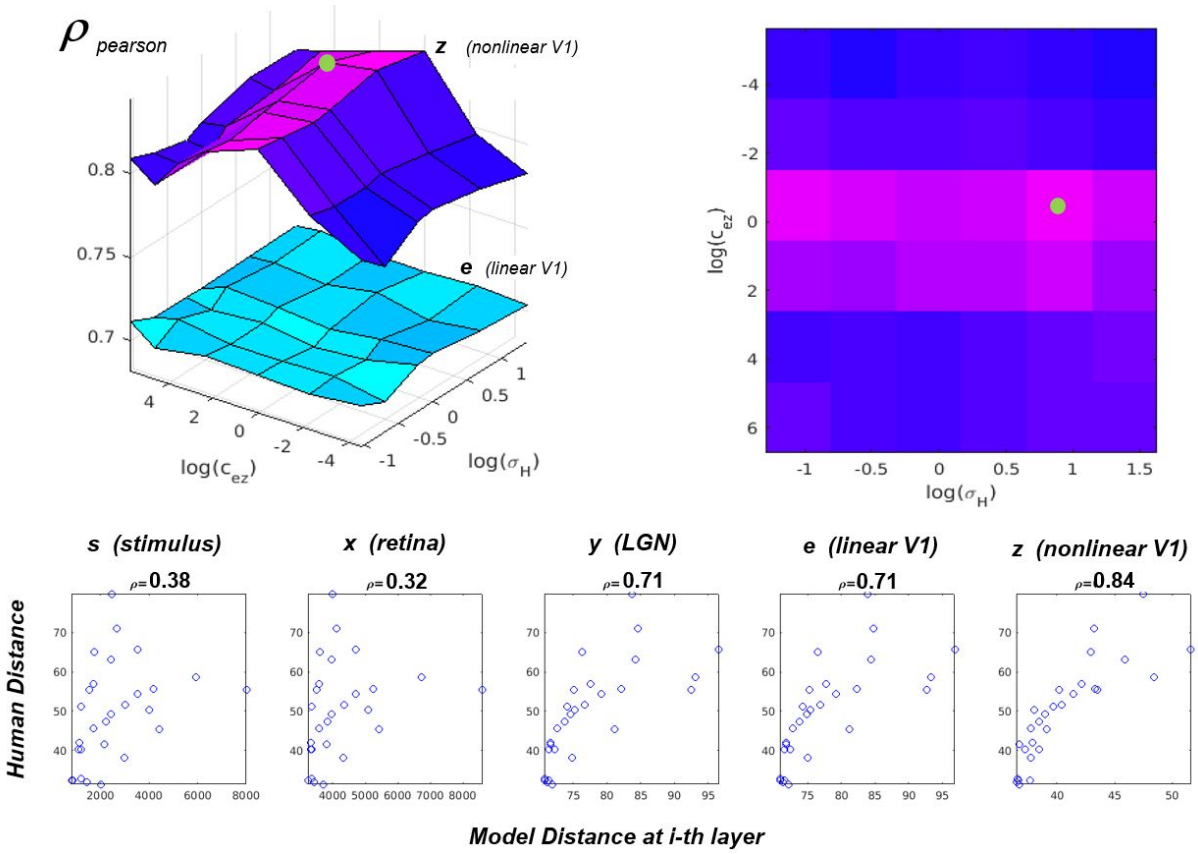


Figure 3: Correlation with human opinion for different cortical connectivity values (surfaces on top) and correlations in previous (linear) layers (scatter plots at the bottom). In the nonlinear cortical case the scatter plot is the one corresponding to the optimum connectivity.

For both models (*I* and *II*) analytical tractability is simple if one considers Gaussian signals. The Gaussian assumption for natural images has been acknowledged as a too rough approximation both in Visual Neuroscience [18–21] and in Image Processing [54, 55]. However, in this section we are going to take this assumption for the sake of analytical tractability. In the experimental section we will compare the results with (synthetic) Gaussian signals and natural inputs. The Gaussian assumption is appropriate and illustrative in this case because (as shown below using a trustable empirical estimator) results for natural images are (1) similar to the Gaussian results, and more important for this work, (2) they confirm the superiority of the description using T also for natural signals.

For the reader convenience, lets recall the definitions of the descriptors considered here: *Total Correlation* [4], and *Mutual Information* [5] in terms of *Entropy*:

$$T(\mathbf{x},\mathbf{y},\mathbf{z}) = \left(\sum_{i=1}^{n} h(x_i)+h(y_i)+h(z_i)\right) - h(\mathbf{x},\mathbf{y},\mathbf{z}) \quad (6)$$

$$I(\mathbf{x},\mathbf{y}) = h(\mathbf{x})+h(\mathbf{y})-h(\mathbf{x},\mathbf{y}) \quad (7)$$

The biggest conceptual difference between these magnitudes is, of course, that T can be applied to *any* number of nodes (or layers). However, even in the case of just two nodes, $T(\mathbf{x},\mathbf{y}) \neq I(\mathbf{x},\mathbf{y})$ because, for multivariate nodes, T considers the redundancy among the coefficients (or neurons) of each node, which is disregarded by I. This difference is key when the signals in each layer are not independent, which is the more interesting situation in visual neuroscience.

As joint and marginal entropy are easily computed for Gaussian signals from the covariance matrices or from the marginal variances [5], Eqs. 6 and 7 imply that, if variables are Gaussian, analytical results are straightforward. This is the case in *Model II*, but, due to the nonlinearity, it is not the case in *Model I*.

3.1 T and I as descriptors of intra-cortical connectivity (*Model I*)

For the sake of simplicity, in this section, on top of the Gaussian assumption mentioned above, we will also consider $c_{zx} = 0$ in our nonlinear *Model I*, i.e. it does not consider feedback. We leave feedback for the results of *Model II* in Section 3.2.

With these assumptions, the variables $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{e}$ are Gaussian because they are sum of linearly-transformed Gaussian variables plus white Gaussian noise. However, the Divisive Normalization nonlinearity $f(\cdot)$ implies that the variable $\mathbf{z}$ is non-Gaussian. In this setting, expressions for T and I involving $\mathbf{z}$ (where the intra-cortical connectivity is) require the application of specific properties of these magnitudes under transforms of the random variables.

The Total Correlation does depend on intra-cortical connectivity:

In order to get an analytical result for $T(\mathbf{x}, \mathbf{y}, \mathbf{z})$, lets concatenate the variables that represent the considered nodes into column vectors of dimension $3n$: $\mathbf{a} = [\mathbf{x}; \mathbf{y}; \mathbf{e}]$, and $\mathbf{a}' = [\mathbf{x}; \mathbf{y}; \mathbf{z}] = [\mathbf{x}; \mathbf{y}; f(\mathbf{e})]$, and consider,

$$\mathbf{a} \xrightarrow{\mathcal{F}} \mathbf{a}'$$

where we are interested in computing $T(\mathbf{a}')$. In this situation, one may use the following property of the variation of Total Correlation when the variables undergo a transformation $\mathcal{F}$ [11, 56]:

$$\Delta T(\mathbf{a}, \mathbf{a}') = T(\mathbf{a}) - T(\mathbf{a}') = \sum_i^{3n} h(a_i) - \sum_i^{3n} h(a_i') + \frac{1}{2}\mathbb{E}_{\mathbf{a}}\big\{ log|\nabla_{\mathbf{a}}\mathcal{F}^\top \cdot \nabla_{\mathbf{a}}\mathcal{F}| \big\} \tag{8}$$

where $\mathbb{E}_{\mathbf{a}}\big\{\cdot\big\}$ is the average over the samples $\mathbf{a}$. Then, taking into account that,

$$\nabla_{\mathbf{a}}\mathcal{F} = \begin{pmatrix} \mathbb{I} & 0 \\ 0 & \nabla_{\mathbf{e}} f \end{pmatrix}$$

and considering that $T(\boldsymbol{a}) = T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{e})$ only depends on Gaussian variables and hence with known entropy in terms of the covariance matrix[2], we obtain the desired result (in *nats*):

$$T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \frac{1}{2}\sum_i^{3n} log(\Sigma^a_{ii}) - \frac{1}{2}log|\Sigma^a| - \frac{n}{2} - \frac{n}{2}log(2\pi) - \frac{1}{2}log|\Sigma^e| + \sum_{i=1}^{n} h(z_i) - \frac{1}{2}\mathbb{E}_{\boldsymbol{e}}\{ log|\nabla_{\boldsymbol{e}} f \cdot \nabla_{\boldsymbol{e}} f^\top| \} \tag{9}$$

where the covariance matrices Σ^e and Σ^a do not depend on the intra-cortical connectivity, because they only depend on $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{e}$:

$$\Sigma^a = \Sigma^{xye} = \begin{pmatrix} \Sigma^x & c_{xy} \cdot \Sigma^x \cdot K^\top & c_{ye} \cdot c_{xy} \cdot \Sigma^x \cdot (F \cdot K)^\top \\ c_{xy} \cdot K \cdot \Sigma^x & \Sigma^y & c_{ye} \cdot \Sigma^y \cdot F^\top \\ c_{ye} \cdot c_{xy} \cdot F \cdot K \cdot \Sigma^x & c_{ye} \cdot F \cdot \Sigma^y & \Sigma^e \end{pmatrix}$$

but, according to [33], $\nabla_{\boldsymbol{e}} f$ does depend on the intra-cortical connectivity due to the interactions in the Divisive Normalization, c_{ez} and H:

$$\nabla_{\boldsymbol{e}} f = \mathbb{D}_{sign(\boldsymbol{e})} \cdot \mathbb{D}^{-1}_{\left(b + c_{ez} \cdot H \cdot |\boldsymbol{e}|\right)} \cdot [\mathbb{I} - c_{ez} \cdot \mathbb{D}_{\boldsymbol{z}} \cdot H] \cdot \mathbb{D}_{\left(\gamma \ sign(\boldsymbol{e}) |\boldsymbol{e}|^{\gamma - 1}\right)} \tag{10}$$

where $\mathbb{D}_{\boldsymbol{v}}$ is a diagonal matrix with the vector $\boldsymbol{v}$ in the diagonal.

Eqs. 9 and 10 explicitly show that $T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ *does* depend on the intra-cortical connectivity.

Another way to see the dependence with the intra-cortical connectivity consist of identifying these two terms in Eq. 9: the (Gaussian) $T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{e})$, using the definition in Eq. 6, and the variation of T under the transform $\mathbf{z} = f(\mathbf{e})$, using the property in Eq. 8. By doing that, it is easy to see that:

$$T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \Big(T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{e}) - T(\boldsymbol{e})\Big) + T(\boldsymbol{z}) \tag{11}$$

where the term in the parenthesis obviously does not depend on the intra-cortical connectivity (because $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{e}$ are previous to that interaction), but $T(\mathbf{z})$ *does* depend on the Divisive Normalization.

The Mutual Information measures do not capture the effect of intra-cortical connectivity: This is easy to see using the following property: the mutual information is invariant to non-singular differentiable transforms of the random vectors [57]:

$$I(\mathbf{a}, f(\mathbf{b})) = I(\mathbf{a}, \mathbf{b}) \tag{12}$$

[2] If $\mathbf{x}$ is a Gaussian variable, its entropy in *nats* is $h(\mathbf{x}) = \frac{1}{2}log|2\pi e\Sigma^x|$ where Σ^x is the covariance of $\mathbf{x}$ [5].

This property is easy to see by considering that $I(\mathbf{a}, \mathbf{b})$ measures the KL-divergence between the densities $p(\mathbf{a}, \mathbf{b})$ and $p(\mathbf{a})p(\mathbf{b})$ [5]. Taking into account that the Jacobian that appears in the variation of the probability under transforms [58] is compensated (in the integral of the KL-divergence) by the change of the differential volume, one gets the invariance.

As a result, no *pairwise* measure I involving $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ depends on the intra-cortical connectivity:

$$\begin{aligned} I(\boldsymbol{x}, \boldsymbol{y}) &= \tfrac{1}{2}\log|\Sigma^x| + \tfrac{1}{2}\log|\Sigma^y| - \tfrac{1}{2}\log|\Sigma^{xy}| \\ I(\boldsymbol{x}, \boldsymbol{z}) = I(\boldsymbol{x}, f(\boldsymbol{e})) &= I(\boldsymbol{x}, \boldsymbol{e}) = \tfrac{1}{2}\log|\Sigma^x| + \tfrac{1}{2}\log|\Sigma^e| - \tfrac{1}{2}\log|\Sigma^{xe}| \\ I(\boldsymbol{y}, \boldsymbol{z}) = I(\boldsymbol{y}, f(\boldsymbol{e})) &= I(\boldsymbol{y}, \boldsymbol{e}) = \tfrac{1}{2}\log|\Sigma^y| + \tfrac{1}{2}\log|\Sigma^e| - \tfrac{1}{2}\log|\Sigma^{ye}| \end{aligned} \tag{13}$$

where,

$$\Sigma^{xy} = \begin{pmatrix} \Sigma^x & c_{xy}\cdot\Sigma^x\cdot K^\top \\ c_{xy}\cdot K\cdot\Sigma^x & \Sigma^y \end{pmatrix}$$

$$\Sigma^{xe} = \begin{pmatrix} \Sigma^x & c_{ye}\cdot c_{xy}\cdot\Sigma^x\cdot(F\cdot K)^\top \\ c_{ye}\cdot c_{xy}\cdot F\cdot K\cdot\Sigma^x & \Sigma^e \end{pmatrix}$$

$$\Sigma^{ye} = \begin{pmatrix} \Sigma^y & c_{ye}\cdot\Sigma^y\cdot F^\top \\ c_{ye}\cdot F\cdot\Sigma^y & \Sigma^e \end{pmatrix}$$

Therefore, we have demonstrated an important concept: in the biologically plausible *Model I*, Eq. 13 means that the conventional I measures *do not capture* the intra-cortical connectivity, which is critical to get that high correlation with biology. On the contrary, Eqs. 9 and 10 explicitly show that T *does* depend on the intra-cortical connectivity.

3.2 T and I as descriptors of feedback (*Model II*)

In *Model II* there is no nonlinearity so, if the source $\mathbf{s}$ is Gaussian and so are the noises injected at the different layers, all the variables (in the forward pass) will be Gaussian including $\mathbf{z}$. Then, the considered feedback from $\mathbf{z}$ to $\mathbf{x}$ just injects an extra Gaussian variable back into $\boldsymbol{x}$. As a result, $\mathbf{x}$ will be Gaussian too for any strength of the feedback. For slow-varying inputs (as natural images at the retina) the feedback signal (coming from the past) is not totally independent of the current value of the source, so the covariance at the retina is not the sum of the covariance matrices of the separate terms in the sum in the first equation of *Model II*. However, this does not modify the Gaussian assumption.

All these considerations imply that the definitions in terms of entropy given in Eqs. 6 and 7 can be applied together with the expression of the entropy for Gaussian signals that only depends on the corresponding covariance matrices. As a result, in order to make explicit the dependence on feedforward and feedback connectivity one only has to consider all possible covariance matrices, which is what we list below for *Model II*.

Assuming that signal and noise are not correlated, the covariance matrices of the signal at each isolated layer are:

$$\begin{aligned} \Sigma^x &= \mathbb{E}\{x\cdot x^\top\} = \Sigma^s + \Sigma^{n_x} + \Big(\frac{c_{zx}}{c_{xy}c_{ye}c_{ez}}\Big)^2 F^{-1}\cdot\Sigma^z\cdot F^{-1\top} + \frac{c_{zx}}{c_{xy}c_{ye}c_{ez}}M(s,z) \\ \Sigma^y &= c_{xy}^2\cdot K\cdot\Sigma^x\cdot K^\top + \sigma^2(n_y)\,\mathbb{I} \\ \Sigma^e &= c_{ye}^2\cdot F\cdot\Sigma^y\cdot F^T + \sigma^2(n_e)\,\mathbb{I} \\ \Sigma^z &= c_{ez}^2\cdot\Sigma^e + n_e^2\cdot\mathbb{I}_d \end{aligned} \tag{14}$$

where $M(s,z)$ is a symmetric matrix that describes the relation between s and z (they are not independent), and it is given by: $M(s,z) = F^{-1}\cdot\mathbb{E}\{s\cdot z^\top\} + \big(F^{-1}\cdot\mathbb{E}\{s\cdot z^\top\}\big)^\top$.

Additionally, the covariance matrices of *two* concatenated vectors that have not been given in Section 3.1 are:

$$\Sigma^{xz} = \begin{pmatrix} \Sigma^x & c_{ye}\cdot c_{xy}\cdot c_{ez}\cdot\Sigma^x\cdot(F\cdot K)^\top \\ c_{ye}\cdot c_{xy}\cdot c_{ez}\cdot F\cdot K\cdot\Sigma^x & \Sigma^z \end{pmatrix}$$

$$\Sigma^{yz} = \begin{pmatrix} \Sigma^y & c_{ye}\cdot c_{ez}\cdot\Sigma^y\cdot F^\top \\ c_{ye}\cdot c_{ez}\cdot F\cdot\Sigma^y & \Sigma^z \end{pmatrix} \tag{15}$$

$$\Sigma^{ez} = \begin{pmatrix} \Sigma^e & c_{ez}\cdot\Sigma^e \\ c_{ez}\cdot\Sigma^e & \Sigma^z \end{pmatrix}$$

Similarly, the covariance matrices of *three* and *four* concatenated vectors that have not been given in Section 3.1 are:

$$\Sigma^{xyz} = \begin{pmatrix} \Sigma^x & c_{xy} \cdot \Sigma^x \cdot K^\top & c_{ye} \cdot c_{xy} \cdot c_{ez} \cdot \Sigma^x \cdot (F \cdot K)^\top \\ c_{xy} \cdot K \cdot \Sigma^x & \Sigma^y & c_{ye} \cdot c_{ez} \cdot \Sigma^y \cdot F^\top \\ c_{ye} \cdot c_{xy} \cdot c_{ez} \cdot F \cdot K \cdot \Sigma^x & c_{ye} \cdot c_{ez} \cdot F \cdot \Sigma^y & \Sigma^z \end{pmatrix}$$

$$\Sigma^{xez} = \begin{pmatrix} \Sigma^x & c_{xy} \cdot c_{ye} \cdot \Sigma^x \cdot (F \cdot K)^\top & c_{ye} \cdot c_{xy} \cdot c_{ez} \cdot \Sigma^x \cdot (F \cdot K)^\top \\ c_{xy} \cdot c_{ye} \cdot F \cdot K \cdot \Sigma^x & \Sigma^e & c_{ez} \cdot \Sigma^e \\ c_{ye} \cdot c_{xy} \cdot c_{ez} \cdot F \cdot K \cdot \Sigma^x & c_{ez} \cdot \Sigma^e & \Sigma^z \end{pmatrix} \tag{16}$$

$$\Sigma^{xyez} = \begin{pmatrix} \Sigma^x & c_{xy} \cdot \Sigma^x \cdot K^\top & c_{ye} \cdot c_{xy} \cdot \Sigma^x \cdot (F \cdot K)^\top & c_{ye} \cdot c_{xy} \cdot c_{ez} \cdot \Sigma^x \cdot (F \cdot K)^\top \\ c_{xy} \cdot K \cdot \Sigma^x & \Sigma^y & c_{ye} \cdot \Sigma^y \cdot F^\top & c_{ye} \cdot c_{ez} \cdot \Sigma^y \cdot F^\top \\ c_{ye} \cdot c_{xy} \cdot F \cdot K \cdot \Sigma^x & c_{ye} \cdot F \cdot \Sigma^y & \Sigma^e & c_{ez} \cdot \Sigma^e \\ c_{ye} \cdot c_{xy} \cdot c_{ez} \cdot F \cdot K \cdot \Sigma^x & c_{ye} \cdot c_{ez} \cdot F \cdot \Sigma^y & c_{ez} \cdot \Sigma^e & \Sigma^z \end{pmatrix}$$

Given the matrices in Eqs. 14-16, in *Model II* both variables T and I depend on the intra-cortical connectivity c_{ez} and on the feedback c_{zx}. However, the sensitivity of the descriptors is not that obvious from these equations plugged into Eqs. 6 and 7. Therefore, in order to figure out which descriptor is better (which one is more sensitive) one should consider specific values of the parameters (as for instance what we considered in Section 2), and compute T and I in a range of connectivity values.

We do that in the next experimental section where we find that, in *Model II*, our descriptor, T, is substantially more sensitive than I to the feedback, c_{zx}, and the intra-cortical connectivity, c_{ez}. And this happens both for Gaussian signals and also for natural images.

4 Empirical results

In this experimental section[3] we address the following points:

- We use the theoretical expressions to illustrate the behaviors of T and I, both in the case where the superiority of T is analytically obvious (as in Eqs. 9-11 versus Eqs. 13 for the intra-cortical connectivity in *Model I*), and in the case where the behavior is not easy to see directly from Eqs. 14-16 plugged into Eqs. 6-7 (in Model II). In these experiments we use Gaussian sources with the same mean and covariance as natural images and the model parameters discussed in Section 2.4.
- We confirm the theoretical results presented in Section 3 for both models (*I* and *II*) through a specific empirical estimator of T and I [12, 59] that has been already used in visual neuroscience [10, 11]. This empirical confirmation of the theory uses sets of $0.5 \cdot 10^5$ Gaussian samples injected into the models (*I* and *II*), and then, the empirical estimator is applied to the responses of the models. Incidentally, the presented pair *theory-data* is a good test-bed for empirical estimators of T and I.
- We explore how the empirical estimations of T and I behave for natural (non-Gaussian) images where, in principle, the theory would not be applicable. We also use sets of $0.5 \cdot 10^5$ natural image patches and the same variations of *Model I* and *Model II*.
- We explore the behavior of T and I in real fMRI signals from cortical regions V1, V2, V3, V4 responding to natural images so that we can discuss possible connectivity schemes.

The structure of this section is as follows: (1) We describe the experimental methods: the empirical estimator, the natural and synthetic image data, and computational issues of the theoretical expressions. (2) We present T and I surfaces for different intra-cortical connectivity c_{ez} and α_H that controls H in *Model I*. (3) We present T and I surfaces for different feedforward and feedback connectivity c_{ez} and c_{zx} in *Model II*. Finally, (4) we present the empirical estimations of T and I from real fMRI recordings.

4.1 Methods: empirical estimator, data, and computational issues

Empirical estimation of T and I from samples: here we use the *Rotation-Based Iterative Gaussianization* (RBIG). This method, originally proposed for PDF estimation [12], is able to transform data following any multivariate PDF

[3] Code and data at `http://isp.uv.es/docs/CODE_connectivity.zip`, `Samples.tar.gz`, and `DATA_connect_2.zip`

into data that follows a unit-covariance multivariate Gaussian. In this way, RBIG is useful to estimate the redundancy among coefficients because it accumulates the variations in redundancy while transforming the original dataset into the final Gaussian dataset where all coefficients are independent. The advantages of RBIG with regard to other information estimators [16, 17] has been shown in [11, 59, 60]. RBIG has also been used in visual neuroscience to check the *Efficient Coding Hypothesis* in Wilson-Cowan networks [10], in Divisive Normalization networks [11], and in color appearance networks [61]. However, any other empirical estimator of T and I from samples [13–17] could be used in the experiments below.

Natural and synthetic data: In the experiments we used $0.5 \cdot 10^5$ image patches of size 8×8, i.e. $n = 64$, randomly taken from the luminance component of two colorimetrically-calibrated datasets: the IPL dataset [62, 63], and the Barcelona dataset [64]. In the IPL dataset only images under the CIE D65 (daylight-like) illuminant were considered. The two datasets were linearly scaled so that the average luminance in both was equal to 40 cd/m^2. This separate global normalization ensures that image patches from both sets are equivalent and can be safely mixed. Then, we randomly extracted the samples $0.25 \cdot 10^5$ from each dataset, and we computed the covariance from this joint set of $0.5 \cdot 10^5$ samples: see Σ^s in Fig. 4. This matrix, Σ^s, is the starting point of all the theoretical results presented in Section 3. Our data has the classical covariance of the luminance in natural images (see for instance [65]), which is diagonalized by DCT-like basis functions (see Fig. 4, consistently with [25, 63, 66]). Then, we generated $0.5 \cdot 10^5$ Gaussian vectors of dimension $n = 64$ with the mean and covariance of the natural samples. Of course, both sets (natural and synthetic) are not the same (as can be seen in Fig. 4, consistently with [18, 19]). Then, we inject the synthetic and natural samples through *Model I* and *Model II* to get the corresponding responses $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{e}$, and $\boldsymbol{z}$, for the range of connectivity values considered in Section 2.

Computational issues: All the analytical results (e.g. Eq. 9) depend on the computation of determinants of large matrices (either covariance matrices or the Jacobian $\nabla_{\boldsymbol{e}} f^\top \cdot \nabla_{\boldsymbol{e}} f$). The computation of determinants in high-dimensional scenarios is very prone to divergences to 0 or ∞. Therefore, it is better to avoid its computation: given the fact that the

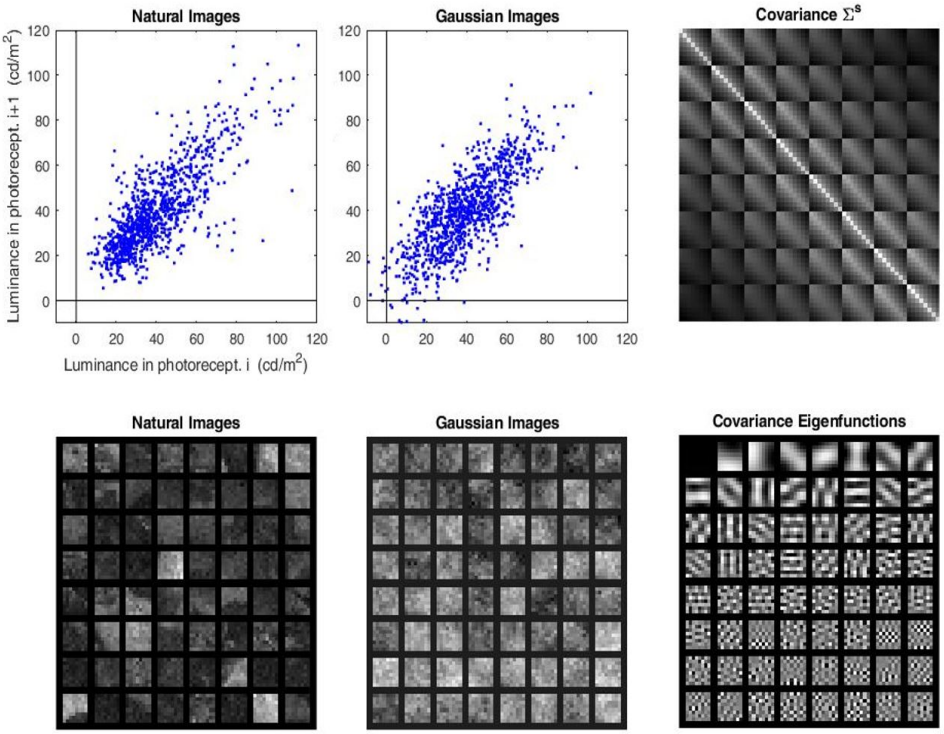


Figure 4: Natural and synthetic image data (the source $\boldsymbol{s}$). The bottom-left mosaic shows illustrative samples from the colorimetrically-calibrated databases IPL and Barcelona. The top-left scatter plot illustrates the joint PDF of the luminance at neighbor photoreceptors. Images and scatter plot show the (non-Gaussian) bias towards low-luminance, and the spatial smoothness of the signal (predominance of low spatial frequency). The non-diagonal nature of the covariance matrix (at the top-right) captures the spatial smoothness, and its eigenfunctions (bottom-right) are similar to the frequency analyzers in the cortex models (in Fig. 1). The order of the functions according the eigenvalue confirms the low-frequency nature of the signal. The central column shows Gaussian samples with the same mean and covariance.

considered matrices, A, are symmetric (either Σ or $\nabla_{\boldsymbol{e}} f^{\top} \cdot \nabla_{\boldsymbol{e}} f$), they are diagonalizable by an orthonormal transform (with unit determinant). Therefore, it holds $log|A| = \sum_{i=1}^{d} log(\lambda_i)$ where λ_i are the eigenvalues of A (whatever the dimension $d \times d$ of the matrix A). Note that this sum is more robust than the naive computation of the determinant.

4.2 Results for I and T in terms of nonlinear intra-cortical connectivity (*Model I*)

Figure 5 shows the results of Mutual Information for different intra-cortical connectivity scenarios in the nonlinear *Model I*. Specifically, we show (a) the theoretical results for Gaussian signals, (b) the empirical results computed with RBIG for Gaussian signals, and (c) the empirical results computed with RBIG for natural signals.

We see two basic trends in the results (both in the theory and in the empirical estimations):

1. $I(\boldsymbol{x}, \boldsymbol{y}) \approx I(\boldsymbol{x}, \boldsymbol{z}) \ll I(\boldsymbol{y}, \boldsymbol{z})$. This could be expected because the shared information is reduced with the noise introduced in each layer and $\sigma(\boldsymbol{n_y}) \gg \sigma(\boldsymbol{n_e})$, and no noise is introduced in $\boldsymbol{z}$, i.e. $f(\cdot)$ is invertible. Therefore, more information is lost between $\boldsymbol{x}$ and inner layers (either $\boldsymbol{y}$ or $\boldsymbol{z}$), than the information lost between $\boldsymbol{y}$ and $\boldsymbol{z}$, which have an almost invertible relation: only a small fraction of bits is lost due to $\boldsymbol{n_e}$.
2. More important for the description of connectivity is the fact that (as predicted by the theory), Mutual Information is *totally insensitive* to the differences in intra-cortical connectivity. Therefore, this pairwise measure is not a good descriptor of connectivity for this kind of nonlinearity.

It is important to note that these global trends in the theory are consistently confirmed by the empirical estimations. Beyond a small bias (overestimation) in I_{RBIG}, it identifies the substantially bigger connection between $\boldsymbol{y}$ and $\boldsymbol{z}$ rather than between $\boldsymbol{x}$ and inner layers. Moreover, I_{RBIG} is also constant over the range of nonlinear connectivity values.

Interestingly, the empirical results for natural images also follow these trends even though the signals are no longer Gaussian. In this case, the non-Gaussianity only introduces a reduction in the I_{RBIG} estimates and a small variation over the explored models, which is negligible in terms of describing changes in the connectivity.

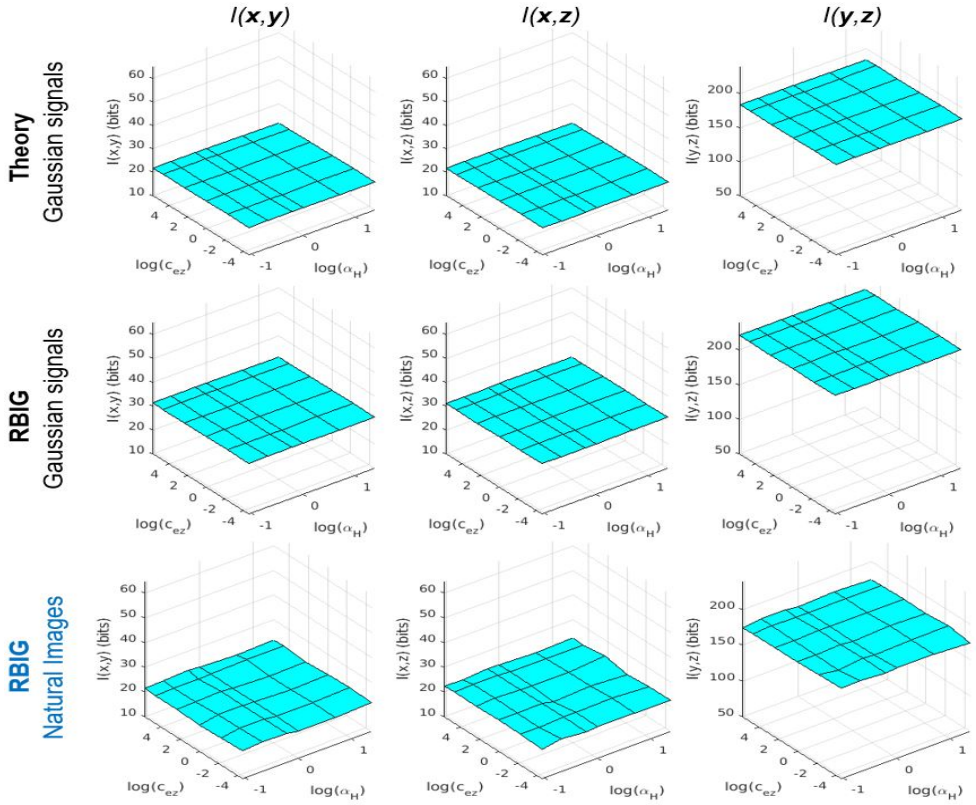


Figure 5: Mutual Information does not describe intra-cortical connectivity in *Model I*. Plots of I as a function of intra-cortical connectivity for Gaussian signals (theory and RBIG estimates), and empirical results for natural images.

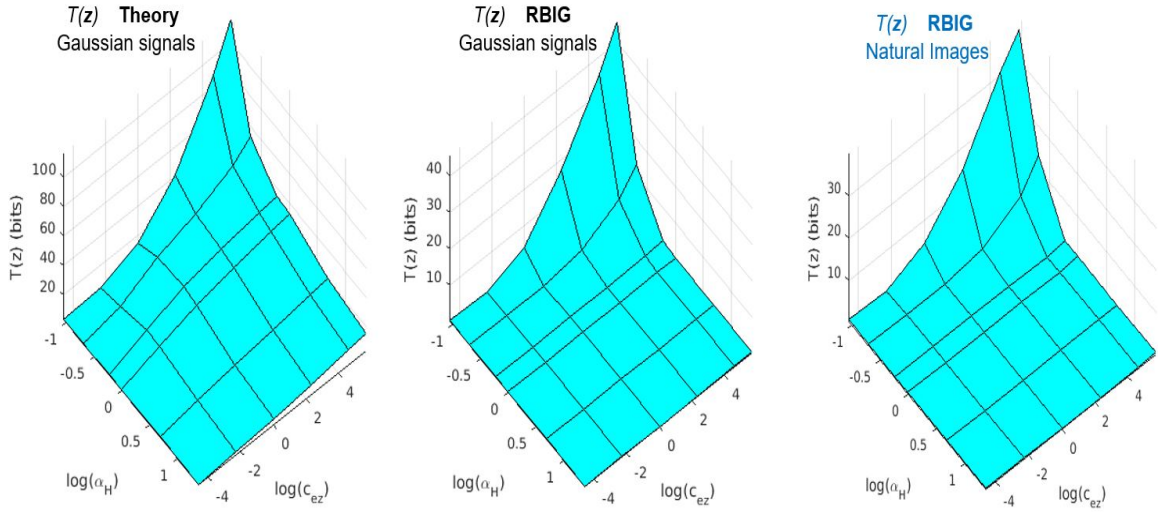


Figure 6: Total Correlation does capture variations in intra-cortical connectivity in *Model I*. Plots of $T(\boldsymbol{z})$ as a function of intra-cortical connectivity for Gaussian signals (theory and RBIG estimates), and empirical results for natural images.

Figure 6 shows the part of $T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ that depends on the nonlinear connectivity: $T(\boldsymbol{z})$ according to Eq. 11. In this case, as opposed to I, the *Total correlation* strongly depends on the intra-cortical connectivity.

Again, (beyond a subestimation bias in RBIG) the general trend of the empirical estimations over the connectivity range confirms the theoretical predictions. The non-Gaussianity of natural signals does not introduce major deviations in the trend of the surface.

A technical comment on the estimation of $T(\boldsymbol{z})$: as the variables $\boldsymbol{z} = f(\boldsymbol{e})$ are non-Gaussian, and this non-Gaussianity is particularly strong in some regions of the explored domain of connectivity, it is important to use a large number of iterations in the Gaussianization algorithm to get a good estimate of T. In particular here we used 500 iterations.

4.3 Results for I and T in terms of feedforward and feedback connectivity (*Model II*)

Figure 7 shows the results of Mutual Information for different feedforward and feedback connectivity scenarios: different combinations of c_{ez} and c_{zx} in *Model II*. Specifically, we show: (a) the theoretical results for Gaussian signals, (b) the empirical results computed with RBIG for Gaussian signals, and (c) the empirical results computed with RBIG for natural signals.

As in the recurrent *Model II* the interpretation of the analytical results is more complicated, in this section each surface is given in a relative scale with regard to its maximum so that the sensitivity of the different descriptors can be fairly compared. Moreover, the variation of the descriptor, Δ, both in percentage and in bits, is also given. As the explored range is the same for every descriptor, Δ is a good measure of the sensitivity to the considered variation of the connectivity.

In *Model II*, due to the top-down feedback signal, $\boldsymbol{x} \xleftarrow{c_{zx}} \boldsymbol{z}$, in principle, all the layers can be affected by an enhanced transmission at a deep layer such as, $\boldsymbol{e} \xrightarrow{c_{ez}} \boldsymbol{x}$. Specifically, the results show the following trends:

1. In general, the shared information increases with c_{ez}. This is obvious in the cases where $\boldsymbol{z}$ is one of the considered nodes (e.g. the last three columns $I(\boldsymbol{x}, \boldsymbol{z})$, $I(\boldsymbol{y}, \boldsymbol{z})$ or $I(\boldsymbol{e}, \boldsymbol{z})$) because an increased c_{ez} improves the presence of the source in the inner representation. More interestingly, we can see that when $\boldsymbol{z}$ is not considered, the effect of c_{ez} is only relevant when there is also significant feedback (as in the two first columns $I(\boldsymbol{x}, \boldsymbol{y})$ and $I(\boldsymbol{x}, \boldsymbol{e})$). This is also the case when considering nodes that are far away, as in $I(\boldsymbol{x}, \boldsymbol{z})$.
2. When considering nodes that are far from the considered interactions (e.g. $\boldsymbol{y}$ and $\boldsymbol{e}$) the descriptor is insensitive to the variations of connectivity (see $I(\boldsymbol{y}, \boldsymbol{e})$ in the third column).
3. In summary, the average percentage of variation of the measures based on I in the theoretical expressions is $\Delta_I = 47 \pm 30\ \%$.

For *Model II* the global trends in the theory are consistently confirmed by the empirical estimations. Similarly to what we found in *Model I*, parallel results in the theory and empirical estimates also form *Model II* confirm the correctness of the theory and the appropriateness of RBIG in this application.

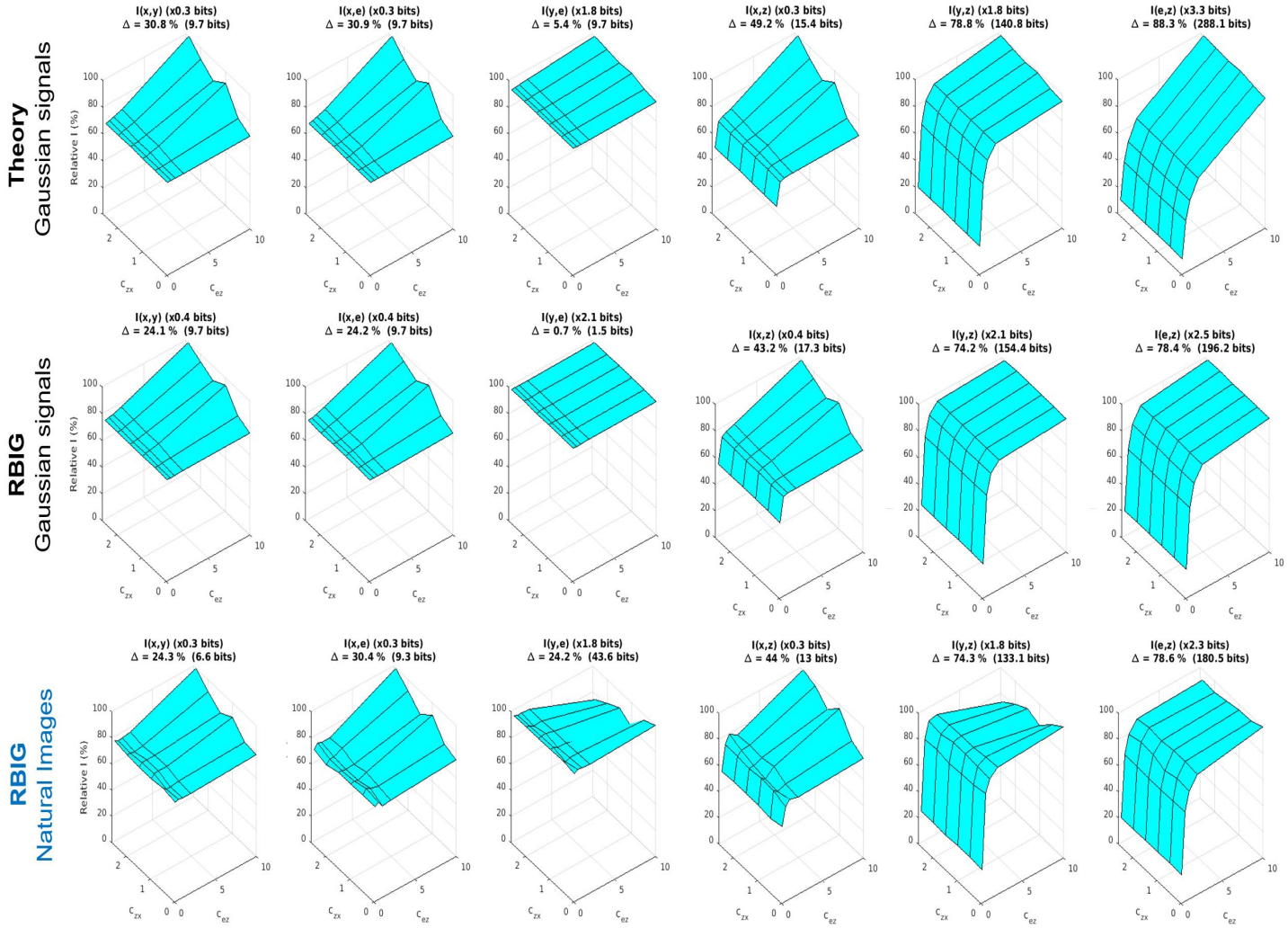


Figure 7: Mutual Information has mild dependence with feedforward and feedback connectivity in *Model II*. Plots of I as a function of the feedforward connectivity, c_{ez}, and feedback, c_{zx}, for Gaussian signals (theory and RBIG estimates), and empirical results for natural images. The plots display relative values of I in percentage with regard to the maximum together with a factor (e.g. $\times 0.3$ in the top-left plot) that allows to express this percentage in absolute values (in bits). Moreover, the plots display the variation (in bits) of the considered descriptor over the range of connectivity values (e.g. $\Delta = 9.7$ bits in the top-left plot). This is a measure of the sensitivity of the descriptor.

Moreover, the empirical results for natural images also follow these trends even though the signals are no longer Gaussian. In this case, the non-Gaussianity only introduces a noticeable variation in $I(\boldsymbol{y}, \boldsymbol{e})$. However, this does not change much the global sensitivity of I, as described by Δ.

Figure 8 shows the results of Total Correlation for different feedforward and feedback connectivity scenarios: different combinations of c_{ez} and c_{zx} in *Model II*. As above, we show: (a) the theoretical results for Gaussian signals, (b) the empirical results computed with RBIG for Gaussian signals, and (c) the empirical results computed with RBIG for natural signals. Here we also present the T surfaces in relative scale for a simpler comparison with the I surfaces in Fig. 7.

The overall percentage of variation of the measures based on T in the theoretical expressions is $\Delta_T = 75 \pm 11$ %.

In the recurrent *Model II*, the sensitivity of T to connectivity and feedback is stronger than the sensitivity of I. Note that $\Delta_T > \Delta_I$ with substantially lower variance over the considered nodes. Therefore, T is more appropriate than I to describe the connectivity in the recurrent *Model II*.

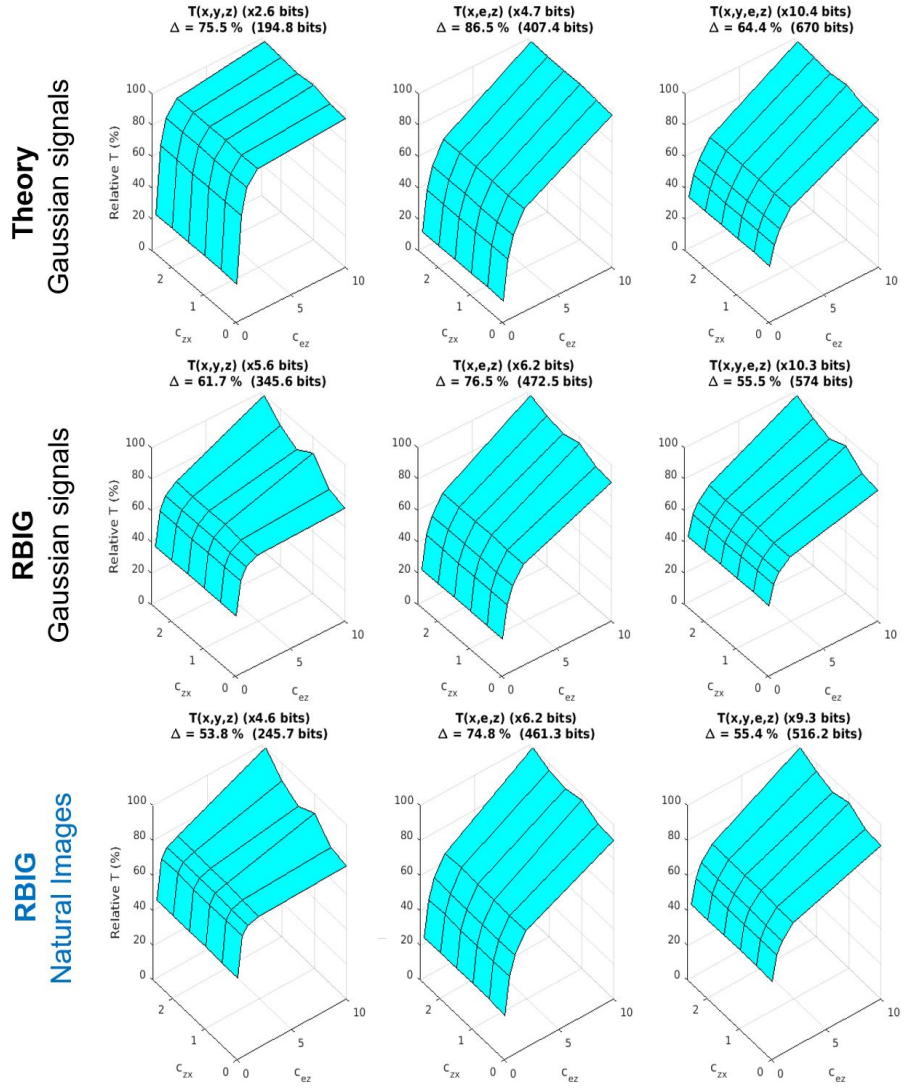


Figure 8: Total Correlation strongly depends on the feedforward and feedback connectivity in *Model II*. Plots of T as a function of the feedforward connectivity, c_{ez}, and feedback, c_{zx}, for Gaussian signals (theory and RBIG estimates), and empirical results for natural images. The plots display relative values of T in percentage with regard to the maximum together with a factor (e.g. $\times 2.6$ in the top-left plot) that allows to express this percentage in absolute values (in bits). Moreover, the plots display the variation (in bits) of the considered descriptor over the range of connectivity values (e.g. $\Delta = 195$ bits in the top-left plot). This is a measure of the sensitivity of the descriptor.

Moreover, as in the previous examples, the empirical estimates from the samples confirm the general trends of the theory, and the results for natural signals approximately follow the results for Gaussian sources.

4.4 Results with real fMRI signals from visual regions V1, V2, V3 and V4

Here we measure the information shared by different visual regions the visual cortex starting from V1. This represents an interesting application because (1) there is a debate on how these regions actually interact [67–71], and (2) there is a long-standing concept in visual neuroscience that relates neural connectivity with information transmission: the *Efficient Coding Hypothesis*, where redundancy reduction has a central role [72, 73]. Specifically, here we take neural data from the *Algonauts Project 2021 challenge* [22], and we consider fMRI signals from V1, V2, V3 and V4 while the observers were looking at natural videos. In our experiments we consider pairwise and multivariate relations among

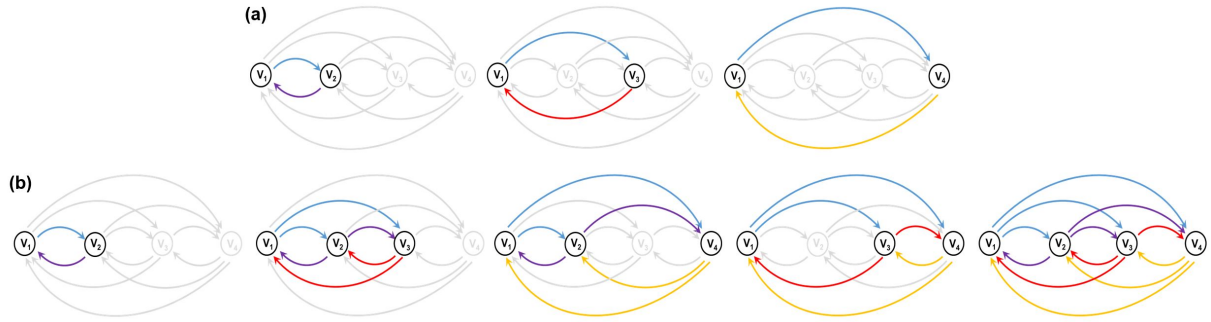


Figure 9: Examples of the (a) pairs of nodes, and (b) groups of nodes, that we consider in our measurements. The top row (a) considers the relation between certain node (in this case V1), and, from left to right, nodes progressively distant: V2, V3, and V4. The bottom row (b), from left to right, adds nodes to the group under consideration. If we start from V_i, the added nodes can be close to it, e.g. (V_i, V_{i+1}, V_{i+2}) in the 2nd diagram, or they can be progressively farther away, as (V_i, V_{i+1}, V_{i+3}) or (V_i, V_{i+2}, V_{i+3}), as in the 3rd and 4th diagrams. Finally, the last diagram at the right shows that we can consider all nodes at the same time, namely (V_1, V_2, V_3, V_4).

regions which (anatomically) are progressively farther away. However, our descriptors of functional links do not make any prior assumption of the possible feedforward or feedback connections.

Ensembles: The considered dataset provides 3 responses of 9 observers for 1000 natural videos in a number of voxels of the considered regions (V1, V2, V3 and V4). In this database there is a one-to-one relation between input and responses, but the number of available voxels depends on the observer and the cortical region. Therefore, just for illustrative purposes, we take 20 randomly selected voxels per region for each observer. This means 20-dimensional signals associated to one input. By considering the data of all trials, all observers, and all input videos, we have $3 \times 9 \times 1000 = 27000$ samples of these 20-dimensional vectors for each region. In these ensembles, the i-th vector of each region corresponds to the same input and the same observer, but the j-th dimension of the vector is the response of a randomly chosen voxel in that region (and observer). We assume all the observers and all the voxels in a region are equivalent. By rerunning this random selection of voxels we get equivalent ensembles.

Empirical estimation: Given the fact that the marginal PDFs of the signals are approximately Gaussian (results not shown), in the estimations of T and I based on iterative Gaussianization we chose only 20 iterations (as opposed to the 500 iterations used in *Model I* where z is non-Gaussian). We re-estimate T and I from equivalent, randomly chosen, ensembles 30 times and we report the average and standard deviation of the results.

Measurements of functional links: we measured I and T in all possible distinct combinations of nodes. Figure 9 illustrates pairwise and multivariate relations among regions which (anatomically) are progressively farther away. Note that the functional link of the configurations in the top row can be addressed by the pairwise $I(\boldsymbol{v_i}, \boldsymbol{v_j})$ or $T(\boldsymbol{v_i}, \boldsymbol{v_j})$. However, progressive consideration of additional nodes, as in the bottom row, can only be quantified using a multivariate descriptor $T(\boldsymbol{v_i}, \boldsymbol{v_j}, \boldsymbol{v_k}, \ldots)$. Note that in a case where the connections are unknown, the shared information (either I or T) is not only affected by the *direct* connections between the considered nodes (in our figure *direct* connections are in color), but also by all other possible *indirect* connections (depicted in gray). The *indirect* connections imply communication through alternative regions that may re-inject the relevant signal into the considered nodes and have a positive effect in the functional link.

On top of the two-node and multi-node cases, mono-mode references are convenient to know if the information is lost through the network or, on the contrary, there are positive synergies. To this end, we report three additional numbers: $T(\boldsymbol{v_i})$, which is a measure of the redundancy within the node $\boldsymbol{v_i}$; and also $I(\boldsymbol{v_i}, \boldsymbol{v_i})$, and $T(\boldsymbol{v_i}, \boldsymbol{v_i})$. In principle, the information shared by a variable with itself, as in $I(\boldsymbol{v_i}, \boldsymbol{v_i})$, and $T(\boldsymbol{v_i}, \boldsymbol{v_i})$, is ∞ [4]. However, given the uncertainty we introduce when using random voxels from each region/observer, two (randomly chosen) sets of $\boldsymbol{v_i}$ are not aligned and then $I(\boldsymbol{v_i}, \boldsymbol{v_i})$, and $T(\boldsymbol{v_i}, \boldsymbol{v_i})$ do not diverge to ∞. Instead, they are measures of the common information present in every realization of the ensemble of responses of that node $\boldsymbol{v_i}$. Therefore, they are a convenient reference to know if the consideration of extra nodes increases or decreases this mono-mode amount of information.

[4] Given any n-dimensional variable $\boldsymbol{a}$, the samples of $(\boldsymbol{a}, \boldsymbol{a})$ are aligned in a $2n$-dimensional space, and then the joint differential entropy terms of Eqs. 6-7 is $-\infty$, leading to $I(\boldsymbol{a}, \boldsymbol{a}) = T(\boldsymbol{a}, \boldsymbol{a}) = \infty$.

For a more intuitive comparison of the results corresponding to configurations with different number of nodes, we report the shared information *per node*. This means: $I(\boldsymbol{v_i}, \boldsymbol{v_j})/2$, $T(\boldsymbol{v_i}, \boldsymbol{v_j})/2$, $T(\boldsymbol{v_i}, \boldsymbol{v_j}, \boldsymbol{v_k})/3$, and, $T(\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}, \boldsymbol{v_4})/4$. In the case of $T(\boldsymbol{v_i})$ the definition already has a single node, so *bits* and *bits/node* are the same.

Finally, we report not only the absolute values in *bits/node*, but (more interestingly to describe the connectivity) how the information per node increases or decreases when we go way from one node or include progressively distant nodes in the measure. We give this deviation in $\%$ with regard to the information per node in V1 (either $I(\boldsymbol{v_1}, \boldsymbol{v_1})$ or $T(\boldsymbol{v_1}, \boldsymbol{v_1})$).

Results: Tables 1-2 show the measures of shared information in three panels: the top panel shows the pair-wise measures $I(\boldsymbol{v_i}, \boldsymbol{v_j})$, the middle panel shows the single-node measure $T(\boldsymbol{v_i})$, and the bottom panel shows the multi-node measures $T(\boldsymbol{v_i}, \boldsymbol{v_j}, \ldots)$. Table 1 has absolute measures in *bits/node*, and Table 2 displays the variation (in $\%$) of the considered configuration with regard to the corresponding measure in V1. The $T(\boldsymbol{v_i}, \boldsymbol{v_j}, \ldots)$ panels have a pair-wise part (at the left) and a multi-node part (the last four columns). This multi-node parts have to be read *row-wise*: each number reports how the node in the row interacts with the nodes in the different columns. Moreover, the consideration of extra nodes is done in cyclic way: in the 3rd row $v_i = v_3$, and hence the 5th column, $(v_{i+1}, v_{i+2}) = (v_4, v_1)$, refers to the connectivity among the nodes (v_3, v_4, v_1).

Not all the values in the tables are independent because of the symmetry of the measures. Note that I and T are invariant to the permutation of the variables: $I(\boldsymbol{v_i}, \boldsymbol{v_j}) = I(\boldsymbol{v_j}, \boldsymbol{v_i})$, and $T(\boldsymbol{v_i}, \boldsymbol{v_j}, \boldsymbol{v_k}) = T(\boldsymbol{v_j}, \boldsymbol{v_k}, \boldsymbol{v_i}) = \ldots$ This implies that the I panels are symmetric and so it is the pairwise part of the T panels. Also as a consequence of the invariance to permutation, some multi-node configurations are equivalent. As the order does not matter, we have combinations of 4 nodes taken 3 at a time, i.e. only 4 independent node configurations. For the sake of clarity the non-redundant values of the tables are highlighted in blue. Also for clarity, the standard deviation over the 30 realizations of the estimation has been reported only in the independent values of Table 1.

The discussion of the results will be focused on the variations of information as we depart from a node (Table 2). Departure, as in the top row of Fig. 9, means *moving away from the diagonal* (along rows/columns) in the pairwise parts of the tables. Departure, as in the bottom row of Fig. 9, means *moving to the right (for the highlighted numbers)* in the multi-node parts. Table 1, with the original absolute measures, is just given for completeness and for the reader convenience.

A final comment on the absolute magnitudes: in every case, the estimated $T(\boldsymbol{v_i}, \boldsymbol{v_j}) > I(\boldsymbol{v_i}, \boldsymbol{v_j})$, which is consistent with the definitions because (as discussed in Eqs. 6-7) T includes the redundancy within the nodes and hence the information is necessarily bigger.

Information flow and conjectures on connectivity: Results show that the redundancy within each node $T(\boldsymbol{v_i})$ is smaller in deeper layers than in V1 (see the negative increments in the middle panel of Table 2). This is consistent with the *Efficient Coding Hypothesis* [72, 73].

$\mathbf{I}(\mathbf{v_i}, \mathbf{v_j})$ (in *bits/node*)	v_1	v_2	v_3	v_4
v_1	**2.4** $\pm$ 0.3	**1.3** $\pm$ 0.2	**1.0** $\pm$ 0.2	**0.8** $\pm$ 0.1
v_2	1.3	**2.0** $\pm$ 0.2	**1.2** $\pm$ 0.2	**0.7** $\pm$ 0.1
v_3	1.0	1.2	**1.7** $\pm$ 0.3	**0.8** $\pm$ 0.1
v_4	0.8	0.7	0.8	**2.2** $\pm$ 0.3

$\mathbf{T}(\mathbf{v_i})$ (in *bits/node*)	v_1	v_2	v_3	v_4
	3.6 $\pm$ 0.3	**3.2** $\pm$ 0.2	**3.0** $\pm$ 0.2	**3.5** $\pm$ 0.3

$\mathbf{T}(\mathbf{v_i}, \mathbf{v_j}, \ldots)$ (in *bits/node*)	v_1	v_2	v_3	v_4	v_{i+1}, v_{i+2}	v_{i+1}, v_{i+3}	v_{i+2}, v_{i+3}	$v_{i+1}, v_{i+2}, v_{i+3}$
v_1	**6.0** $\pm$ 0.3	**5.0** $\pm$ 0.2	**4.7** $\pm$ 0.2	**4.6** $\pm$ 0.3	**6.1** $\pm$ 0.2	**5.9** $\pm$ 0.3	**5.7** $\pm$ 0.3	**6.6** $\pm$ 0.2
v_2	5.0	**5.4** $\pm$ 0.3	**4.7** $\pm$ 0.3	**4.5** $\pm$ 0.3	**5.7** $\pm$ 0.3	6.1	5.9	6.6
v_3	4.7	4.7	**5.1** $\pm$ 0.2	**4.5** $\pm$ 0.2	5.7	5.7	6.1	6.6
v_4	4.6	4.5	4.5	**5.9** $\pm$ 0.3	5.9	5.7	5.7	6.6

Table 1: $I(\boldsymbol{v_i}, \boldsymbol{v_j})$ between pairs of areas, $T(\boldsymbol{v_i})$ in each area, and $T(\boldsymbol{v_i}, \boldsymbol{v_j}, \ldots)$ among multiple areas in *bits/node*. See the ***Results*** paragraph in the text for the interpretation of pairs and triplets with progressively distant nodes.

Reduction in $T(\boldsymbol{v_i})$ in the middle panel is not the same as the reductions of $T(\boldsymbol{v_i}, \boldsymbol{v_i})$ or $I(\boldsymbol{v_i}, \boldsymbol{v_i})$ along the diagonal of the pairwise parts of the top and the bottom panels. While redundancy reduction in $T(\boldsymbol{v_i})$ means better information encoding, reduction in $T(\boldsymbol{v_i}, \boldsymbol{v_i})$ or $I(\boldsymbol{v_i}, \boldsymbol{v_i})$ means a decay in the information content. This decay is more apparent in $I(\boldsymbol{v_i}, \boldsymbol{v_i})$, because the reduction of $T(\boldsymbol{v_i}, \boldsymbol{v_i})$ is biased by the simultaneous reduction of the intra-node redundancy in $T(\boldsymbol{v_i})$. Actually, if we discount $\Delta T(v_i)$ from $\Delta T(v_i, v_i)$, the corrected $T(v_i, v_i)$ is more constant[5]. This smaller reduction in the information content may be a positive effect of connectivity seen in T and not in I.

However, the mono-node measures mentioned above only describe the information in each node, but not how much of this information comes from another region. This second concept, more related to connectivity, is measured by pairwise and multi-node measures. In this regard, progressively bigger reductions in the pairwise $\Delta I(\boldsymbol{v_i}, \boldsymbol{v_j})$ and $\Delta T(\boldsymbol{v_i}, \boldsymbol{v_j})$ away from the diagonal mean information loss along the way (or reduced functional connectivity). This information loss seems consistent with the *data processing inequality* [5] to a certain extent. However, as discussed below, the results (particularly T in multiple nodes) confirm the existence of relevant feedback in these regions.

The *data processing inequality* [5] states that information lost between two nodes cannot be recovered by further processing (with no additional input from the original node). This inequality strictly holds in purely feedforward schemes $\boldsymbol{v_1} \rightarrow \boldsymbol{v_2} \rightarrow \boldsymbol{v_3} \rightarrow \boldsymbol{v_4}$, where, due to the absence of feedback connections and skip connections, the response in inner layers conditioned to the previous layer is independent of the early layers. In such systems, it holds $I(\boldsymbol{v_1}, \boldsymbol{v_2}) > I(\boldsymbol{v_1}, \boldsymbol{v_3}) > I(\boldsymbol{v_1}, \boldsymbol{v_4})$. This behavior is what is observed in the rows of the I panel when moving away from the diagonal to inner layers. This suggest that the feedforward component of the connectivity can be strong, and in such simplistic situation, one could deduce the strength of each connection from the different decays in $I(v_i, v_j)$.

However, in our case (where feedback and skip connections may exist) the *data processing inequality* may not hold. Reductions in I do not necessarily mean that the other connections are not present. This is more clear looking at the results of T. While the behavior of the pairwise T moving to deeper layers is negative (similarly to I), something different happens by considering extra nodes. Under the purely feedforward assumption extra nodes should share less information with the previous and the global T should decrease, particularly if the intra-node redundancy does not increase (as in this pathway). However, we see that in some cases the consideration of extra nodes implies an increase of the shared information per node, as for instance when going from (v_1, v_2) to (v_1, v_2, v_3) or from there to (v_1, v_2, v_3, v_4) (see the positive increments highlighted in blue in Table 2).

Multi-node results obtained from the proposed measure T are interesting because we can see that the connections in the group (v_1, v_2, v_3) are substantially stronger than the connections in the group (v_2, v_3, v_4) despite they are at similar anatomical distance. This suggests some top-down feedback from v_3 or v_2 or feedforward skip connections from v_1 to v_3. The same is true when considering all the nodes together with a substantial increment (by 11%).

These two different synergistic behaviors that can be seen using the proposed *Total Correlation* clearly mean that one can rule out a pure feedforward scheme in the V_1, V_2, V_3, V_4 regions, and more complex connectivity schemes do exist. This is not that obvious just using the conventional I.

5 Discussion and conclusions

Analytical results: T is a better descriptor of connectivity than I. The goal of this paper is addressing the fundamental limitation of the seminal work that proposed T as a measure of functional connectivity [3]: namely the lack of analytical results that can justify the superiority of the T over the conventional I beyond the multivariate versus pairwise definitions. Here we did that analytical study in the context of the early visual brain with simple models of the retina-V1 cortex pathway.

For mathematical convenience we considered two variations of the general framework presented in Eq. 2: *Model I* and *Model II*. These models were chosen to illustrate two fundamental properties of neural architectures in early vision: (1) the Divisive Normalization nonlinearity in *Model I*, in Section 2.1, and (2) an eventual top-down recurrence in *Model II* in Section 2.2.

It is important to stress again that the models are not arbitrary: according to the results in Section 2.4 the nonlinearity in *Model I* is key to improve the explanation of the psychophysics, and the explored range of intra-cortical connectivity actually covers different behaviors (with substantial differences in the explained variance of human data). The top-down connection in *Model II* was not specifically justified, but given the observed behavior of the steady state in $\boldsymbol{e}$, the explored feedback does not reduce substantially the $\rho = 0.7$ result. This indicates that *Model II* has certain biological plausibility, so that it can be used to illustrate the study of recurrent connections. The plausibility of the models and the

[5]Losses ΔT corrected in this way (-8.3%, -12.5%, and 0%, for v2, v3, v4), are smaller than the original values (-8.8%, -14.5%, -0.1%), seen in the diagonal of the pairwise part of the T panel (Table 2).

$\mathbf{\Delta I(v_i, v_j)}$ (in %)	v_1	v_2	v_3	v_4
v_1	**0.0**	**-45.0**	**-55.6**	**-67.0**
v_2	-45	**-16.1**	**-47.7**	**-71.4**
v_3	-55.6	-47.7	**-26.1**	**-67.8**
v_4	-67.0	-71.3	-67.8	**-7.1**

$\mathbf{\Delta T(v_i)}$ (in %)	v_1	v_2	v_3	v_4
	0.0	**-11.1**	**-15.5**	**-2.1**

$\mathbf{\Delta T(v_i, v_j, \dots)}$ (in %)	v_1	v_2	v_3	v_4	v_{i+1}, v_{i+2}	v_{i+1}, v_{i+3}	v_{i+2}, v_{i+3}	$v_{i+1}, v_{i+2}, v_{i+3}$
v_1	**0.0**	**-16.0**	**-21.3**	**-21.7**	**2.3**	**-1.4**	**-3.8**	**11.3**
v_2	-16.0	**-8.8**	**-21.4**	**-23.8**	**-3.9**	2.3	-1.4	11.3
v_3	-21.3	-21.4	**-14.5**	**-24.9**	-3.8	-3.9	2.3	11.3
v_4	-21.7	-23.8	-24.9	**-0.1**	-1.4	-3.8	-3.9	11.3

Table 2: Variations of I and T (in % with regard to V1) when considering progressively distant nodes or adding extra nodes. See the ***Results*** paragraph in the text for the interpretation of pairs and triplets with progressively distant nodes.

generality and relevance of the facts they illustrate (nonlinearities and recurrence) implies that a proper descriptor of functional connectivity should be sensitive to the different variations of the models.

Sections 3.1, 4.2, and 4.3 explicitly show the superiority of T over I in the considered nonlinear and recurrent models. The conclusion of these analytical results (confirmed by the experimental simulations) is that while the conventional *Mutual Information* is not useful to capture the intra-cortical connections in *Model I*, the proposed measure, *Total Correlation*, is quite sensitive to this connectivity. Similarly, the proposed *Total Correlation* is more sensitive than *Mutual Information* to the feedforward and feedback connectivity explored in the recurrent *Model II*. From a general perspective, the considered nonlinearity is ubiquitous in the visual pathway [8, 33, 34, 74, 75]. Therefore, the success of the proposed multivariate *Total Correlation* in describing this connectivity is a substantial advantage with regard to the conventional, pairwise, *Mutual Information*.

Results with real data: T highlights synergies in V_1, V_2, V_3, V_4. The positive results of T (and the corresponding RBIG estimates) in the analytical settings presented above not only address a limitation of [3], but really justify its use in real scenarios. In the case of fMRI data from the visual regions V_1, V_2, V_3, V_4, our measurements of T show that: (1) the redundancy within each layer, $T(v_i)$, is reduced along the way, which is consistent with the *Efficient Coding Hypothesis*, (2) the information content measured through $T(v_i, v_i)$ is more stable along the way than the measures given by $I(v_i, v_i)$, particularly if the inner redundancy is discounted. (3) The variation of the pairwise measures of $I(v_i, v_j)$ seems compatible with the *data processing inequality* in a purely feedforward setting $\boldsymbol{v_1 \to v_2 \to v_3 \to v_4}$, however, (4) the multi-node T shows synergies that rule out the purely feedforward scheme. Moreover, it suggests stronger functional connectivity between the nodes V_1, V_2, V_3 than between V_2, V_3, V_4 despite a similar anatomical distance. All this complex behavior is not easy to see just using the conventional I.

Relations with previous work. Firstly, this is the necessary analytical companion of the proposal of *Total Correlation* to measure connectivity [3]. Then, here we have applied this tool to visual areas extending the works that first used Mutual Information to assess the connectivity between pairs of visual areas [76] or those that measured Mutual Information between V1 and MT (or V5) under Divisive Normalization transforms [77]. The analysis of Mutual Information between progressively deeper visual layers is also related with previous works focused on quantifying the information flow in different nonlinear models of retina-V1 pathway [10, 11], which were restricted to purely feedforward models.

On the other hand, the approach we took here (quantifying the statistical properties of the responses of real brains or psychophysically plausible models) is is related with a body of literature that follows Barlow's *Efficient Coding Hypothesis* in a non-classical direction. Note that the classical direction is *from-statistics-to-biology*: a system optimized for a sensible statistical goal may display biological-like behavior [72, 73]. This is the direction that explained linear receptive fields [18, 31, 32, 63] and sensory nonlinearities [20, 36, 56, 62, 78, 79] from statistics. However, there is literature that reasons in the opposite direction *from-biology-to-statistics*: look at the statistical properties of the responses of biologically plausible systems and you will find statistically interesting behavior. In this regard, redundancy

reduction [21, 37, 80], and efficient information transmission [10, 11, 61, 81] has been found in in real and biologically plausible models. And this is similar to the information-theoretic analysis that we did of real and simulated responses.

Limitations and future research. This study has three main limitations that should be addressed by future research. First (and most important) is the limitation of the analytical examples: they addressed fundamental issues such as nonlinearities and recurrence, but they did it in *separate* examples (*Model I* and *Model II*). Moreover, *Model I* didn't include noise after the divisive normalization so that one could apply the property of the variation of T under deterministic transforms, Eq. 8, and the invariance of I under transforms of one of the variables, Eq. 12. Future research should try to get unified expressions for a general nonlinear and recurrent model with noise at all layers.

The second, more instrumental, limitation is related to the specific empirical estimator of T which is necessary in real scenarios. Here we used our *Rotation-Based Iterative Gaussianization* [12, 59], and it proved to follow the trends of the theoretical surfaces in the analytical scenarios. However, RBIG may suffer from errors when the signals are strongly non-Gaussian with multiple modes separated by low probability regions as may happen after Divisive Normalization (see the PDFs of natural images in [11, 21]). An approximate knowledge of the PDF of the signals is required to set the number of iterations in RBIG. Of course, future research can use other empirical estimators as for instance [13–17]. In this regard, the analytical results presented here are a good test-bed for current or future empirical estimators.

Finally, regarding the results with real data, it is important to acknowledge that there are more comprehensive databases. The one we used (*the Algonauts 2021 Challenge* [22]) only considers 1000 videos and has a restricted set of voxels because we wanted a simple proof of concept for our measure T and estimator RBIG on low level regions. The work done here could be extended in different ways. First, the database could be segmented depending on the properties of the stimuli (e.g. color, texture and motion content) because the functional connectivity between the considered regions may depend on these low-level features of the input. This could tell us about the specialization of these regions in different dimensions of the stimuli. Moreover, the computation of connectivity based on T depending on the structure of the scene could clarify the differences in the feedback signals found in figure-ground contexts [69, 71]. And second, larger databases (such as [82]) may be convenient to confirm the current results and be more appropriate to study the connectivity depending on the properties of the input so that the subsets are big enough to trust the information estimates. Databases like [83] can be used to address the relation between V1 and higher-level regions (FFA, PPA,...).

Conclusions: In this work we derived analytical results that show that *Total Correlation* is a better descriptor of connectivity than *Mutual Information* in plausible models of the retina-LGN-V1 that include nonlinearities due to intra-cortical connectivity and top-down feedback. T is better because it is more sensitive than I to connectivity. Analytical results are derived for Gaussian signals but, as confirmed by empirical estimates, they also hold for natural inputs. Our T results for real responses recorded from V1,V2,V3,V4 rule out a naive feedforward-only information flow and suggest stronger feedback connections in V1,V2,V3, than in V2,V3,V4.

The proposed measure opens several possibilities: (1) it can be applied to assess the connectivity in complex models that have been developed to reproduce feedforward and feedback oscillations [70], and (2) it can be used to examine signal-dependent feedback in stimuli with figure-ground or spatially segregated textures, which is an interesting open question in visual neuroscience [69, 71].

6 Author contributions

QL checked the plausibility of the neural models reproducing the psychophysical data, implemented the experiments to compare Total Correlation and Mutual Information, computed the measures in the fMRI data using RBIG and contributed to the writing of the manuscript. **GVS** made some comments and suggestions on earlier versions of this manuscript, and contributed to the writing of the manuscript. **JM** had the idea of the work, developed the theoretical results, implemented the neural model, and prepared the first draft of the manuscript.

7 Acknowledgements

The authors thank Dr. Valero Laparra for his comments on early stopping of RBIG depending on data dimensionality and Dr. Olga Stefanska for her support in the writing process. We thank the organizers of the *Algonauts Project 2021 Challenge* for providing their interesting fMRI dataset which was used in this study. This work was partially funded by these spanish/european grants from GVA/AEI/FEDER/EU: MICINN PID2020-118071GB-I00, MICINN PDC2021-121522-C21, and GVA Grisolía-P/2019/035 (for JM and QL), and by the Defense Advanced Research Projects Agency (DARPA) under award FA8750-17-C-0106 (for GVS).

References

[1] Karl Friston. "Functional and Effective Connectivity: A Review". In: *Brain connectivity* 1 (Jan. 2011), pp. 13–36.

[2] J.T. Lizier et al. "Multivariate Information-Theoretic Measures Reveal Directed Information Structure and Task Relevant Changes in fMRI Connectivity". In: *J. Comput. Neurosci.* 30.1 (2011), pp. 85–107.

[3] Qiang Li. "Functional connectivity inference from fMRI data using multivariate information measures". In: *Neural Networks* 146 (2022), pp. 85–97.

[4] Satosi Watanabe. "Information theoretical analysis of multivariate correlation". In: *IBM Journal of research and development* 4.1 (1960), pp. 66–82.

[5] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006. ISBN: 0471241954.

[6] M. Carandini and D. Heeger. "Summation and Division by Neurons in Visual Cortex". In: *Science* 264.5163 (1994), pp. 1333–6.

[7] N.C. Rust and J.A. Movshon. "In praise of artifice". In: *Nat. Neurosci.* 8 (2005), pp. 1647–1650.

[8] M. Carandini and D. Heeger. "Normalization as a canonical neural computation". In: *Nat. Rev. Neurosci.* 13.1 (2012), pp. 51–62.

[9] M. Martinez, M Bertalmío, and J. Malo. "In Praise of Artifice Reloaded: Caution with Natural Image Databases in Modeling Vision". In: *Front. Neurosci. doi: 10.3389/fnins.2019.00008* (2019).

[10] A. Gomez-Villa, M. Bertalmio, and J. Malo. "Visual Information Flow in Wilson-Cowan Networks". In: *J. Neurophysiol. doi:10.1152/jn.00487.2019* (2020).

[11] J. Malo. "Spatio-chromatic information available from different neural layers via Gaussianization". In: *J. Math. Neurosci.* 10.18 (2020). DOI: 10.1186/s13408-020-00095-8.

[12] V. Laparra, G. Camps-Valls, and J. Malo. "Iterative gaussianization: from ICA to random rotations". In: *IEEE Trans. Neural Networks* 22.4 (2011), pp. 537–549.

[13] Greg Ver Steeg and Aram Galstyan. "Discovering Structure in High-Dimensional Data Through Correlation Explanation". In: *Advances in Neural Information Processing Systems, NIPS'14*. 2014.

[14] Greg Ver Steeg. "Unsupervised Learning via Total Correlation Explanation". In: *IJCAI*. 2017.

[15] Greg Ver Steeg and Aram Galstyan. "Maximally Informative Hierarchical Representations of High-Dimensional Data". In: *AISTATS'15*. 2015.

[16] Iván Marín-Franch and David H. Foster. "Estimating Information from Image Colors: An Application to Digital Cameras and Natural Scenes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 78–91.

[17] Z. Szabó. "Information Theoretical Estimators Toolbox". In: *Journal of Machine Learning Research* 15 (2014), pp. 283–287.

[18] Bruno A. Olshausen and David J. Field. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". In: *Nature* 381 (1996), pp. 607–609.

[19] Eero P Simoncelli and Bruno A Olshausen. "Natural Image Statistics and Neural Representation". In: *Annual Review of Neuroscience* 24.1 (2001), pp. 1193–1216.

[20] J. Malo and J. Gutiérrez. "V1 non-linear properties emerge from local-to-global non-linear ICA". In: *Network: Computation in Neural Systems* 17.1 (2006), pp. 85–102.

[21] J. Malo and V. Laparra. "Psychophysically tuned divisive normalization approximately factorizes the PDF of natural images". In: *Neural computation* 22.12 (2010), pp. 3179–3206.

[22] Radoslaw Martin Cichy et al. "The Algonauts Project 2021 Challenge: How the Human Brain Makes Sense of a World in Motion". In: *CoRR* abs/2104.13714 (2021). arXiv: 2104.13714. URL: https://arxiv.org/abs/2104.13714.

[23] D Cai, Gregory C. DeAngelis, and Ralph D. Freeman. "Spatiotemporal receptive field organization in the lateral geniculate nucleus of cats and kittens." In: *Journal of neurophysiology* 78 2 (1997), pp. 1045–61.

[24] Andrew B. Watson. "DCT quantization matrices visually optimized for individual images". In: *Electronic Imaging*. 1993.

[25] P.J.B. Hancock, R.J. Baddeley, and L.S. Smith. "The principal components of natural images". In: *Network* 3 (1992), pp. 61–70.

[26] Jose Juan Esteve-Taboada et al. "Psychophysical Estimation of Early and Late Noise". In: *arXiv 10.48550/ARXIV.2012.06608* (2020). URL: https://arxiv.org/abs/2012.06608.

[27] David H Brainard and Spatial Vision. "The psychophysics toolbox". In: *Spatial vision* 10 (1997), pp. 433–436.

[28] Eric R. Kandel, James H. Schwartz, and Thomas M. Jessell, eds. *Principles of Neural Science*. Third. New York: Elsevier, 1991.

[29] Li Zhaoping. "A new framework for understanding vision from the perspective of the primary visual cortex". In: *Current Opinion in Neurobiology* 58 (2019), pp. 1–10.

[30] E. Martinez-Uriegas. "Chromatic-achromatic multiplexing in human color vision". In: ed. by D H Kelly. CRC Press, 1994. Chap. Chapt. 4 in Visual Science and Engineering: Models and Applications, pp. 117–187.

[31] J. Atick, Z. Li, and A. Redlich. "Understanding Retinal Color Coding from First Principles". In: *Neural Comput.* 4.4 (1992), pp. 559–572.

[32] Qiang Li et al. "Contrast sensitivity functions in autoencoders". In: *Journal of Vision* 22.6 (May 2022), pp. 8–8.

[33] M. Martinez et al. "Derivatives and inverse of cascaded linear+nonlinear neural models". In: *PLOS ONE* 13.10 (Oct. 2018), pp. 1–49.

[34] A. B. Watson and J. A. Solomon. "Model of visual contrast gain control and pattern masking". In: *JOSA A* 14.9 (1997), pp. 2379–2391.

[35] Rajesh Rao et al. "Natural Image Statistics and Divisive Normalization: Modeling Nonlinearities and Adaptation in Cortical Neurons". In: *Statistical Theories of the Brain* (Jan. 2001).

[36] O. Schwartz and E.P. Simoncelli. "Natural signal statistics and sensory gain control". In: *Nature Neurosci.* 4.8 (2001), pp. 819–825.

[37] J. Malo and E Simoncelli. "Nonlinear image representation for efficient perceptual coding". In: *IEEE Trans.Im.Proc.* 15.1 (2006), pp. 68–80.

[38] Ruben Coen-Cagli, Peter Dayan, and Odelia Schwartz. "Cortical Surround Interactions and Perceptual Salience via Natural Scene Statistics". In: *PLOS Comp. Biol.* 8.3 (Mar. 2012), pp. 1–18.

[39] A. B. Watson and J. Malo. "Video quality measures based on the standard spatial observer". In: *IEEE Proc. Int. Conf. Im. Proc.* Vol. 3. 2002, pp. III–III.

[40] J. Malo et al. "Characterization of the human visual system threshold performance by a weighting function in the Gabor domain". In: *Journal of Modern Optics* 44.1 (1997), pp. 127–148.

[41] Andrew Watson. "Image Compression Using the DCT". In: *Mathematica Journal* 4 (Aug. 1994).

[42] Albert Ahumada and Heidi Peterso. "Luminance-Model-Based DCT Quantization for Color Image Compression". In: *Proc SPIE Human Vision, Visual Process Display III* 1666 (Dec. 1997).

[43] Andrew B. Watson, ed. *Digital Images and Human Vision*. Cambridge, MA, USA: MIT Press, 1993. ISBN: 0262231719.

[44] J. Malo, AM Pons, and J.M. Artigas. "Subjective image fidelity metric based on bit allocation of the human visual system in the DCT domain". In: *Im. Vis. Comp.* 15.7 (1997), pp. 535–548.

[45] G. Camps-Valls et al. "Kernel-based Framework for Multi-Temporal and Multi-Source Remote Sensing Data Classification and Change Detection". In: *IEEE TGRS* 46.6 (2008), pp. 1822–1835.

[46] Hugh R Wilson and Jack D Cowan. "A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue". In: *Kybernetik* 13.2 (1973), pp. 55–80.

[47] Mark A. Georgeson and Timothy S. Meese. "Fixed or variable noise in contrast discrimination? The jury's still out..." In: *Vision Research* 46.25 (Nov. 2006), pp. 4294–4303.

[48] M.D. Fairchild. *Color Appearance Models*. The Wiley-IS&T Series in Imaging Science and Technology. Wiley, 2013.

[49] N. Ponomarenko et al. "Color image database for evaluation of image quality metrics". In: *2008 IEEE 10th Workshop on Multimedia Signal Processing*. 2008, pp. 403–408.

[50] V. Laparra, J. Muñoz, and J. Malo. "Divisive normalization image quality metric revisited". In: *JOSA A* 27.4 (2010), pp. 852–864.

[51] A. Hepburn et al. "Perceptnet: A Human Visual System Inspired Neural Network For Estimating Perceptual Distance". In: *IEEE ICIP*. 2020, pp. 121–125.

[52] P.C. Teo and D.J. Heeger. "Perceptual image distortion". In: *Proceedings of 1st International Conference on Image Processing*. Vol. 2. 1994, 982–986 vol.2.

[53] Z. Wang and A. C. Bovik. "Mean squared error: Love it or leave it? A new look at signal fidelity measures". In: *IEEE Signal Processing Magazine* 26.1 (2009), pp. 98–117.

[54] Vasily Strela et al. "Adaptive Wiener denoising using a Gaussian scale mixture model in the wavelet domain". In: *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)* 2 (2001), 37–40 vol.2.

[55] J. Gutiérrez, F. J Ferri, and J. Malo. "Regularization operators for natural images based on nonlinear perception models". In: *IEEE Trans. Im. Proc.* 15.1 (2006), pp. 189–200.

[56] Siwei Lyu and Eero P. Simoncelli. "Nonlinear Extraction of Independent Components of Natural Images Using Radial Gaussianization". In: *Neural Computation* 21.6 (2009), pp. 1485–1519.

[57] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. "Estimating mutual information". In: *Phys. Rev. E* 69 (6 June 2004), p. 066138.

[58] Henry Stark and John Woods. *Probability, Random Processes, and Estimation Theory for Engineers*. Vol. 90. Jan. 1994.

[59] Valero Laparra et al. "Information Theory Measures via Multidimensional Gaussianization". In: *CoRR* abs/2010.03807 (2020). arXiv: 2010.03807. URL: https://arxiv.org/abs/2010.03807.

[60] J.E. Johnson et al. "Information Theory in Density Destructors". In: *7th Int. Conf. Mach. Learn., ICML 2019, Workshop on Invertible Normalization Flows*. 2019.

[61] Jesús Malo. "Information flow in Color Appearance Neural Networks". In: *Proceedings of Entropy 2021: The Scientific Tool of the 21st Century* (2019).

[62] V. Laparra et al. "Nonlinearities and adaptation of color vision from sequential principal curves analysis". In: *Neural Computation* 24.10 (2012), pp. 2751–2788.

[63] M. U Gutmann et al. "Spatio-chromatic adaptation via higher-order canonical correlation analysis of natural images". In: *PloS one* 9.2 (2014), e86481.

[64] J. Vazquez-Corral et al. "Color Constancy Algorithms: Psychophysical Evaluation on a New Dataset". In: *Journal of Imaging Science and Technology* 53.3 (2009), pp. 31105-1-31105–9.

[65] I. Epifanio, J. Gutierrez, and J. Malo. "Linear transform for simultaneous diagonalization of covariance and perceptual metric matrix in image coding". In: *Patt. Recog.* 36.8 (2003), pp. 1799–1811.

[66] R.J. Clarke. "Relation between the Karhunen Loève and cosine transforms". In: *IEE Proceedings F (Comm. Radar Sig. Proc.)* 128 (6 1981), pp. 359–361.

[67] João Semedo et al. "Feedforward and feedback interactions between visual cortical areas use different population activity patterns". In: *Nat Commun* 1099 (2022), p. 13.

[68] Timo van Kerkoerle et al. "Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex". In: *PNAS* 111.40 (2014), pp. 14332–14341.

[69] P. Christiaan Klink et al. "Distinct Feedforward and Feedback Effects of Microstimulation in Visual Cortex Reveal Neural Mechanisms of Texture Segregation". In: *Neuron* 95.1 (2017), 209–220.e3.

[70] Jorge F. Mejias et al. "Feedforward and feedback frequency-dependent interactions in a large-scale laminar network of the primate cortex". In: *Science Advances* 2.11 (2016), e1601335.

[71] Hulusi Kafaligonul, Bruno Breitmeyer, and Haluk Öğmen. "Feedforward and feedback processes in vision". In: *Front. Psychol.* 6 (Mar. 2015).

[72] H. Barlow. "Possible Principles Underlying the Transformations of Sensory Messages". In: *Sensory Comm.* 1 (Jan. 1961).

[73] H. Barlow. "Redundancy reduction revisited". In: *Network: Comp. Neur. Syst.* 12.3 (2001), pp. 241–253.

[74] J. M. Hillis and D.H Brainard. "Do common mechanisms of adaptation mediate color discrimination and appearance?" In: *JOSA A* 22.10 (2005), pp. 2090–2106.

[75] E. Simoncelli and D. Heeger. "A model of neuronal responses in visual area MT". In: *Vis. Res.* 38.5 (1998), pp. 743–761.

[76] B. Chai et al. "Exploring Functional Connectivities of the Human Brain using Multivariate Information Analysis". In: *NIPS*. Ed. by Y. Bengio. Vol. 22. Curran Associates, Inc., 2009, pp. 270–278.

[77] Sameer Saproo and John T. Serences. "Attention Improves Transfer of Motion Information between V1 and MT". In: *J. Neurosci.* 34.10 (2014), pp. 3586–3596.

[78] T.v.d. Twer and D.I.A. MacLeod. "Optimal nonlinear codes for the perception of natural colours". In: *Network: Computation in Neural Systems* 12.3 (2001), pp. 395–407.

[79] V. Laparra and J. Malo. "Visual aftereffects and sensory nonlinearities from a single statistical framework". In: *Front. Human Neurosci.* 9 (2015), p. 557.

[80] A. Renart et al. "The asynchronous state in cortical circuits". In: *Science* 327(5965) (2010), pp. 587–590.

[81] D.H. Foster, I. Marin-Franch, and S.M.C. Nascimento. "Coding efficiency of CIE color spaces". In: *Proc. 16th Color Imag. Conf.* Soc. Imag. Sci. Tech. 2008, pp. 285–288.

[82] E.J. Allen, G. St-Yves, and Y. Wu. "A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence." In: *Nature Neurosci.* 25 (2022), pp. 116–126.

[83] N. Chang, J.A. Pyles, and A. Marcus. "BOLD5000, a public fMRI dataset while viewing 5000 visual images." In: *Sci. Data* 6 (2019), p. 49.

.

Paper IV

Contrast sensitivity functions in autoencoders

Qiang Li Image Processing Lab, Parc Cientific, Universitat de Valéncia, Spain

Alex Gomez-Villa Computer Vision Center, Universitat Autónoma de Barcelona, Spain

Marcelo Bertalmío Instituto de Óptica, Spanish National Research Council (CSIC), Spain

Jesús Malo Image Processing Lab, Parc Cientific, Universitat de Valéncia, Spain

Three decades ago, Atick et al. suggested that human frequency sensitivity may emerge from the enhancement required for a more efficient analysis of retinal images. Here we reassess the relevance of low-level vision tasks in the explanation of the contrast sensitivity functions (CSFs) in light of 1) the current trend of using artificial neural networks for studying vision, and 2) the current knowledge of retinal image representations. As a first contribution, we show that a very popular type of convolutional neural networks (CNNs), called autoencoders, may develop human-like CSFs in the spatiotemporal and chromatic dimensions when trained to perform some basic low-level vision tasks (like retinal noise and optical blur removal), but not others (like chromatic) adaptation or pure reconstruction after simple bottlenecks). As an illustrative example, the best CNN (in the considered set of simple architectures for enhancement of the retinal signal) reproduces the CSFs with a root mean square error of 11% of the maximum sensitivity. As a second contribution, we provide experimental evidence of the fact that, for some functional goals (at low abstraction level), deeper CNNs that are better in reaching the quantitative goal are actually worse in replicating human-like phenomena (such as the CSFs). This low-level result (for the explored networks) is not necessarily in contradiction with other works that report advantages of deeper nets in modeling higher level vision goals. However, in line with a growing body of literature, our results suggests another word of caution about CNNs in vision science because the use of simplified units or unrealistic architectures in goal optimization may be a limitation for the modeling and understanding of human vision.

Introduction

The human contrast sensitivity function (CSF) characterizes the psychophysical response to visual gratings of different frequency (Campbell & Robson, 1968). Filter characterizations in the Fourier domain are complete only for linear, shift-invariant systems. Human vision certainly is more complicated than that, however, this simple measure of the bandwidth of the system is still of paramount significance in biological vision: the CSF filter is an image-computable model that roughly describes the kind of visual information that is available for humans (Watson & Ahumada, 2016). Moreover, although it is defined for threshold conditions, there are many examples that illustrate the relevance of the CSF in more general situations (Watson et al., 1986; Watson & Malo, 2002; Watson & Ahumada, 2005), so it has shaped image engineering over decades (Mannos & Sakrison, 1974; Hunt, 1975; Wallace, 1992; Taubman & Marcellin, 2001). This theoretical and practical relevance motivated the measurement of CSFs, not only for spatial gratings (Campbell & Robson, 1968), but also for moving gratings (Kelly, 1979), chromatic gratings (Mullen, 1985), spatiotemporal chromatic gratings (Díez-Ajenjo et al., 2011), at different luminance levels (Wuerger et al., 2020), and for alternative basis of the image space (Malo et al., 1997).

Principled explanations of the human CSFs

Of course, the psychophysical CSFs have physiological roots in the spatiotemporal bandwidths of the center-surround cells tuned to achromatic and chromatic stimuli (Enroth-Cugell & Robson, 1966;

Citation: Li, Q., Gomez-Villa, A., Bertalmío, M., & Malo, J. (2022). Contrast sensitivity functions in autoencoders. *Journal of Vision*, *22*(6):8, 1–45, https://doi.org/10.1167/jov.22.6.8.

https://doi.org/10.1167/jov.22.6.8 Received February 28, 2021; published May 19, 2022 ISSN 1534-7362

de Valois & Pease, 1971; Ingling & Martinez-Uriegas, 1983; Martinez-Uriegas, 1994; Cai et al., 1997; Reid & Shapley, 1992, 2002). However, the physiological basis of psychophysical phenomena does not explain the functional role (or *goal*) of the underlying computation (Marr & Poggio, 1976; Marr, 1982). The discussion about the goal of certain mechanism relies on deriving the biological behavior from a computational principle. In the specific case of the CSFs, the classical work of Atick et al. (Atick et al., 1992; Atick & Redlich, 1992; Atick, 2011) derived the spatiochromatic CSFs from the maximization of the information transferred from the input to the response of the system that, under certain conditions, is equivalent to optimal deblurring and denoising of the retinal signals. These classical explanations were based on clever observations about the 2nd order properties of natural images, but relied on linear filtering models. As a result, the consideration of more flexible (nonlinear) models could lead to a better fulfillment of the computational goal and, eventually to better explanations of the CSFs. A step forward in a more general (nonlinear) derivation of these phenomena from low-level principles was given by Karklin and Simoncelli (2011), where they obtained sensors with center-surround receptive fields optimizing the information transferred by a linear+nonlinear layer of neurons with noisy inputs. However, this work did not consider the chromatic or the temporal dimensions of the problem, and no explicit comparison with the psychophysical CSFs was done. Similarly, (Lindsey et al., 2019) also reproduced center-surround sensors close to the retina when training anatomically constrained artificial neural nets (in this case, training for a higher level task, such as object recognition). Again, these center-surround cells eventually would induce CSFs, but this was not analyzed in that paper.

Emergence of CSFs in artificial neural networks

Automatic differentiation (Baydin et al., 2018) has simplified the search of computational principles in vision science because it allows the optimization of complex models according to different goals without the burden of obtaining the analytical derivatives of the goals w.r.t. the model parameters. Automatic differentiation is at the core of the current explosion of *deep learning* (Goodfellow et al., 2016). A full analytical description of the derivatives of realistic nonlinearities in visual neuroscience is certainly possible (Martinez et al., 2018), but the widespread availability of deep-learning tools for simplified neurons makes the exploration of these artificial architectures much easier. Conventional convolutional neural networks (CNNs) are too simplistic from the neuroscience perspective,[1] but the freedom to combine multiple of such simplified layers in any possible way may compensate this shortcoming. In the end, one has a flexible system that can be optimized with automatic differentiation to fulfill whatever computational goal under consideration. As a result, deep learning models are becoming standard in visual neuroscience (Kriegeskorte, 2015; Yamins & DiCarlo, 2016; Cadena et al., 2019).

According to the above, the study of the CSF of artificial neural networks is interesting for two reasons: 1) CNNs are flexible and easily optimizable tools that may allow us to investigate principled explanations of the human CSFs with more generality than the classical methods considered, and 2) given the widespread use of CNNs in computer vision and their recent use in visual neuroscience, the eventual emergence of human-like sensitivities in these artificial systems has intrinsic interest.

Very recently, two groups have reported complementary results on the emergence of CSFs in deep networks: first, in order to explain the human-like nature of some of the brightness and color illusions in CNNs trained for *low-level* visual tasks found in Gomez-Villa et al. (2019), a novel eigenanalysis of the networks was proposed (Gomez-Villa et al., 2020). This analysis revealed the emergence of human-like chromatic channels and achromatic and chromatic CSFs in these channels. Then Akbarinia et al. (2021) have found that networks trained for *high-level* visual tasks, such as classification, also may develop an achromatic CSF, in this case not explicitly imposing low-level constraints.

Contributions and scope of this work

- First, following Atick et al. (1992), Atick and Redlich (1992), Atick (2011), Karklin and Simoncelli (2011), we reconsider principled explanations of the CSFs from *low-level* visual tasks in light of new available methods: i) the current tools from deep-learning, and ii) the current knowledge of retinal image representations. We check the emergence of spatiotemporal chromatic CSFs in a wider range of low-level (goal/architecture) situations with more realistic inputs.

  Regarding the retinal input, we use recent models of the human modulation transfer function (MTF) (Watson, 2013), and recent calibrated estimations of the noise in the cones (Esteve et al., 2020) obtained via the retina models implemented in ISETbio (Cottaris et al., 2019, 2020). In this way, here we generate realistic spatiotemporal noisy inputs to the visual pathway in a plausible representation: the cones of (Stockman & Sharpe, 2000) tuned to long, medium, and short wavelengths (LMS cones).

  Regarding the deep learning tools, we use spatiotemporal extensions of the *convolutional autoencoders* used in our analysis of color illusions in CNNs (Gomez-Villa et al., 2019, 2020). We

elaborate on the proper determination of the CSF for convolutional autoencoders: instead of the linear characterization of the autoencoder used in Gomez-Villa et al. (2020), which hides its nonlinear nature into a single matrix, here we stimulate the networks with gratings of different contrast. In this way, the changes of the attenuation functions describe the nonlinearities of the system.

Regarding the architectures, in this work we focus on autoencoders that reconstruct the signal in the input domain as opposed to the consideration of more general architectures that encode the images into more abstract representations to achieve higher level tasks, such as classification. This limitation in scope is reasonable if one wants to model early vision stages like the lateral geniculate nucleus (LGN), which do not imply change of domain and may function according to error minimization and signal enhancement principles (Martinez-Otero et al., 2014). If the CSFs are related to the response of LGN neurons, as is usually assumed, autoencoders seem a reasonable computational framework to use.

Regarding the tasks, in this low-level context with autoencoder tools, we consider different visual tasks which may be implemented as early as in the retina-LGN path: a) the enhancement of the retinal signal (related to information maximization) when the input is subject to different degrees of degradation owing to different pupil diameters or different plausible levels of retinal noise, b) the compensation of changes in the spectral illumination of the scene in a reasonable range of color temperature, and c) the reconstruction of the signal when some information may be lost in eventual bottlenecks.

- Second, here we provide experimental evidence of "deeper CNNs are not necessarily better" (in representing this abstraction level). The bigger generality of flexible CNN models over fixed linear models is obvious, but one may ask: do more flexible architectures necessarily lead to more human CSFs? or does better accuracy in the goal imply more human CSFs and masking behavior? Consistently with previous results in low-level tasks (Flachot et al., 2020; Gomez-Villa et al., 2020), our CSF results presented also seem to favor shallow networks (in the explored range of architectures).

  Our findings at this low level of abstraction complement other results where deeper architectures actually imply closer resemblance to human behavior (Yamins et al., 2014; Cichy et al., 2016; Cadena et al., 2019; Lindsey et al., 2019). But this is not contradictory because they refer to different abstraction levels (high-level object recognition vs. our low-level color constancy and error minimization goals).

The structure of this article is as follows. We used estimating contrast sensitivity in autoencoders that extends the theory proposed in Gomez-Villa et al. (2020) to obtain the CSFs of autoencoders with an analysis of the energy (or standard deviation) of the input and the output gratings. Experiments describe the considered low-level visual tasks (compensation of biodistortions, chromatic adaptation, and signal reconstruction after bottlenecks) and the setting of the numerical experiments. Results show the main empirical findings of the work: the emergence of human-like CSFs in the spatiotemporal and chromatic dimensions in shallow CNN autoencoders trained to minimize the distortion introduced by the optics of the eye and the noise in the cones. Finally, we discuss the implications of the empirical results: on the one hand, statements about the goal or organization principles are difficult to separate from the implementation because the final behavior very much depends on the algorithmic level (or selected architecture). On the other hand, special care has to be taken in using deep models in low-level vision science: their ability for function approximation may make them excel in the performance of a sensible score, but without the appropriate architecture constraints, this does not guarantee the similarity with humans. Appendix A provides details of the implementation of the models. Appendix B describes the image/video datasets to train the models and the sinusoids used to probe the networks. Appendix C illustrates the proper training and convergence of all the considered CNNs in all experiments. It shows the learning curves and explicit examples of the responses (reconstructed signals in test) for all the considered goal/architecture scenarios.

Methods: Estimating contrast sensitivity in autoencoders

Here we consider different linear characterizations of the autoencoders including the eigenanalysis proposed in Gomez-Villa et al. (2020). That theory is extended with the explicit consideration of the image acquisition process in the human eye, which leads us to propose a procedure to estimate the autoencoder CSF that is more connected to the definition of the CSF in human observers.

Autoencoders

Autoencoders are artificial networks that transform the signal into an inner representation through an *encoder*, and a *decoder* transforms this inner representation back in the input domain.

$$\boldsymbol{x} \xrightarrow{N_\theta(\boldsymbol{x})} \boldsymbol{y} \quad (1)$$

In Equation 1, we do not made explicit the encoding and decoding operations, that is, $\boldsymbol{x}$ and $\boldsymbol{y}$ are in the image space. In this work, we will not make any assumption on the nature of the inner representation of the autoencoder. This is because the basic goal function in autoencoders (reconstruction error) is defined in the image domain, shared by input and response. Moreover, with the appropriate stimuli, the CSF characterization can be defined in this image domain.

Following Gomez-Villa et al. (2019, 2020), we focus on convolutional autoencoders. We discuss and explore different computational goals, but, for now, let's consider that the parameters θ are trained to compensate the blur and noise introduced in the signal by the image acquisition process. In this context, given a clean image, $\boldsymbol{x}_c$, the input to the neural system would be a distorted version: $\boldsymbol{x} = H \cdot \boldsymbol{x}_c + \boldsymbol{n}_r$, where H is a blurring operator related to the optics of the eye and $\boldsymbol{n}_r$ is the noise associated with the response of the LMS photodetectors at the retina. Both H and $\boldsymbol{n}_r$ are unknown to the neural system. The goal of the network at this early stage is inferring $\boldsymbol{x}_c$ from $\boldsymbol{x}$. Accurate models of LGN cells show that this may be one of the goals of the biological processing after retinal detection (Martinez-Otero et al., 2014).

In the supervised learning setting of artificial neural networks, the parameters θ of the network are selected so that the average reconstruction error $\varepsilon = |\boldsymbol{x}_c - N_\theta(\boldsymbol{x})|_2$ is minimized over a set of training images (Goodfellow et al., 2016). In our case, we refer to the average reconstruction error as ε_{LMS} because the input signal is expressed in the LMS cone space (Stockman & Sharpe, 2000). Of course, supervised learning and parameter updates using backpropagation may not be biologically plausible (Lillicrap et al., 2020). However, our initial aim here is looking for statistical explanations of human frequency sensitivity and hence artificial neural networks can be seen as convenient tools to optimize the selected goal. With this focus on the *goal*, the specific learning algorithm is not as important as ensuring that the final network actually fulfills the goal. We will see that the situation may not be that simple because networks optimizing the same goal with equivalent performance may display *human* or *non-human* CSFs, depending on their architecture.

Filter definition of the CSF and linearized autoencoders

The CSF describes the linear response of human viewers for low-contrast sinusoids (Campbell & Robson, 1968; Kelly, 1979; Mullen, 1985). In that linear setting, the CSF describes an input-output mapping where an input sinusoid of frequency f, the basis function $\boldsymbol{b}^f$, leads to an output, $\boldsymbol{y}^f$, with attenuated contrast (or attenuated standard deviation, σ). The output standard deviation is given by, $\sigma(\boldsymbol{y}^f) = \mathrm{CSF}(f)\,\sigma(\boldsymbol{b}^f)$. In the case of humans the attenuation factor, $\mathrm{CSF}(f)$, has to be obtained from contrast thresholds because there is no access to the output. However, for autoencoders, the computation of the output is straightforward. If the degradation of the acquisition is taken into account, the sinusoids, $\boldsymbol{b}^f$, used to simulate the measurement of the CSF have to undergo the degradation as well, and we should consider an *eye+network* system, S:

$$b^f \xrightarrow{H\cdot b^f + n_r} b^f_\star \xrightarrow{N_\theta(b^f_\star)} y^f \qquad S_\theta(b^f) \tag{2}$$

Therefore, one could check the attenuation factor by comparing the standard deviation of output and input:

$$\mathrm{CSF}(f) = \frac{\sigma(S_\theta(\boldsymbol{b}^f))}{\sigma(\boldsymbol{b}^f)} = \frac{\sigma(N_\theta(H\cdot\boldsymbol{b}^f + \boldsymbol{n}_r))}{\sigma(\boldsymbol{b}^f)} \tag{3}$$

Note that the CSF ratio in Equation 3 (which uses degraded sinusoids to probe the network) is different from checking the Fourier response of the network, where one would use clean sinusoids at the input:

$$\mathcal{N}^F(f) = \frac{\sigma(N_\theta(\boldsymbol{b}^f))}{\sigma(\boldsymbol{b}^f)} \tag{4}$$

The relation of Equation 3 with the regular determination of the CSF in humans is illustrated in Figure 1. Of course, the ratio in Equation 3 should be computed for low-contrast sinusoids to keep parallelism with human CSF and keep the (eventually) nonlinear autoencoder in the low-energy range. For chromatic sinusoids the deviations have to be computed separately over the achromatic, red–green, and blue–yellow color channels (Mullen, 1985). In this work we use a classical opponent color space (Hurvich & Jameson, 1957) to generate achromatic and purely chromatic gratings and to decompose the corresponding responses.

Of course, plain attenuation for sinusoids in Equation 3 may not provide a full description of the action of nonlinear systems. In principle, it is not obvious why we should perform the analysis in a specific basis. Therefore, one should check to what extent waves are indeed eigenfunctions of the system.

A way to test this point is linearizing the response of the autoencoders in the low-contrast regime and check that it is shift invariant. Using a Taylor expansion, the response for low-contrast images can be approximated by the Jacobian around the origin (the zero-contrast image, $\mathbf{0}$, which is just a flat gray patch):

$$\begin{aligned} \boldsymbol{y} &= S_\theta(\boldsymbol{x}) \\ \boldsymbol{y}_{\mathrm{low}} &= S_\theta(\mathbf{0} + \boldsymbol{x}_{\mathrm{low}}) \\ &\approx S_\theta(\mathbf{0}) + \nabla_x S_\theta(\mathbf{0}) \cdot \boldsymbol{x}_{\mathrm{low}} \\ \boldsymbol{y}_{\mathrm{low}} &\approx \nabla_x S_\theta(\mathbf{0}) \cdot \boldsymbol{x}_{\mathrm{low}} \end{aligned} \tag{5}$$

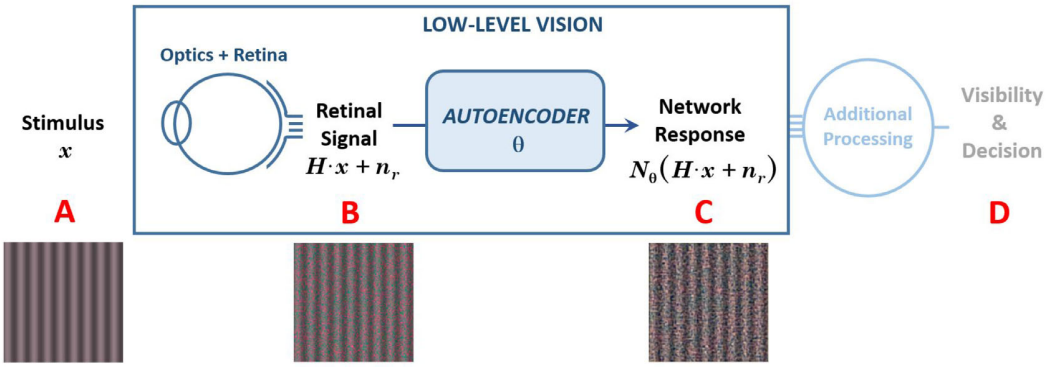


Figure 1. Definition of the CSF as a frequency-dependent attenuation factor in a system to develop low-level vision tasks. The diagram illustrates transforms in the visual signal from the input stimulus (A), the degraded signal owing to optical blur and retinal noise (B), the process of the early neural path where the output is still in a spatial LMS representation (C) modeled here by autoencoders, and additional mechanisms that compute a decision on the visibility (D). In the conventional view of the CSF as a filter, the process from A to C is assumed to be linear and (in humans) the visibility of gratings is assumed to be based on the amplitude of the response at the point C. In human psychophysics (with no access to C) the observer makes visibility decisions, and the attenuation factors are determined from the thresholds. When dealing with artificial systems we do have access to the response in C so we do not need to model the decision mechanism and we can simply estimate the CSF from the ratio in Equation 3.

where we assumed that the response for zero-contrast images is zero. If the behavior of the system at this low-energy regime is shift invariant, the Jacobian matrix can be diagonalized as $\nabla_x S_\theta(\mathbf{0}) = B \cdot \lambda \cdot B^{-1}$, with extended oscillatory basis functions in the columns of B (and rows of B^{-1}). Fourier basis and cosine basis are examples of extended (nonlocal) oscillatory functions that diagonalize shift invariant systems. The reason for this result is equivalent to the emergence of cosine basis when computing the principal components of stationary signals (shift-invariant autocorrelation) (Clarke, 1981). As a result, the slope of the response for low-contrast sinusoids (the CSF) will be related to the eigendecomposition of the Jacobian of the system at $\mathbf{0}$. Let's compute the response for a sinusoid in this Taylor/Fourier setting to see the relation. A basis function $\boldsymbol{b}^f$ with specific frequency f is orthogonal to all rows (sinusoids) in B^{-1} except that of the same frequency, that is, $B^{-1} \cdot \boldsymbol{b}^f = \boldsymbol{\delta}^{f'f}$. And this delta selects the corresponding column (of frequency f) among all the columns in the matrix B:

$$\begin{aligned} \boldsymbol{y}^f &= S_\theta(\boldsymbol{b}^f) \approx \nabla_x S_\theta(\mathbf{0}) \cdot \boldsymbol{b}^f \\ &\approx B \cdot \lambda \cdot B^{-1} \cdot \boldsymbol{b}^f \\ &\approx B \cdot \lambda \cdot \boldsymbol{\delta}^{f'f} \\ &\approx \lambda_f \boldsymbol{b}^f \end{aligned} \qquad (6)$$

So the slope of the response for basis functions of frequency f is λ_f (the corresponding eigenvalue of the Jacobian of the autoencoder). As a result, for systems with shift invariance in the low-contrast regime, the eigenvalues of the linear approximation of the system (eigenvalues of the Jacobian) are conceptually similar to the CSF. A direct comparison of the eigenvalue spectrum with the CSF may not be simple because the eigenfunctions may differ from Fourier sinusoids. Examples of this include isotropic systems (with a constant sensitivity for certain $|f|$ independent of orientation). In this case, the eigenbasis may be not sinusoids, but arbitrary linear combinations of sinusoids of the same frequency and different orientation.

Nevertheless, if the linearized version of the system (the Jacobian at $\mathbf{0}$) is shift invariant, which can be seen from a convolutional structure in the Jacobian matrix, oscillatory waves are eigenfunctions of the system, and hence Equation 3 may provide a good description of the behavior of the system.

Alternative linear characterizations of the autoencoders

A two-dimensional (2D) cartoon of the impact of the degradation and restoration processes in the probability density function (PDF) of the signal can illustrate alternative characterizations of the neural networks optimized to enhance the retinal signal (see Figure 2). In this diagram, two-pixel natural scenes (left) follow a PDF obtained from independent Student t-sources mixed by a matrix that introduces strong correlation between the luminance of the pixels. This kind of two-pixel representations is common to describe the statistics of natural images (Simoncelli & Olshausen, 2001), and mixtures of sparse components is a widely accepted model for natural scenes (Hyvärinen et al., 2009; van den Oord & Schrauwen, 2014; Malo, 2020),

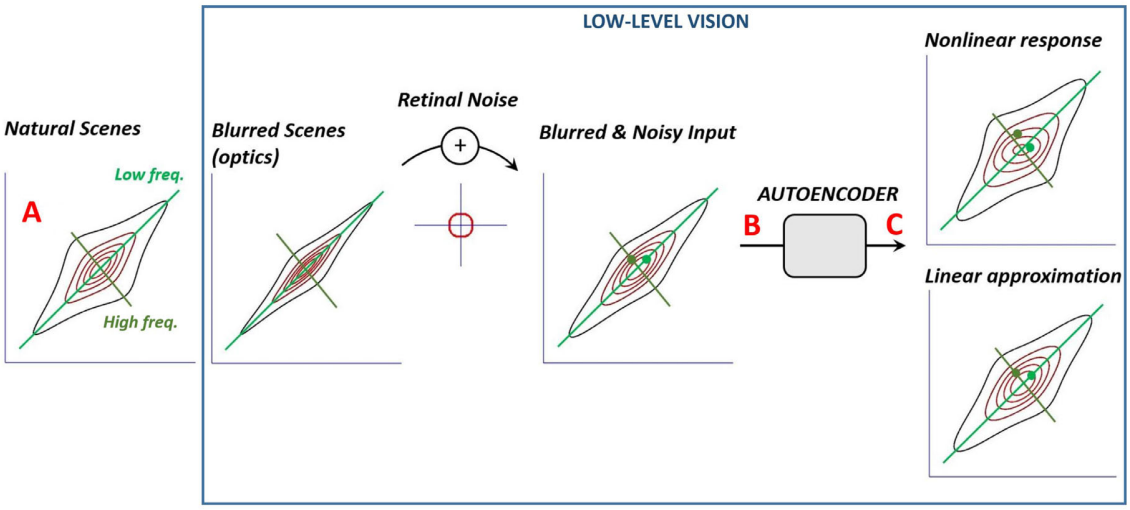


Figure 2. Degradation of the signal PDF and linear and nonlinear strategies to compensate the degradation. The axes of the plots represent the luminance in each of the photodetectors of two-pixel images as in Simoncelli and Olshausen (2001). (A) The PDF of natural scenes: marginal heavy tailed distributions of oscillatory functions mixed to have strong correlation in the pixel domain. Optical blur (second panel) implies contraction of the high frequencies. Additive retinal noise implies the convolution by the PDF of the noise leading to the PDF in (B). The solutions to the restoration problem at point C may include i) nonlinear transforms such as the one represented by the PDF at the right-top, and ii) linear transforms as the one at the right-bottom.

and appropriate enough for this illustration. In this diagram, the low-frequency direction corresponds with the main diagonal (where the two pixels have the same luminance) and the high-frequency direction is orthogonal (for images where one of the pixels is brighter than the other). The zero-contrast image is at the crossing point of the frequency axes.

Optical blur implies the attenuation of high-frequency versus low-frequency components and hence the contraction of the dataset as shown in the second panel. Assuming linear and noisy photoreceptors, the PDF of the retinal response results from the convolution of the PDF of the blurred images with the PDF of the noise (the function with circular support). The result (third panel) is the input to the autoencoder, whose goal is recovering the distribution at the first panel. Linear solutions are limited to global scaling of the domain (for instance, by inverting the contraction introduced by the blur), whereas nonlinear solutions may twist the domain in arbitrary ways.

In this setting, the computation of the CSF according to Equation 3 means putting low-contrast sinusoids (e.g., the samples highlighted in green in the third panel) through the system, and checking the amplitude of the output (green dots at the panels at the right) over the directions of the input. This nonlinear example illustrates the fact that the behavior can be contrast dependent (see the different twist in the concentric contours). This graphical view illustrates the difference between three possible linear characterizations, with $\boldsymbol{y} = M \cdot \boldsymbol{x}$:

- **The optimal linear solution:** The matrix M that better relates the input $\boldsymbol{x}$ with the desired output $\boldsymbol{x}_c$. This is the M that minimizes the expected value $E\{|\boldsymbol{x}_c - M \cdot \boldsymbol{x}|_2\}$. Assuming a representative set of N clean/distorted pairs stacked in the matrices $\boldsymbol{X}_c = [\boldsymbol{x}_c^{(1)}\boldsymbol{x}_c^{(2)} \dots \boldsymbol{x}_c^{(N)}]$ and $\boldsymbol{X} = [\boldsymbol{x}^{(1)}\boldsymbol{x}^{(2)} \dots \boldsymbol{x}^{(N)}]$, the optimal solution in Euclidean terms is given by the pseudoinverse:

$$M = \boldsymbol{X}_c \cdot \boldsymbol{X}^{\dagger} \quad (7)$$

- **Globally linearized network:** The matrix M that better describes the nonlinear behavior of the network over the whole set of natural images. This is the M that minimizes $E\{|S(\boldsymbol{x}, \theta) - M \cdot \boldsymbol{x}|_2\}$. Assuming a representative set of N input/output pairs stacked in the matrices $\boldsymbol{X} = [\boldsymbol{x}^{(1)}\boldsymbol{x}^{(2)} \dots \boldsymbol{x}^{(N)}]$, and $\boldsymbol{Y} = [\boldsymbol{y}^{(1)}\boldsymbol{y}^{(2)} \dots \boldsymbol{y}^{(N)}]$, the solution is given by the pseudoinverse:

$$M = \boldsymbol{Y} \cdot \boldsymbol{X}^{\dagger} \quad (8)$$

- **Locally linearized network at 0:** The matrix M that better describes the nonlinear behavior of the network for low-contrast images. This is the M that minimizes $E\{|\nabla_x S(\boldsymbol{0}, \theta) \cdot \boldsymbol{x}_{low} - M \cdot \boldsymbol{x}_{low}|_2\}$. Of course, this could be empirically approximated by $M = \boldsymbol{Y}_{low} \cdot \boldsymbol{X}_{low}^{\dagger}$, but in this case the obvious exact solution is:

$$M = \nabla_x S_\theta(\boldsymbol{0}) \quad (9)$$

Although the optimal linear solution (or the optimal linear network) is a convenient reference to describe the problem, the other two options are different characterizations of the autoencoder. Equation 8 summarizes the behavior of the network in a single matrix, and Equation 9 is a description only valid around **0**, and hence more closely connected to the low-contrast regime of the CSF. The eigenanalysis cited for $\nabla_x S$ in Equation 6 can be applied for the three matrix characterizations, but it is important to note the differences between them.

The Jacobian of cascades of linear+nonlinear layers (as in autoencoders based on CNNs) can be obtained either analytically,[2] or it can be obtained via automatic differentiation or alternative methods based on system identification (Berardino et al., 2017). However, these procedures are tedious, so in Gomez-Villa et al. (2020), we took the more straightforward approach represented by Equation 8.

The different linear characterizations considered in this section and the diagram in Figure 2 illustrate that the behavior of a nonlinear autoencoder for high contrasts may be substantially different from the threshold behavior. Therefore, the attenuation of sinusoids by the linearized system (by the matrix Equation 8) will be compared with the result of Equation 3.

Limitation of the proposed CSF definition in autoencoders

To maximize the equivalence to human CSFs, the proposed procedure (the ratio in Equation 3, which compares the signals at points C and A in Figure 1) implies the consideration of the retinal degradation process. This consideration of the retinal noise will be shown to improve similarity with human CSFs in the experiment 3, but it comes at a cost. Note that, even if the role of the autoencoder is compensating the retinal noise, complete removal is not possible. Therefore, there is some residual distortion in the response after the autoencoder. As a result, the standard deviation in the numerator of the proposed Equation 3 not only measures the contrast of the output grating, but also measures the energy of the residual noise. In this way, when the contrast of the sinusoids $\boldsymbol{b}^f$ is very small, as expected in threshold conditions, the standard deviation maybe measuring more the residual noise than the contrast of the output. The limitation of Equation 3 is that it has to be applied to sinusoids of relatively high contrasts so that the energy of the response coming from the sinusoid is bigger than the energy of the response coming from the noise.

One can overcome this limitation in two ways: 1) by computing the response many times for different noise evaluations and cancelling the residual noise by averaging over the realizations, and 2) by using relatively high-contrast sinusoids so that the effect of the residual noise is negligible.

In this work (for computational convenience) we used the second approach: we probed the models with sinusoids with contrasts in the range [0.07, 0.6]. The lower limit is certainly higher than the minimum absolute threshold of the standard spatial observer (which is approximately 0.005) (Watson & Malo, 2002; Watson & Ahumada, 2005). Nevertheless, we choose this range for two reasons: first, 0.07 is the average of the threshold achromatic contrasts in the standard spatial observer, and second, we empirically checked that the effect of the noise was negligible above this value.

Experiments

The Introduction raised questions on the role of low-level vision goals to explain the CSFs, the emergence of the CSFs in autoencoders working to solve these goals, and the eventual advantage of progressively more flexible models in explaining the CSFs. To address these issues in the more general spatiotemporal chromatic case, 1) we perform two extensive experiments (one with images, and one with video) to compensate biologically sensible degradation of the retinal signal (compensation of biodistortion), using a range of CNN architectures of different depth or flexibility, 2) we consider alternative low-level functional goals such as chromatic adaptation and the compensation of the effect of bottlenecks, 3) we consider different levels of biodistortion, chromatic shifts in different directions, and bottlenecks with different restrictions, and 4) we consider the consistency of the results under changes in the statistics of the signal. In this section, we describe the experimental setting of these simulations.

Functional goals

Compensation of retinal biodistortion (biological blur and noise) consists of overcoming the degradation introduced in the acquisition of the visual signal. Specifically, the top panel in Figure 3 shows how a natural scene is degraded at the output of the retina according to the variations of the eye MTF for different pupil diameters (from top to bottom, $d = 2$ mm, $d = 4$ mm and $d = 6$ mm), and a sensible range of Poisson retinal noise levels (from left to right, Fano factors $F = 0.25$, $F = 0.5$ and $F = 1$). Variations of the MTF have been simulated with the expression in Watson (2013), and the noise in LMS sensors has been estimated in the discrete representation of the

Figure 3. Functional goals. Possible low-level goals of the autoencoders are compensating the following distortions in the visual input. *Top panel:* different levels of retinal biodistortion. *Bottom panel (first row)* Changes in the spectral illumination. *Bottom panel (second row)* Changes in illumination + retinal biodegradation. An alternative low-level goal is the reconstruction of the signal in presence of bottlenecks (as in the architectures considered below, Figure 4 right). Note: This selfie of the corresponding author was sent to the first author to test the models on a copyright-free image.

input digital image as in Esteve et al. (2020). In that work, noise was obtained by stimulating the ISETBio retinal model (Cottaris et al., 2019, 2020) with flat stimuli of controlled size and tristimulus values over short and long exposure times. Cartesian resampling of the random cone mosaic of the retinal model and integration of the photocurrents over space/time reveals the effective Poisson nature of the noise (in the original LMS units) and allows the estimation of the effective Fano factor in the original discrete grid of the input image. In that way, we can easily generate calibrated noisy retinal images by adding this effective

Poisson noise in the LMS representation of the digital image. The illustrations in Figure 3 come from the transformation of the LMS tristimulus images into the RGB digital counts for proper display.

Chromatic adaptation

Consists of the compensation of the deviations of the signal induced by the change of illuminant. The bottom panel shows how the image of a natural scene changes under changes in the shape of the spectral radiance of the illuminant. The change of illuminant in a digital image was simulated in this way: each pixel of the image was associated with a reflectance chosen from a large database of natural reflectances so that under an equienergetic illuminant led to the tristimulus values of the pixel. Then, a black-body radiator, which simulates natural ambient light along the day, was used to generate spectra of the same energy but different shape. From there, we could get versions of the scene under arbitrary color temperatures. This process is straightforward using the functionalities and databases of Colorlab (Malo & Luque, 2002). Of course, this process is just an approximation because it disregards the (unknown) geometry of the scene and assumes a flat Lambertian world. Nevertheless, as illustrated in the examples in Figure 3, it does a good qualitative job to generate controlled samples to check chromatic adaptation in large image databases.

Compensation of chromatic shifts and biodistortions

The reason to consider this combination is that pure chromatic shifts with no additional distortion is not a realistic input for the visual pathway: the image acquisition front-end does exist and, hence, what we called biodistortion has to be taken into account. Such combination of distortions is illustrated by the second row of the bottom panel in Figure 3. Note that in the examples involving chromatic deviations (bottom panel) we introduced a flat-reflectance frame to help the models to cope with the chromatic adaptation[3].

Compensation of bottlenecks (pure reconstruction)

Consists of recovering the input after the signal has gone through a bottleneck. Examples of bottlenecks include the restriction of the spatial resolution or the restriction of the number of features (or channels) in the representation. Figure 4 (right panel) shows an illustrative range of architectures: from cases that expand the number of features (no bottleneck) to a variety of cases that introduce local pooling, reduce the number of features, or try to compensate the effect of spatial undersampling by increasing the number of features. Bottlenecks may imply severe information loss if the representation is not optimized. Therefore, pure reconstruction of the signal in presence of bottlenecks is a sensible low-level goal to explore.

Compensation of bottlenecks and biodistortion

As stated, errors in the acquisition front end (biodistortion) do exist, so their consideration together with bottleneck compensation makes the goal more realistic.

All in all, we explored nine levels of biodegradation, chromatic adaptation in the blueish and the reddish directions ($T = 8600\ K$ and $T = 4400\ K$, respectively), and the combination of the central biodistortion with the considered chromatic deviations. We considered a pure reconstruction task with the eight bottleneck configurations in Figure 4, and the compensation of these bottlenecks was also combined with the central biodistortion case. The optical/retinal degradation in movies was applied in a frame-by-frame basis. No experiments involving chromatic adaptation or bottlenecks were done in movies, but only in natural and cartoon images.

The above computational goals are all measured in *distortion* terms, or how well the deviations ε_{LMS} were compensated. However, even within this low abstraction level, other computational goals could be considered together with the distortion, as for instance the information or the energy of the signal. In the experiments we restrict ourselves to the considered cases of distortion minimization and purely architectural bottlenecks. The discussion suggests how the goals considered here could be related or combined with other kind of goals or more general (energy or information) bottlenecks.

Architectures

In this work, we consider 2D CNNs that act on spatiochromatic signals and three-dimensional (3D) CNNs that act on spatiotemporal signals (color images and color videos).

The set of explored architectures is shown in Figure 4. These are variations of the basic toy networks studied in Gomez-Villa et al. (2019, 2020): autoencoders with convolutional layers made of 8 feature maps with kernels of spatial size 5×5 and sigmoids or rectified linear units (ReLU) as activation functions. From that starting point, here we consider a range of nets of increasing *depth* and flexibility: from the linear network in Equation 7 (as a convenient baseline reference of one layer with no flexibility), and CNNs with two layers to eight layers, both for the 2D and 3D cases. Moreover, we also consider a range of architectures with different bottlenecks, in this case, only 2D.

Of course, the range of possible architectures is virtually infinite and an exhaustive exploration of the architecture space is out of the scope of this work.

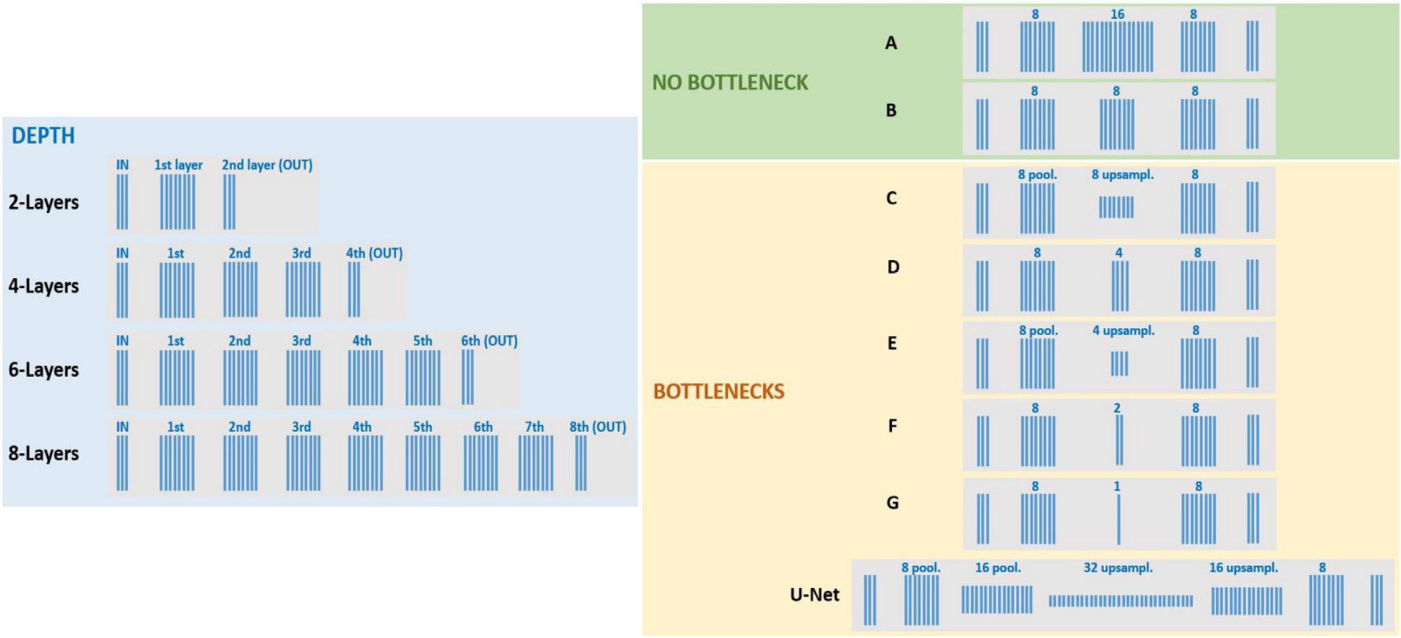


Figure 4. Architectures. (Left) Range of architectures of different *depth* following the basic structure of the nets studied in Gomez-Villa et al. (2019, 2020): three channels at the input and at the last (output) layer. In this work the input is in an LMS color representation. The rest of the layers have eight features with no undersampling or bottleneck, as represented by the eight blue lines of the same length. These architectures of increasing depth are used to study the compensation of the biodistortion in images and videos. The best of these architectures in terms of CSFs (which turns out to be the two layers example) is used with images to explore chromatic adaptation. (Right) Range of architectures used to illustrate the effect of *bottlenecks*. In this case, the inner layer in the four layer architecture at the left is systematically expanded (in A) or contracted (from C to G, either in the number of features or in the spatial resolution. See numbers and indications of pooling/upsampling and corresponding length of the layers-) to generate a range of bottlenecks. Finally, an illustrative U-Net with no residual connections is also considered in the experiments with bottlenecks.

However, note that the considered set of architectures of progressive flexibility and constraints is appropriate for the aim of this work for two reasons: 1) these architectures do a good job in fulfilling the goal so they are good examples to reason about systems that work according to the considered function, and 2) they display a range of flexibility and accuracy in the goal which is appropriate to illustrate the proposed questions (eventual emergence of the CSF and other nonlinearities, and qualitative effect in the CSF of increased flexibility and improvements in the goal accuracy).

The first point (the considered *toy* models do a reasonably good job in fulfilling the goal) is a technical issue that is demonstrated by the performance tables shown below and by the specific learning curves and reconstructions included in Appendix A. However, to put this quantitative performance in context, it is interesting to note that the retinal biodistortion is not an easy task to solve for general-purpose state-of-the-art image restoration CNNs. In particular, following Gomez-Villa et al. (2020), on top of the described toy networks, the computation of the CSFs of cutting-edge deeper models designed for restoration could be an illustrative limit to consider. However, we found that the combination of representative examples of generic CNNs for denoising (Zhang et al., 2017; Soh & Cho, 2021) and deblurring (Tao et al., 2018), which gave excellent results with arbitrary Gaussian noise and blur in Gomez-Villa et al. (2020), is not satisfactory with biological distortion. In particular, generic enhancement algorithms did not produce better results than the considered simple architectures (specifically trained for this biodistortion).

Of course, this does not mean that the toy models used here are better than the state of the art, nor that state-of-the-art models are intrinsically unable to deal with this biological degradation. One could certainly fine-tune these deep architectures for the biodistortion and then get a better result than with the considered set of architectures, but that is not the goal of this work. The relevant argument in favor of the considered (toy) architectures for our purposes here is this: the fact that generic blind restoration CNNs need to be retrained to get better results than the proposed models means that these simple models can be considered as good (enough) examples of systems actually fulfilling the goal.

Regarding the second point (the considered set of architectures is good enough to illustrate interesting questions), consider that i) according to the results presented below (Results, Tables 1 and 4) the toy

nonlinear models decrease by up to 35% and 48% the error of the optimal linear solution in images and video, respectively, and ii) the best nonlinear model reduces the error of the shallower nonlinear model by 21% and 12% in images and video respectively.

In summary, the considered set of architectures (progressively deeper CNNs and a range of bottlenecks) does a reasonable job in optimizing the goals, and it is wide enough to illustrate changes in the achievement of the goals. As a result, the considered set of architectures is appropriate to address the questions raised in the introduction.

See Appendix A for implementation details. Data and code are available at http://isp.uv.es/code/visioncolor/autoencoderCSF.html

See Appendix B for details on the databases to generate the training stimuli and the stimuli used to probe networks.

Assessing the quality of the CSF results

The CSFs defined for the autoencoders may be subject to two arbitrary scale factors. On the one hand, the response of the network could be multiplied by an arbitrary global scale factor and hence, the numerator in Equation 3 (and the CSF amplitude) would be multiplied by this scale factor as well. We refer to this global scale factor in the amplitude as α_{CSF}. On the other hand, the sampling assumptions (or assumptions on the extent of the signal, or the viewing distance) introduced in the description of the stimuli are arbitrary and they imply an arbitrary scaling in the frequency axis of our Fourier domains. We refer to this scale factor on frequency as α_f.

The factor on amplitude is not a major problem: one network and a modified version with its outputs multiplied by α_{CSF} are equivalent and their quality should be rated the same. The factor on frequency does not reduce the validity of the results either as long as it is moderate. Note that using the MTF expressions in Watson (2013), if the filter corresponding with a pupil of 3.5 mm is modified by applying $\alpha_f = 0.75$ or $\alpha_f = 4.5$, the resulting MTF is similar to what would have been obtained with d = 2 mm or d = 6 mm, respectively. Therefore, as changes in the MTF (the only element where the scaling in frequency matters) are plausible if $\alpha_f \in [0.75, 4.5]$, one should also discount moderate variations of this factor when assessing the quality of the CSFs.

The similarity between the model and the human CSFs will be measured by the Euclidean distance between the CSF vectors, averaging over the frequency, f, and the chromatic channels, c (achromatic, red–green and yellow–blue), which will be referred to as:

$$\text{RMSE} = \left(\sum_{f,\,c} (\text{CSF}_c^{\text{scaled}}(f) - \text{CSF}_c^{\text{human}}(f))^2 \right)^{\frac{1}{2}}, \quad (10)$$

where the *scaled* attenuation factors of the model are related to the *raw* attenuation factors of the model as:

$$\text{CSF}^{\text{scaled}}(f) = \alpha_{\text{CSF}} \cdot \text{CSF}^{\text{raw}}(\alpha_f \cdot f) \quad (11)$$

In the following, we report the scaled CSFs together with the scaling factors that minimize the distance with human CSFs.

It is important to mention that the relative scaling between the CSFs in the three chromatic channels is a characteristic feature of a network (or system) and it should not be modified. Therefore, the same factors in Equation 11 are applied to the three CSFs. With these considerations, the CSFs reported below represent the closest approximation the models may give to the human CSFs, and hence the comparison between them is fair.

The magnitude of the RMSE errors has to be understood in reference to the maximum value of the human sensitivity. As a convenient example to have in mind, RMSE – 22 corresponds with an average deviation of 10% of the scale of the human spatiotemporal CSF at every frequency and chromatic channel. This is because the maximum sensitivity is approximately 200 for stationary gratings and about 220 for moving gratings (Watson & Malo, 2002; Kelly, 1979).

List of experiments

The empirical exploration of the considered architectures consists of six experiments. Experiments 1 through 5 deal with spatiochromatic stimuli and 2D networks, and experiment 6 deals with spatiotemporal chromatic stimuli and 3D networks. As stated, the computational goals are measured by the Euclidean distance between the reconstructed image and the original image, referred to as ε_{LMS}. The similarity with the human behavior is measured in terms of the Euclidean distance between the model CSFs and the human CSFs, that is, the RMSE defined in Equation 10.

- **Experiment 1: Spatiochromatic CSFs from biodistortion compensation by a range of architectures.** This experiment is focused on the central degradation shown in the first panel of Figure 3 (d = 4 mm, F = 0.5) and analyzes in detail the CSFs for nine architectures: the optimal linear network, and eight CNN architectures with two, four, six, and eight layers with either sigmoid or ReLU activations, all optimized according to this distortion–compensation goal. Once the architectures are properly trained (using 20 $\cdot 10^3$ images of the ImageNet database cited in Appendix B, 18 $\cdot 10^3$ for training and 2 $\cdot 10^3$ for validation), we get the numerical performance

of the models in the independent test set of 10^3 images. The sizes of the train/validation/test sets are the same in all experiments with images, experiments 1 to 5. Throughout all the experiments, the performance is expressed as the average ε_{LMS} of the reconstruction in LMS space over 20 batches of 50 randomly chosen images per batch. The standard deviation over these 20 computations is also reported. The learning curves (train/validation) and the reconstructions of one representative test image are given in Appendix C. Then, the CSFs (attenuation factors) of the trained models are computed according to the method described in the Methods: Estimating contrast sensitivity in autoencoders for gratings of different contrasts. The eventual variation of the attenuation reveals the nonlinear nature of the contrast response for gratings. In experiment 1, we also show the CSFs of the linear network and the linearized versions of the nonlinear networks introduced in the Methods: Estimating contrast sensitivity in autoencoders. From the results of experiment 1, one of the nonlinear models is chosen as having representative resemblance with human behavior in terms of the CSFs (two layers with ReLU activation). Experiments 2, 3, and 4 further explore the behavior of this specific model in a number of conditions.

- **Experiment 2: Consistency of the CSFs from biodistortion compensation over a range of distortion levels.** This experiment is focused on the representative architecture selected after experiment 1, and checks its CSFs when trained for the nine different degradation levels considered in the first panel of Figure 3.
- **Experiment 3: CSFs from chromatic adaptation and biodistortion compensation.** This experiment checks the CSFs of the representative architecture selected after experiment 1, when it is trained for i) the biodegradation compensation alone ii) the degradation compensation together with compensation of a bluish illuminant, iii) the degradation compensation together with compensation of a reddish illuminant, iv) pure compensation of a bluish illuminant, and v) pure compensation of a reddish illuminant. In the illustration of Figure 3, these correspond with the five distorted versions closer to the clean image under equienergetic illuminant. As stated, the purely chromatic deviations are not realistic because they disregard the optics and retinal noise. However, they represent an illustrative reference. In the same vein, as a convenient reference, in this experiment we compute the CSF in two ways: a) the proposed (realistic) way, Equation 3, by putting the clean gratings through the retinal degradation before entering the network, and b) the idealized way, Equation 4, in which we simply put the clean gratings through the considered network. This will stress the difference in the obtained CSFs when considering realistic spatial degradations or not.
- **Experiment 4: Consistency of the human/non-human CSFs under change in signal statistics.** Here we reconsider the chromatic adaptation and the degradation–compensation goals of experiment 3 now using stimuli of (apparently) quite different spatiochromatic statistics: the images from the Pink Panther cartoons. All the other settings remain the same as in experiment 3.
- **Experiment 5: CSFs from bottleneck–compensation and biodistortion compensation.** This experiment shows the CSFs of the systems that emerge from imposing pure reconstruction of the signal in presence of bottlenecks in the network (the eight examples in Figure 4, right). Pure reconstruction is compared with the compensation of biodistortion in the same architectures. Given the similarity between activation options found in experiment 1, here we just explore the ReLU case.
- **Experiment 6: Spatiotemporal chromatic CSFs from biodistortion compensation by a range of architectures.** Here we check the fundamental findings of experiment 1 for spatiotemporal chromatic gratings on 3D networks optimized for degradation–compensation. Given the similarity between activation options found in experiment 1, here we just explore the sigmoid case. Therefore, we explored five architectures: the linear one and two, four, six, and eight layers with sigmoid. In this spatiotemporal case we used $22 \cdot 10^3$ video patches in the learning ($20 \cdot 10^3$ for training and $2 \cdot 10^3$ for validation), and $3 \cdot 10^3$ for test.

Results

Results in all the experiments have two parts: 1) the perception part, with the CSFs and the contrast responses of the networks, and 2) the technical part, with evidences of the convergence of the models, numerical performance in reconstruction, and visual examples of the performance in reconstructing images. The main text is focused on the perception part, while all the technical material is given in the Appendix C.

Experiment 1: Spatiochromatic CSFs from biodistortion compensation

Figure 5 shows the achromatic and chromatic CSFs of the considered models (the linear solution and the eight CNNs) together with the human CSFs for convenient reference. The human data come from

	Comput. goal ε_{LMS}		CSFs RMSE		Comput. goal ε_{LMS}		CSFs RMSE	
Distortion	15.5 ± 0.2							
Linear net	**13.1** ± 0.1		**24.4**					
CNNs	Sigmoid nonlinear	ReLU nonlinear	Sigmoid nonlinear	ReLU nonlinear	Sigmoid linearized	ReLU linearized	Sigmoid linearized	ReLU linearized
2 Layers	**10.3** ± 0.1	**10.8** ± 0.1	**26.9**	**24.5** ± 0.3	12.5 ± 0.1	12.6 ± 0.2	24.1	22.8
4 Layers	**8.9** ± 0.1	**9.1** ± 0.1	**26.6**	**28.5**	12.5 ± 0.1	12.5 ± 0.1	23.2	23.1
6 Layers	**8.7** ± 0.1	**8.5** ± 0.1	**29.7** ± 0.7	**33.1**	12.5 ± 0.2	12.5 ± 0.2	23.2	23.1
8 Layers	**8.9** ± 0.2	**8.7** ± 0.1	**31.2**	**31.6**	12.6 ± 0.1	12.7 ± 0.1	27.0	27.4

Table 1. Experiment 1: Emergence of CSFs from biodistortion compensation (computational goal and error in CSFs). The achievement of the computational goal is described by ε_{LMS} (error of the reconstructed signal in LMS) for batches of images of the independent test set (averages and standard deviations from 20 realizations with 50 images/batch). The distance between the CSFs of the networks and the human CSFs is measured by the RMSE between the functions, Equation 10. Uncertainty of the RMSE was estimated only in two cases (two layer ReLU and six layer sigmoid) and is represented here by the standard error of the corresponding means. It is interesting to note that the optimal linear solution (computed from the train set) has worse performance in the test set than the linearized versions of the networks. Numbers in bold style refer to nonlinear networks and numbers in regular style refer to linearized networks.

the achromatic standard spatial observer (Watson & Malo, 2002; Watson & Ahumada, 2005) and from the measurements in Mullen (1985). The plots for the nonlinear models show the attenuation factors (CSFs) for gratings of different contrast (dark to light colors mean lower to higher contrasts).

These plots include the RMSE measure of the difference of the artificial CSFs with the human CSFs. The insets also show the optimal values of the arbitrary scaling factors ($\alpha_f, \alpha_{\text{CSF}}$) applied to the axes of the raw CSFs of the network to minimize the distance with the human CSFs. Because these optimal scaling factors values were found exhaustively in all cases, the comparison of the final CSFs and RMSE values is fair.

Results show the emergence of a band-pass sensitivity in the achromatic channel and low-pass sensitivities in the chromatic channels. The bandwidth of the chromatic channels is always substantially narrower than the achromatic bandwidth. These properties are qualitatively in line with human behavior.

Shallower networks (either ReLU or sigmoid) display a greater resemblance with human CSFs. In particular, deeper nets introduce substantial distortion in the chromatic channels: note that the red–green channel is over attenuated (particularly for the eight layer architectures but also in the six layer cases). The RMSE scores summarize these differences and show that shallower nets (two and four layers) provide better explanations of the CSFs than deeper nets (six and eight layers).

Interestingly, the optimal linear solution (a single dense layer with identity activation) is the one that better reproduces the CSFs. However, by its linear nature, it cannot include contrast-dependent behavior. In this regard, the shallow networks (two layers) display a consistent decay of the gain (attenuation factor) with contrast. This decay has an impact on the contrast response curves for gratings. The contrast response curves describe the evolution of the amplitude of the response to a grating as a function of the contrast of the grating. In humans, contrast response curves are increasing saturating functions both for achromatic gratings (Legge & Foley, 1980; Legge, 1981) and for chromatic gratings (Martinez-Uriegas, 1997). The decay found in two layer CNNs implies a saturation of the contrast response curves for these shallow CNNs, in line with human behavior. Figure 6 shows representative examples of these response curves: Although the two layer network (top row) consistently displays saturating behavior for every frequency, the deeper net (bottom row) shows non-human (linear or expansion) responses.

Finally, Figure 7 shows the CSFs corresponding with the linearized versions of the nonlinear CNNs, Equation 8. Of course, the linear approximations have contrast-independent behavior and hence the same CSF for all contrasts. The global linear approximations of the nonlinear models improve the resemblance of the CSFs with human behavior: the linearized shallow nets are closer to humans than the linear model, and linearization corrects over attenuation of the chromatic channels in the six layer models. However this increased similarity with human CSF comes at the cost of a significant drop in the performance (see the increase in ε_{LMS} error in Table 1). The linearization leads to rigid models that disregard the differences between the original nonlinear models and behave more similarly. In any case, linearization does not overcome the

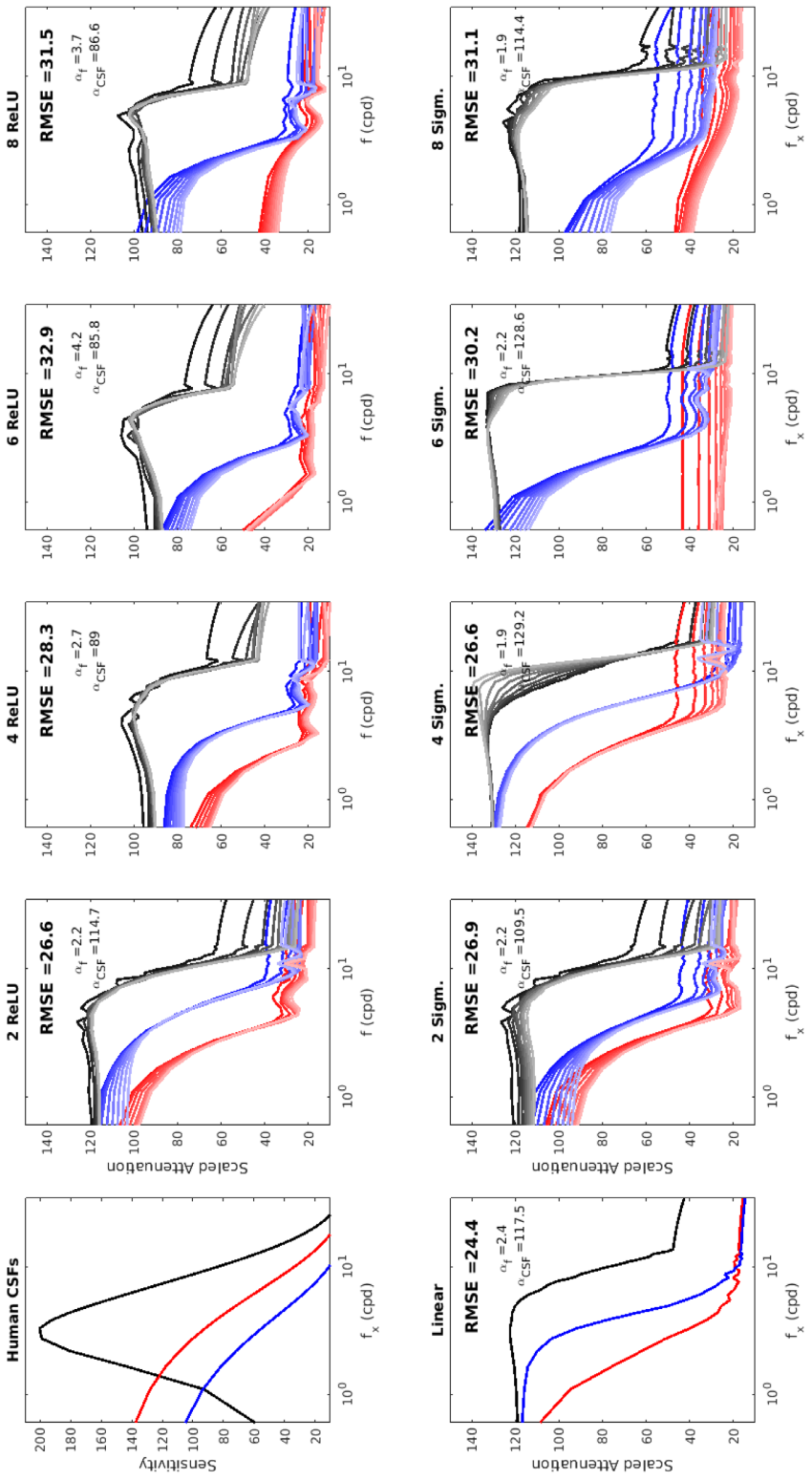


Figure 5. Experiment 1 (spatiochromatic CSFs from the compensation of biodistortion). Attenuation factors for gratings of different contrast for a range of CNN autoencoders trained for biodistortion compensation. Achromatic, blue, and red lines refer to the CSFs of the achromatic, red–green, and blue–yellow channels, respectively. Dark to light colors refer to progressively higher contrasts (evenly spaced in the range [0.07–0.6]). The human CSFs (top left) and the CSFs of the optimal linear solution (bottom left) are also shown as a convenient reference. RMSE measures describe the differences between the model and the human CSFs. The plots also display the values for the scaling factors in frequency and amplitude described in the text, Equation 11.

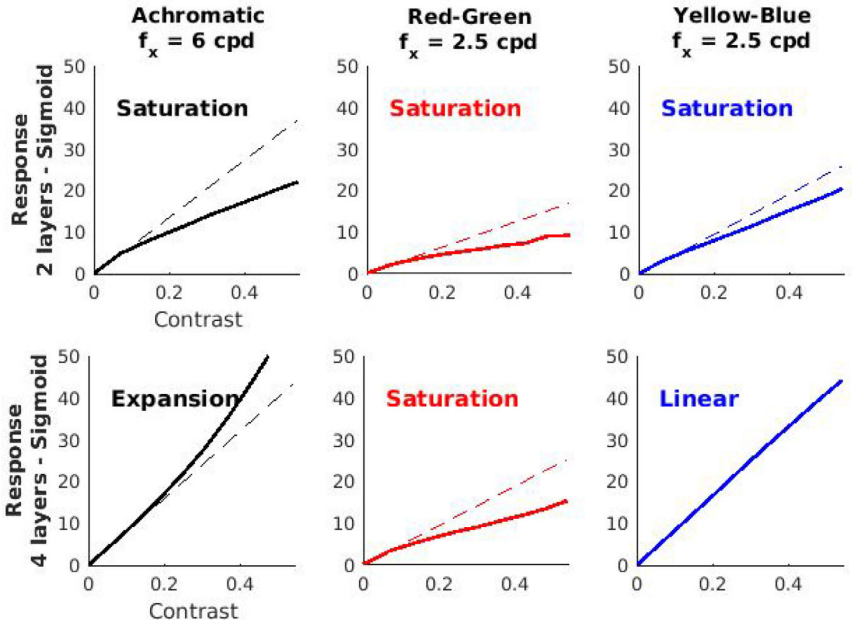


Figure 6. Experiment 1 (Illustrative contrast responses from the compensation of biodistortion). Representative examples of nonlinear responses for achromatic and chromatic sinusoids found. Saturation in these responses comes from the decay in the attenuation factors with contrast in Figure 5. Similarly, expansion comes from the increase in the attenuation factors. The linear behavior at the low-contrast regime has been plotted with dashed line as useful reference to highlight the nonlinear behavior. It is interesting to note that the saturating or expansive nature of the final contrast nonlinearity is a collective behavior that is not trivially attached to the specific (saturating or expansive) nonlinearities in individual neurons.

overattenuation of the red–green channel in the eight layer models.

Table 1 shows that while deeper networks are significantly better at fulfilling the computational goal (as expected from their increased flexibility), they are worse than shallow nets in reproducing the human behavior (as seen in Figures 5–7).

Progressive improvement in the goal for increasing depth is numerically substantial (and also visible in the reconstructed signals in Figure 14 in Appendix A) from two, four, to six layers, and the numerical performance stays (statistically) the same for eight layers. For this last case, there are chromatic issues in line with what was been found in the CSFs: the colorfulness of the reconstruction in Figure 14 is related to the relative gain of the chromatic channels. In particular, the consistent underestimation of the red–green CSF by the eight layer CNNs (either using ReLU or sigmoid activation) leads to low-saturation images. Interestingly, this effect is also visible in the reconstructed images coming from the linearized CNNs (Figure 15) and is consistent with the strong attenuation of the RG channel in the linearized eight layer architectures in Figure 7.

It is important to stress that the deviations in the chromatic CSFs in deep models do not come from not fulfilling the goal or having poor convergence in the training. First, all models (even the linear one) do reduce the error of the original retinal degradation (Table 1) so they are fulfilling the goal. And second, the learning curves in the Appendix C (Figure 14) show that all models achieved a plateau in the training thus indicating proper convergence. Moreover, the asymptotic values achieved in the learning are consistent with the ε_{LMS} in test shown in Table 1.

As stated, RMSE errors in Table 1 have to be interpreted in terms of the scale of the human CSF. For example, the best and the worst CNNs (RMSE = 24.4 and RMSE = 33.1, respectively) have average deviations of 11% and 15% with regard to the maximum human sensitivity. Of course, a single figure of merit averaged over frequencies and chromatic channels may hide an uneven distribution of the errors. For instance, consider the specific six layer sigmoid CSF shown in Figure 5, which displays a clear over attenuation of the red–green channel. In that case, if the global RMSE = 30.2 is broken down into its chromatic components we have 29.0, 35.0, and 26.8 for the achromatic, red–green, and yellow–blue errors, which clearly point out that the biggest problem is in the red–green sensitivity. The same is true for the average over spatial frequencies: the global description does not stress the discrepancy in the low frequencies of the achromatic channel. That is why the (necessarily limited) description in the tables comes together with the explicit plots of the three CSFs for different contrasts.

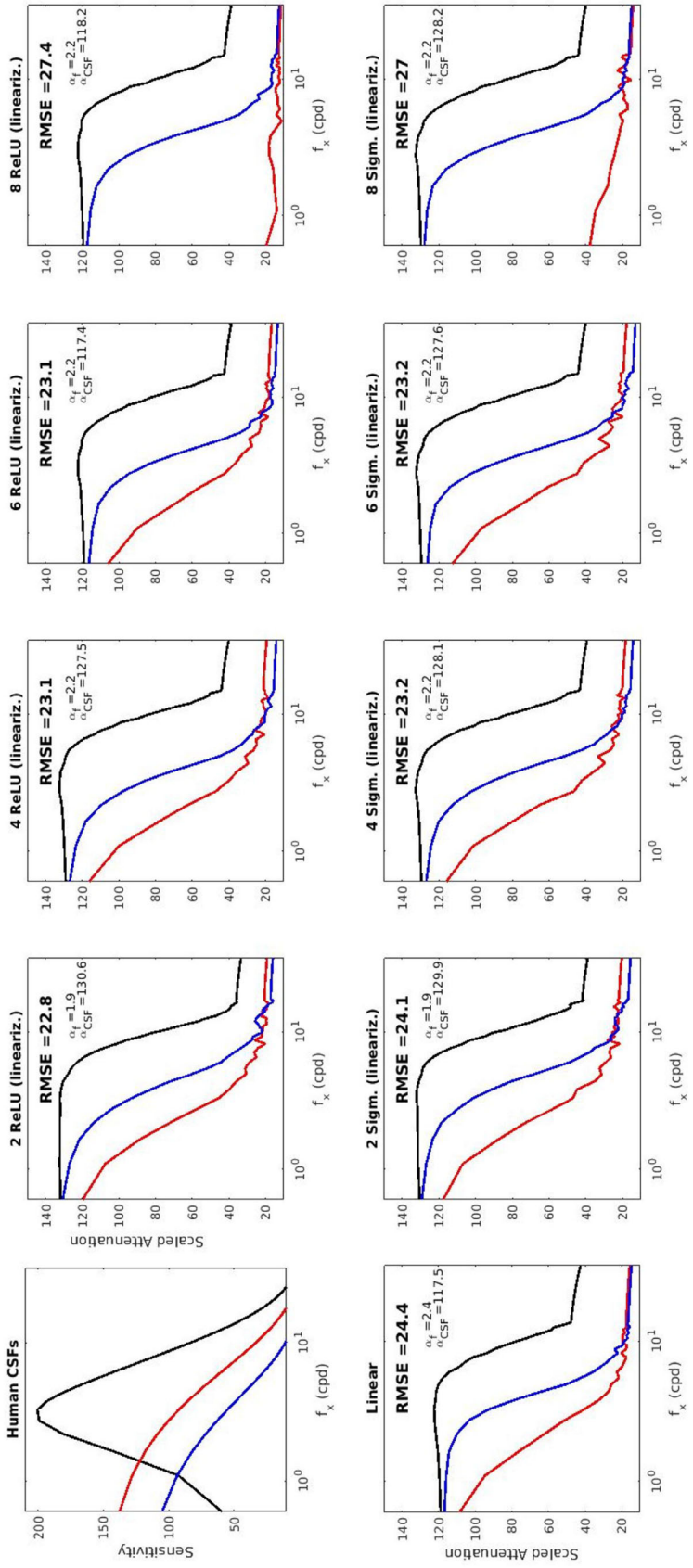


Figure 7. Experiment 1 (spatiochromatic CSFs for linearized networks in biodistortion compensation). The human CSFs and the CSF of the optimal linear network are also included as reference.

Another important technical issue is the consistency of the CSFs over random initialization. This is easy to check by training a number of times the same architecture for the same computational task and over the same set of stimuli but from different initial values of the model parameters. Given the intensive computation required,[4] we checked this variability only in two illustrative models: one with reasonably human-like behavior (two layer ReLU), and another with less-human CSFs (six layer sigmoid). In these two models, we retrained them 20 additional times and recomputed the corresponding CSFs (results not plotted). In the six layer case, all the explored seeds lead to a flat red–green CSF of too low sensitivity (i.e., a non-human behavior), and in some cases even the blue–yellow sensitivity was strongly attenuated too. On the contrary, the two layer case systematically leads to better CSFs, as summarized by the RMSE in Table 1, where the uncertainty is represented by the standard error of the mean. The shape of the sensitivities is pretty consistent in both cases, always better for the two layer case. At the same time, and not surprisingly, the six layer architectures systematically led to lower ε_{LMS} error. Only 1 of the 42 realizations (21 per model) led to a clear outlier (RMSE = 43.8 in the six layer case) and even for this CSF-outlier the distortion ε_{LMS} was not off the distribution. According to the observed consistency, the remaining 49 configurations of task/architecture in the work were studied using a single random initialization of the parameters.

The next experiments explore the consistency of the human-like behavior found in shallow autoencoders in a number scenarios. According to the results found in experiment 1, we select the two layer ReLU autoencoder as a representative example of shallow architecture with reasonable human-like behavior (RMSE of 11% of the maximum sensitivity) so we focus on this architecture in experiments 2 through 4.

Experiment 2: Consistency of CSFs over a range of biodistortions

Figure 8 shows the CSFs obtained when training the two layer ReLU net to compensate a range of retinal degradations (as described by the different pupil diameters and Fano factors). Learning curves that show the good convergence of the models and representative visual examples of the reconstructions are given in the Appendix C (Figures 16 and 17).

The results in Figure 8 show that band pass / low-pass channels with distinct bandwidths consistently appear in all cases, and the RMSE with human CSF (25 ± 2), mean and standard deviation, stays in the low range of the values found in experiment 1.

Regarding the evolution of the CSFs with contrast, it is important to note that for some conditions (low blur and high noise) the gain in the achromatic CSF increases with contrast, which is equivalent to contrast response curves that are not saturating.

Experiment 3: CSFs from chromatic adaptation versus biodistortion compensation

Figure 9 (top row) shows the CSFs emerging when the representative shallow network with human-like behavior in experiment 1 is trained for a range of alternative low-level tasks, some involving compensation of the retinal degradation (first, second, and third cases), and some others only involving chromatic adaptation (fourth and fifth). The corresponding learning curves for the models and visual examples of reconstruction are given in the Appendix A (Figure 18). Table 2 (top) summarizes the numerical performance of the models in this experiment (ε_{LMS} for the computational goal, and RMSE for the CSFs).

First, lets focus on the case where the determination of the CSF faithfully follows Equation 3 and hence we have a realistic eye+retina degradation (solid lines). Results show that only the cases where the task involves the biodegradation imply a clear difference in bandwidth between the achromatic and the chromatic channels. In the cases where there is only chromatic adaptation, the three CSFs are wider and of similar bandwidth. This behavior is clearly non-human, as confirmed by the RMSE measures in the fourth and fifth panels at the right.

Second, this difference is more clear in the idealized cases, Equation 4, where clean sinusoids are used to determine the CSFs (dashed lines). In this situation, the CSFs of the purely chromatic goals are wider and flatter indicating that the networks are not performing any particular spatial modification in any chromatic channel. As a result, the RMSE values for the chromatic adaptation cases (light style numbers below the frequency axis) substantially increase indicating poorer description of human CSFs. In this regard, the errors for the cases in which the task involves biodegradation are lower, but they are even lower if the CSF is measured considering the realistic degradation in the input.

In summary, the results show two trends. On the one hand, human-like features emerge in the CSFs if the degradation–compensation task is considered, but they do not if only chromatic adaptation is considered. On the other hand, the CSFs are closer to human in RMSE if the determination takes the retinal degradation into account in the sinusoids.

Finally, there is an interesting human chromatic feature that is well-captured by all the CNN models that were trained for chromatic adaptation: all of them display a sort of Von-Kries modification of the red–green and yellow–blue channels. Note that, when the red illuminant has to be compensated (third and fifth cases), the red–green CSF is attenuated while the

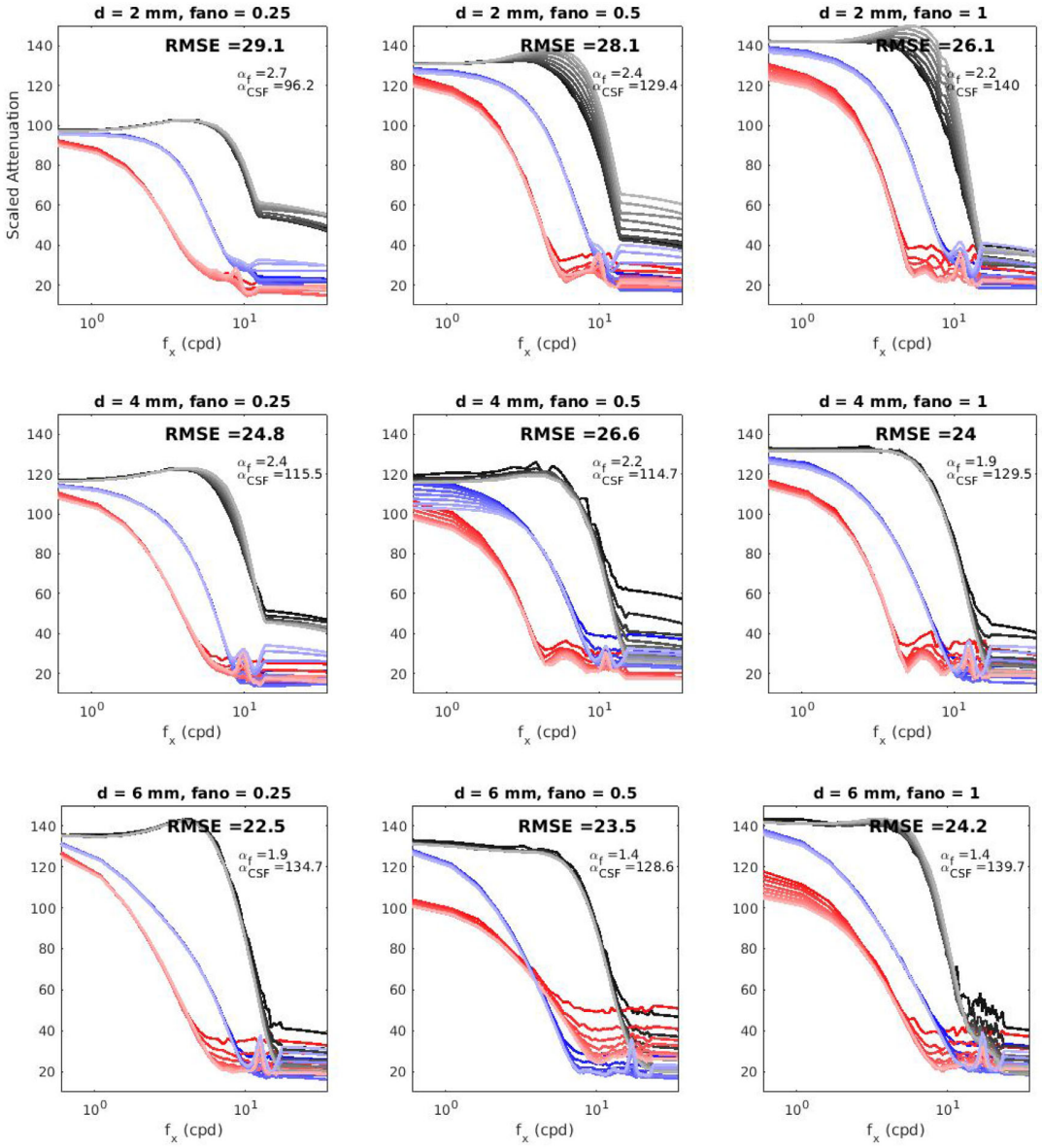


Figure 8. Experiment 2: Consistency of human-like result over a range of retinal degradations. Spatiochromatic CSFs of the two layer ReLU net for a range of retinal degradations. The RMSE distortion over the nine retinal conditions is 25 $\pm$ 2 (mean and standard deviation). The line style conventions and meaning of the numbers is the same as in experiment 1.

blue–yellow CSF is boosted, and the other way around in in the compensation of a bluish illuminant (second and fourth cases, where the blue–yellow channel is attenuated).

Experiment 4: Consistency of spatiochromatic CSFs under changes of signal statistics

Figure 9 (bottom row) shows the CSFs emerging when the representative shallow network with human-like behavior in experiment 1 is trained for the range of low-level tasks considered in experiment 3 optimizing the performances over cartoon images (as opposed to regular photographic images). The corresponding learning curves for the models and visual examples of reconstruction are given in the Appendix C (Figure 19). Table 2 (bottom) summarizes the numerical performance of the models in this experiment.

The parallelism in the results of experiments 3 and 4 confirms the robustness of the behaviors shown in experiment 3 to certain changes in signal statistics. Note that this parallelism does not mean that the CSFs are independent of the signal statistics. It just means that they are invariant to this change of statistics. It is important to remark that the low-level statistics of these (apparently different) sources may not be that different. Colors of the Pink Panther images are

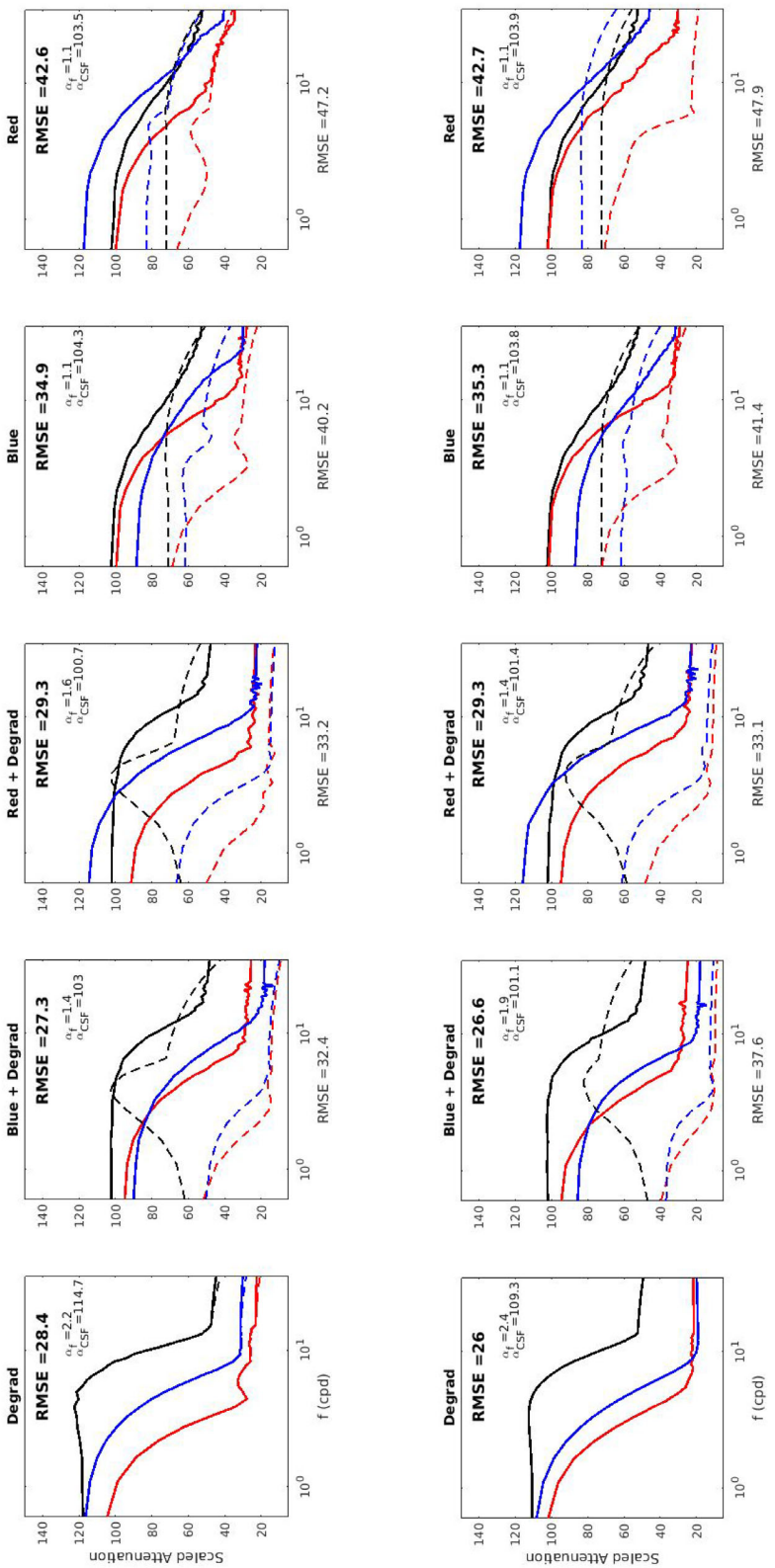


Figure 9. Experiment 3 (top row): CSFs from biodistortion compensation versus chromatic adaptation in natural images. From left to right, we consider the CSFs that emerge from (a) the compensation of biodegradation (left plot), (b) combinations of chromatic adaptation and degradation compensation (second and third plots), and (c) pure chromatic adaptation panels (fourth and fifth plots at the right). The cases in the columns first-to-third involve retinal degradation (with or without color adaptation) while the cases in the fourth and fifth columns only involve chromatic adaptation. For the sake of clarity, only low contrast results are shown. Solid lines

→

←
correspond with CSFs determined using Equation 3 (where the input to the network includes realistic acquisition process). Dashed lines correspond with CSFs determined using Equation 4 (where the network is probed with clean sinusoids). The RMSE values in bold correspond with the CSFs determined in realistic conditions (curves in bold). The RMSE values displayed below the frequency axis correspond to the CSFs determined with clean sinusoids (dashed lines). In the cases involving chromatic adaptation the input sinusoids were color shifted according to the corresponding change in the illuminant. Experiment 4 (bottom row): Consistency of the results for different signal statistics (cartoon images). The computational goals are the same as above. The only difference is that the models are trained with cartoon images (from the Pink Panther show) as opposed to regular photographic images from ImageNet (used in experiments 1–3).

Natural images					
Comput. goals	Degradat. only	Degradat. + blue adapt.	Degradat. + red adapt.	Blue adapt.	Red adapt.
Original ε_{LMS}	12.1 ± 0.2	18.1 ± 0.2	15.2 ± 0.3	13.2 ± 0.2	10.2 ± 0.3
ε_{LMS} after 2-layers ReLU	9.79 ± 0.08	8.7 ± 0.1	8.9 ± 0.1	1.92 ± 0.05	1.01 ± 0.01
RMSE of CSFs					
$\text{CSF}_{\text{realist.}}$	28.4	27.3	29.3	34.9	42.6
$\text{CSF}_{\text{simplist.}}$		32.4	33.2	40.2	47.2
Cartoon Images					
Comput. Goals	Degradat. Only	Degradat. + Blue Adapt.	Degradat. + Red Adapt.	Blue Adapt.	Red Adapt.
Original ε_{LMS}	12.9 ± 0.2	18.1 ± 0.1	15.6 ± 0.2	13.4 ± 0.3	9.1 ± 0.3
ε_{LMS} after 2-layers ReLU	10.14 ± 0.08	9.5 ± 0.1	9.2 ± 0.1	1.34 ± 0.02	0.837 ± 0.06
RMSE of CSFs					
$\text{CSF}_{\text{realist.}}$	26.0	26.6	29.3	35.3	42.7
$\text{CSF}_{\text{simplist.}}$		37.7	33.1	41.4	47.9

Table 2. Experiment 3 (top). Compensation of biodegradation versus chromatic adaptation in natural images. Performance in the goals and eventual human-like CSFs are described by ε_{LMS} and RMSE, respectively. $\text{CSF}_{\text{realist.}}$ and $\text{CSF}_{\text{simplist.}}$ refer to the way the CSF is computed (taking into account or neglecting the retinal degradation in the sinusoidal stimuli). Experiment 4 (bottom). Consistency under change of image statistics. The considered goals and magnitudes have the same meaning as in experiment 3. The only difference is in the scenes used to train the models.

certainly more saturated, but beyond this obvious fact, other differences may be subtle. In particular, we took precautions to get frames from a 5-hour compilation where backgrounds around the whole chromaticity range appear not to bias the chromatic CSFs. Regarding the spatial content, note that there are plenty of edges of arbitrary orientations and also low-frequency transitions and shadows in the cartoons. More radical modifications of spatial information (e.g., edit the cartoon images to make them isoluminant; i.e., zero contrast in the achromatic channel) could lead to substantial variation of the CSFs, but the goal of this illustration is to point out the robustness of the result more than look for its limits.

Experiment 5: CSFs from bottleneck compensation versus biodistortion compensation

Figure 10 shows the CSFs of the systems that emerge from the architectures with bottlenecks considered in Figure 4 (right), when considering two different functional goals: 1) pure reconstruction of the input signal, that is, compensation of the information loss imposed by the bottleneck, and 2) compensation of the bottleneck together with compensation of the biodistortion. Table 3 summarizes the distortions in the CSFs, RMSE, and the performance in the reconstruction, ε_{LMS}.

The Appendix C, Figure 20, confirms that these architectures converged to a plateau of ε_{LMS}. Moreover, consistently with the data in Table 3, Figures 20 and 21 show that these systems achieve the computational goals to an extent that depends on the severity of the bottleneck in a very intuitive fashion (see comments in Appendix C).

More interesting is what happens to the emerging CSFs in Figure 10. In the absence of a bottleneck, pure reconstruction leads to wide filters equal in the three chromatic channels, a clearly non-human result with RMSE approximately 40 (architectures A and B). Similarly to pure chromatic adaptation, unconstrained pure reconstruction induces no spatial

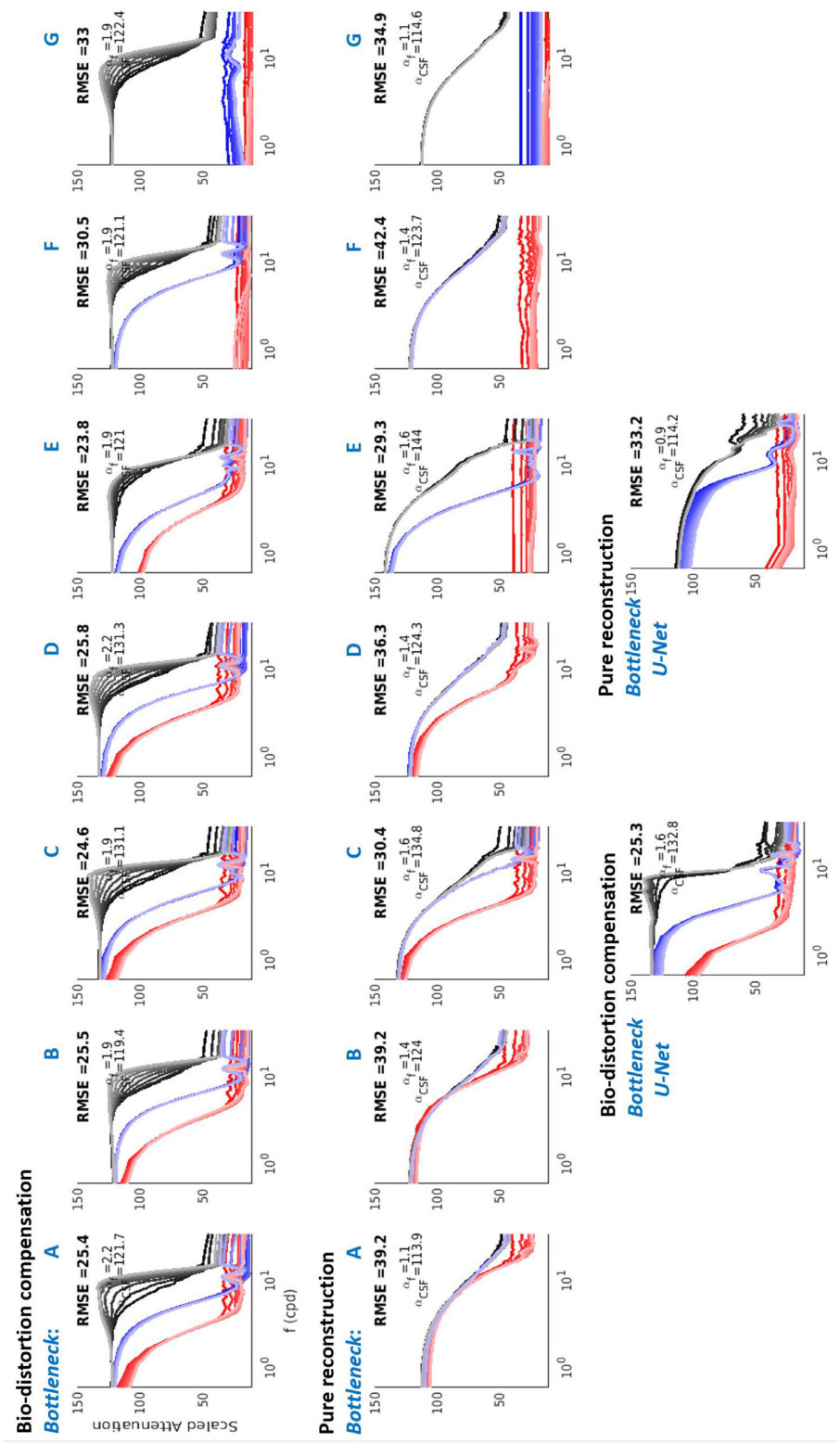


Figure 10. Experiment 5: CSFs from architectures with bottlenecks. Notation of the architectures with bottlenecks (letters in blue) refers to the diagrams shown in Figure 4. (Top row) CSFs emerging from bottleneck compensation and biodistortion degradation in progressively more restrictive bottlenecks in a four-later architecture. (Middle row) CSFs emerging from pure reconstruction of the signal in the same architectures. (Bottom row) CSFs emerging in U-Nets from biodistortion compensation (left) or pure reconstruction (right).

	No bottleneck		Bottlenecks					
	A	B	C	D	E	F	G	U-Net
Bio-Distort.								
ε_{LMS}	9.1 ± 0.3	9.2 ± 0.2	9.6 ± 0.1	9.8 ± 0.2	10.3 ± 0.4	11.2 ± 0.2	15.6 ± 0.5	12.3 ± 0.3
RMSE	25.4	25.5	24.6	25.8	23.8	30.5	33.0	25.3
Pure Recons.								
ε_{LMS}	1.2 ± 0.1	2.2 ± 0.1	4.8 ± 0.3	3.0 ± 0.5	12.8 ± 0.7	5.6 ± 0.8	11.9 ± 0.5	12.2 ± 0.6
RMSE	39.2	39.2	30.4	36.3	29.3	42.4	34.9	33.2

Table 3. Experiment 5: Compensation of biodistortion versus pure reconstruction in networks with bottlenecks. The bottlenecks in the architectures A through G and U-Net are described in Figure 4 (right). Performance in the reconstruction goals is measured by the ε_{LMS} error (in a test set) and the quality of the CSFs is given by the RMSE error. The considered biodistortion was the central case in Figure 3 with an original level of $\varepsilon_{LMS} = 15.5$, which is reduced to the values reported in the first row after the application of the models. In the pure reconstruction case the original retinal distortion is $\varepsilon_{LMS} = 0$ so the errors reported in the 3rd row come from poor reconstruction or an incomplete compensation of the bottleneck.

selectivity and hence small similarity with human vision. Mild bottlenecks restricting the number of features and/or the spatial resolution do introduce differences in the bandwidth of the achromatic/chromatic channels, but the shape of the filters is far from human (RMSE $\geq$ 30 in architectures C and D). Then, more severe bottlenecks (architectures E–G and U-Net) quickly leads to over-attenuation of one or both chromatic CSFs (and hence non-human behavior with a RMSE of approximately 35 in these architectures for reconstruction). On the other hand, the very same architectures trained for the compensation of biodistortion lead to more human-like CSFs. See the band-pass/low-pass shape of the achromatic/chromatic CSFs and the RMSE of approximately 25, except for architectures F and G that overattenuate the chromatic CSFs but still preserve the band-pass nature of the achromatic channel. Better preservation of chromatic CSFs by the systems tuned to compensate the biodistortion is visually confirmed by the reconstructions of a representative image in Appendix C, Figure 21.

In summary, pure reconstruction with the explored bottlenecks induces a difference between the relative bandwidths of the achromatic and chromatic CSFs. However, the results become closer to human (both in the shape of the filters and in RMSE) when considering the compensation of the biodegradation of the retinal signal. And this resemblance remains even if the system is not constrained by a bottleneck.

Experiment 6: Spatiochromatic–temporal CSFs from biodistortion compensation

Figure 11 shows the attenuation factors found for low-contrast moving sinusoids (both achromatic and chromatic) in the plane (f_x, f_t) for a range of 3D CNN autoencoders and for the optimal linear solution. Experimental human CSFs for achromatic moving gratings (Kelly, 1979), and for chromatic moving gratings (Díez-Ajenjo et al., 2011) are also included as a useful reference. The learning curves for the models and visual examples of reconstructions are given in the Appendix C (Figure 22). Table 3 summarizes the numerical performance of the models in this experiment.

The CSF results show that the main feature of the spatiotemporal human window of visibility (its diamond shape), with smaller spatial bandwidth for higher temporal frequencies (or speeds) (Kelly, 1979; Watson, 2013) is reproduced by all the models as well as the substantially lower bandwidth of the chromatic channels, focused on very low spatiotemporal frequencies. The error of the best net is a RMSE of 17% of the maximum sensitivity.

Consistently with the results found in images (experiment 1), resemblance with human CSFs is bigger in shallower models (linear, two layers with a RMSE of approximately 17% or 18%, respectively) than in deeper models (six layers, and eight layers with a RMSE of approximately 22%) despite the performance of the deeper models in the goal is substantially better than the performance of the linear or the two layer model. The major differences are in the scaling of the chromatic CSFs: note that deeper models over attenuate the chromatic patterns. The RMSE measures confirm the superiority of the shallower solutions. For instance, note that the over attenuation of the red–green channel in the CNNs implies that the greenish hue of the background in the visual example of Figure 22 fades away, while it does not in the linear solution (which has obvious problems in other respects).

The linear solution cannot display a contrast dependent behavior, but the two layer architecture displays a consistent decay of the gain with contrast that is in line with the saturating nature of contrast response curves of humans for moving sinusoids

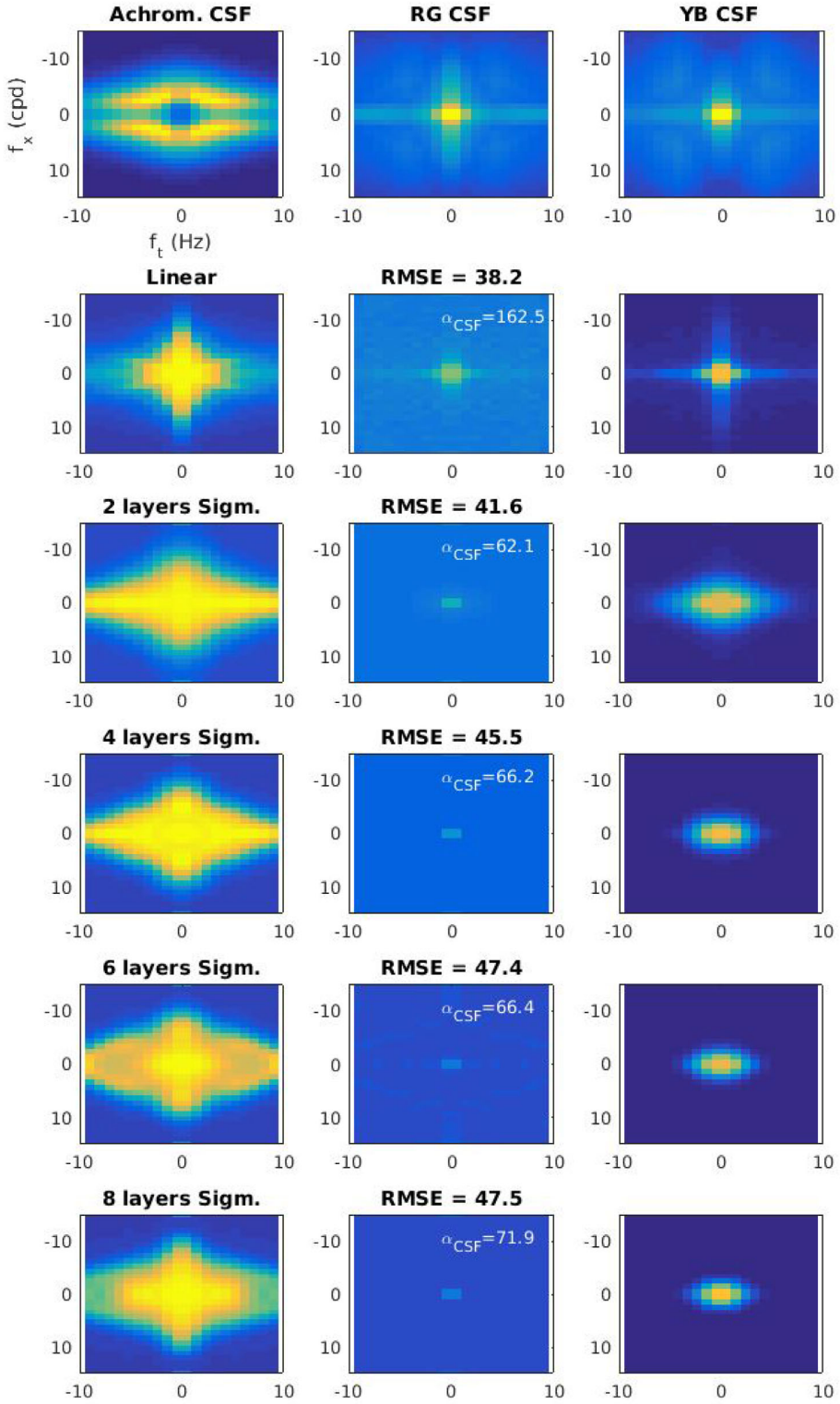


Figure 11. Experiment 6: Spatiotemporal chromatic CSFs from biodistortion compensation. The first row shows the human CSFs in (f_x, f_t), and the following show the model CSFs. The RMSE numbers (average over channels) represent the distance between them. In this experiment, $\alpha_f = 1$ in all cases.

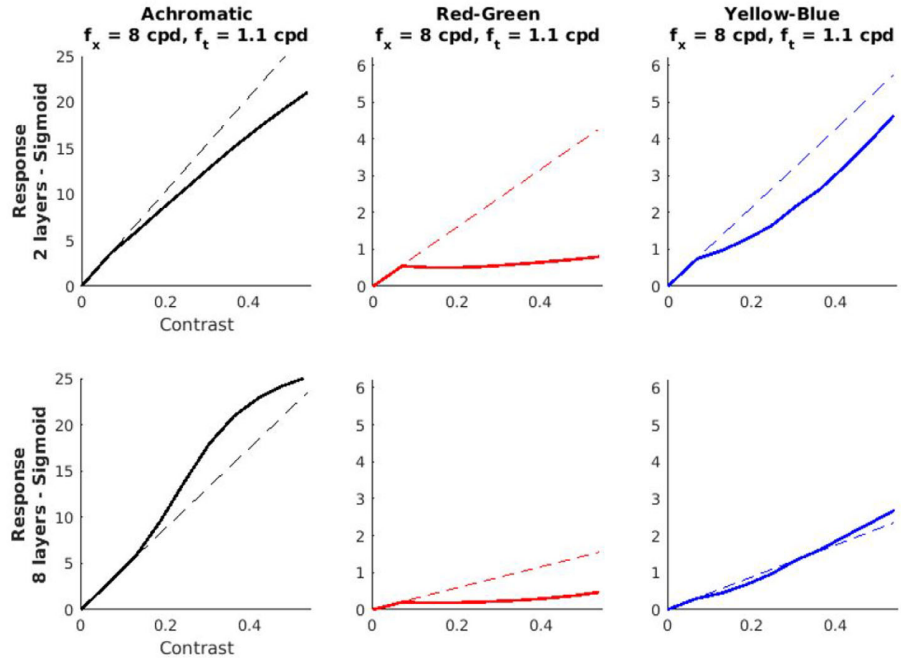


Figure 12. Experiment 6: Representative examples of nonlinear responses for spatiotemporal achromatic and chromatic gratings in shallow (top) and deeper (bottom) networks for biodistortion compensation.

	Comput. goal ε_{LMS}	CSFs RMSE
Distortion	5.2 ± 0.1	
Linear Net	3.5 ± 0.2	38.2
CNNs	Sigmoid nonlinear	Sigmoid nonlinear
2-Layers	2.07 ± 0.06	41.6
4-Layers	1.96 ± 0.07	45.5
6-Layers	1.83 ± 0.06	47.4
8-Layers	1.89 ± 0.07	47.5

Table 4. Experiment 6: Emergence of spatiotemporal chromatic CSF in 3D CNNs for compensation of biodistortion. The measures of the achievement of the goal ε_{LMS} and the distance with human behavior (RMSE difference between model and human CSF) have the same meaning as in the rest of experiments. The degradation included in the movies was the same as the considered for images (same pupil diameter and Fano factor). However, the numerical ε_{LMS} deviations turned out to be lower because the considered movies are *darker* and hence smaller LMS values lead to substantially smaller Poisson noise.

(Simoncelli & Heeger, 1998; Morgan et al., 2006). Figure 12 shows illustrative examples of these response curves: Although the two layer network (top row) consistently displays saturating behavior, the deeper net (bottom row) shows greater variability on the shape of the response.

As in the image case, the deviations in the chromatic CSFs in deep models do not come from not fulfilling the goal or having poor convergence in the training. First, all models (even the linear one) do reduce the error of the original retinal degradation so they are solving the computational problem. And second, the learning curves in the Appendix C (Figure 22) show that all models achieved a plateau in the training thus indicating proper convergence. Moreover, the asymptotic values achieved in the learning are consistent with the ε_{LMS} in test shown in Table 4.

Discussion

Summary of results

In these experiments, we trained a range of CNN autoencoders over natural scenes to solve different low-level vision goals: the compensation of retinal distortions, the compensation of changes in the illumination, the compensation of information loss after simple bottlenecks (or pure reconstruction after bottlenecks), and combinations of these.

Following the analysis of linearized networks presented in Gomez-Villa et al. (2020), it makes sense to stimulate these nets with achromatic, red–green and yellow–blue isoluminant sinusoids and moving sinusoids. The attenuation suffered by these gratings shows that:

- Human-like CSFs may emerge in systems that compensate retinal distortion: specifically, 2D shallow autoencoders trained to compensate retinal distortion display narrow low-pass behavior in the chromatic channels and wider band-pass behavior in the achromatic channel, so the shape and relative bandwidth of these artificial CSFs resemble those of humans (Figures 5, 7, and 8). Of course the match is not complete: the best CSFs obtained from the explored CNNs still deviate from human CSFs (RMSE of approximately 11% of the maximum sensitivity). Deeper autoencoders for the same goal also show CSFs with these basic shapes but the resemblance with human CSFs is consistently lower (RMSE of approximately 15% of the maximum sensitivity), particularly due to poor scaling of the chromatic CSFs (Figures 5 and 7).
- Artificial CSFs obtained from the compensation of retinal distortion differ from human CSFs in two qualitative aspects: a) The decay of network sensitivity found at low frequencies for achromatic gratings is not as big as in humans, and b) The relative amplitude of the red–green and the yellow–blue CSFs in the networks is inverted with regard to the humans. In our networks, the yellow–blue CSF is always bigger than the red–green CSF, and interestingly, this is pretty consistent over different architectures and datasets with different image statistics.
- Similar sensitivities consistently appear in shallow autoencoders for a range of levels in retinal distortions (Figure 8).
- Human-like CSFs with distinct bandwidths in achromatic/chromatic channels do not appear in pure chromatic adaptation tasks, but they do as soon as the retinal distortion compensation goal is considered (with or without chromatic adaptation). The compensation of chromatic shifts together with the compensation of biodistortion leads to systems in which the chromatic CSFs change their global gain similarly to a Von-Kries mechanism (Figure 9, top).
- CSFs emerging from chromatic adaptation and degradation compensation goals are similar for natural images and cartoon images (Figure 9, bottom).
- Pure reconstruction in architectures with a restrictive bottleneck induces changes in the relative bandwidths of the achromatic and chromatic CSFs with regard to trivial all-pass filters found in systems without bottleneck. However (in the explored cases), these CSFs are remarkably non-human. Interestingly, the very same architectures lead to more human-like CSFs as soon as the retinal distortion compensation goal is considered (Figure 10).
- The 3D autoencoders for retinal degradation compensation display a wide diamond-shaped achromatic bandwidth and very narrow chromatic bandwidths in the spatiotemporal Fourier domain, in parallel with humans. And this similarity is larger in the linear and shallow autoencoders (RMSE of approximately 17% of the maximum sensitivity) while it decays for deeper networks (RMSE of approximately 22% of the maximum sensitivity), again owing to poor scaling of the chromatic CSFs (Figure 11).
- The gain in shallow autoencoders decays with contrast and hence the contrast responses for gratings saturate with contrast. This happens both in the spatial and the spatiotemporal cases (Figures 6 and 12). This resembles contrast masking in humans. However, in deeper autoencoders this consistent saturation (and hence similarity with humans) is not found.

The emergence of human-like features in the CSFs (distinct bandwidth and shape of achromatic and chromatic channels) is related to the different properties of achromatic and chromatic patterns in visual scenes. The statistical unbalance towards achromatic patterns is known from long ago in terms of variance (Ruderman et al., 1998) and, more recently, it has been quantified in accurate information theoretic units (Malo, 2020). The eventual problems in preserving the saturation (or poor scaling of chromatic CSFs) in deeper models, do not come from training. Note that, according to the learning curves, all the models achieved proper convergence. On the contrary, the problems may come from the small (statistical) relevance of chromatic textures as opposed to the achromatic textures and the inability of deeper models to deal with this unbalance with a low-level ε_{LMS} goal: (too) flexible networks optimized to compensate the distortions focus (too much) on the spatial achromatic information to optimize the goal and are likely to distort chromatic information. The consequence is a negative impact on the chromatic CSFs. This does not seem to be a problem for more rigid shallower architectures and even the linear solution.

At this low abstraction level, where the minimization of distortion in LMS is simply connected with information maximization, and in the set of architectures considered, shallow networks seem more appropriate to explain the CSFs.

Relation to other accounts of the CSFs

Our results revisit classical work on the statistical grounds of the CSFs (Atick et al., 1992; Atick & Redlich, 1992; Atick, 2011) in light of the new possibilities provided by automatic differentiation.

From the technical point of view, a number of assumptions that had to be done in the 1990s, either have been confirmed with the use of large data sets, or are not necessary with the use of more flexible models. In particular, regarding the signal, Atick et al. assumed translation invariance, independence between color and space-time, and second-order relations (autocorrelation with $1/|f|^2$ decay). Moreover, regarding the model, they restricted themselves to linear solutions similar to Wiener filters. More recent studies with colorimetrically calibrated scenes (Gutmann et al., 2014) have confirmed the correctness of the color/space independence assumption. However, the focus on the power spectrum and the linear solutions has proven to be too limited for denoising (Gutiérrez et al., 2006). Adaptive (nonlinear) models that take into account additional features of the signal are required. Nevertheless, the nonlinear networks considered here turn out to be roughly translation invariant (Gomez-Villa et al., 2020), as expected from their convolutional architectures and the stationary nature of the problems they face. Another technical difference is in the formulation of the statistical goal: Atick et al. maximized the mutual information between the clean signal and the response $I(\boldsymbol{x}_c, \boldsymbol{y})$; while here we minimized ε_{LMS} between the clean signal and the response, $|\boldsymbol{X}_c - \boldsymbol{Y}|_2$. These goals are exactly equivalent when the difference between clean signal and the response is Gaussian, which is not the case in general. However, note that these goals are always related because the limit $|\boldsymbol{X}_c - \boldsymbol{Y}|_2 \to 0$ implies $I(\boldsymbol{x}_c, \boldsymbol{y}) \to \infty$. Beyond the spatiochromatic case, in our work we check the emergence of the CSFs with spatiotemporal signals, which was mentioned but not addressed by Atick et al. Finally, the consideration of nonlinear models allows us to show that the error-minimization goal may also lead to saturation of the contrast responses that, of course, was not possible in the linear framework of Atick et al.

As stated in the Introduction, other group has been working independently on the emergence of CSFs in artificial neural networks (Akbarinia et al., 2021). Their results are restricted to the spatial CSF (no chromatic nor spatiotemporal cases) and are based on networks trained for higher level goals, such as classification. Therefore, their results from higher level goals are complementary to ours, obtained from lower-level goals intended for the analysis of early visual stages such as the LGN. More generally, higher level goals such as classification performance may be an indirect way to impose preservation of colorfulness or a proper scaling of the chromatic CSFs. Although chromatic information may have small relevance to minimize ε_{LMS} in reconstruction, it may be more crucial for recognition.

Other works have obtained center surround sensors by optimizing a linear+nonlinear network with a low-level infomax+energy goal (Karklin & Simoncelli, 2011), or deeper nets with higher level classification goals (Lindsey et al., 2019). These sensors could induce CSF-like bandwidths in the corresponding models but this aspect was not addressed in these works.

Individual non-Euclidean distances from the optimization of average Euclidean distance

An interesting consequence of our low-level result is that the Euclidean measure, ε_{LMS}, averaged over the set of natural images leads to systems that measure individual differences in non-Euclidean ways. Note that in the systems that we trained given two input signals $\boldsymbol{x}$ and $\boldsymbol{x} + \Delta\boldsymbol{x}$, the difference between the corresponding responses is $\Delta\boldsymbol{y} \approx \boldsymbol{M} \cdot \Delta\boldsymbol{x}$. As a network should assess the difference between the two signals from $\Delta\boldsymbol{y}$, the perceived difference for the system will depend on $\boldsymbol{M}$. Specifically (Epifanio et al., 2003; Laparra et al., 2010), the perceived distance for the system will depend on the metric, $\boldsymbol{M}^\top \cdot \boldsymbol{M}$, and hence it will depend on λ^2 (the eigenspectrum of $\boldsymbol{M}$) or, as seen here, on the CSF^2. This metric is non-Euclidean: for instance, high-frequency distortions will be less relevant for the network than medium or low-frequency distortions. Even though here we did not check the correlation between the image distortions perceived by humans and networks, the observed CSFs in the networks are consistent with the fact that the Euclidean distance between images at the retina is not a good representation of human distortion metrics (Wang & Bovik, 2009; Laparra et al., 2010; Hepburn et al., 2020).

The emergence of a non-Euclidean distance from the minimization of the average Euclidean distance over natural images is a counterintuitive consequence of the highly nonuniform distribution of natural scenes: distortions in less populated regions of the image space (e.g., in high-frequency directions or in chromatic channels) have to be under-rated to favor the average match to the data in highly populated or more informative regions (low-frequency, achromatic patterns).

Recent work on autoencoders with low-level rate-distortion constraints on natural images has shown the emergence of non-Euclidean distances correlated with human opinion of distortion (Hepburn et al., 2022). Human opinion of distortion is known to be strongly mediated by the CSFs, but the bandwidth of this autoencoder and its eventual similarity with the CSF was not explored in that work.

Alternative low-level computational goals

Here we considered different low-level alternatives to the retinal signal enhancement goal proposed by Atick et al.: although our results are conclusive regarding the

role of chromatic adaptation, more research is definitely required about the relative relevance of bottlenecks in shaping the CSFs.

On the one hand, given the small role of spatial information in the changes of the LMS image purely owing to changes in illumination, it is not surprising that systems designed for pure chromatic adaptation have wide (all-pass) behavior in all channels (i.e., no spatial effect). As a consequence (as confirmed by our results), pure chromatic adaptation does not lead to CSFs with a human-like shape. This shape and relative bandwidths have to be related to other goals (e.g., the compensation of biodistortion). However, training for chromatic adaptation does introduce an important human-like behavior (which may not emerge from other tasks): it leads to adaptive global scaling of the red–green or yellow–blue CSFs. This effect is consistent with the observations done on spatiochromatic adaptation under changes in spectral illumination (Gutmann et al., 2014): the spatial structure of the receptive fields remains almost constant but their chromatic tuning basically changes according to Von-Kries adaptation.

On the other hand, as opposed to chromatic shifts, the spatial effect of bottlenecks is relevant. However, we only explored a small range of architectural bottlenecks: the toy examples of Figure 4-right. In this restricted set our results suggest that the compensation of the biodegradation at the retina plays a stronger role in the emergence of human-like CSFs than the consideration of the bottlenecks. However, bottlenecks in architectures C, D, and U-Net favor the emergence of nontrivial frequency selectivity. This would be consistent with Lindsey et al. (2019), who reported positive effects of bottlenecks in the emergence of center surround receptive fields. Nevertheless, the specific configuration of the bottleneck that maximizes the human nature of the CSFs and the relative role of bottlenecks in the compensation of the retinal distortion are interesting matters for further research.

More generally, other low-level goals could be considered together with the distortion, as for instance the information or the energy of the signal. Architectural bottlenecks considered here or in Lindsey et al. (2019) indirectly constrain the energy and the entropy of the signal by reducing the dimensionality of the signal. However, one could consider more general factors beyond the dimensionality as for instance the neural noise, the PDFs of signal and noise and the redundancy of the visual signals in the representation. In fact, transmitted information may be modulated by changes of the representation and by the amount of noise even without changes in the dimensionality (Malo, 2020).

In a separate study (Hepburn et al., 2022) we have shown that rate-distortion bottlenecks in autoencoders induce distance measures which are correlated with subjective opinion of distortion. The autoencoders we presented here do not include constraints on information, but the emergence of a non-Euclidean metric depending on M (and hence on the CSFs) suggests that the distance will be correlated with human opinion in line with Hepburn et al. (2022).

Alternative low-level goals could include non-human retinal degradation. Other species have different optical quality and noise in their retinas may be substantially different. This may affect the kind of computations required to extract the appropriate information from this degraded input, and hence their corresponding CSFs.

All these issues, the specific impact of more sophisticated bottlenecks in the CSFs, which was not analyzed here or in Karklin and Simoncelli (2011), Lindsey et al. (2019), or Hepburn et al. (2022), the emergence of human-like image distortion measures from the enhancement of retinal signals, and the consideration of retinal degradation for other species, is a matter for future research.

Goal and architecture are not independent

More important than the technical generalizations over Atick et al. (1992), Atick and Redlich (1992), and Atick (2011), is that the current freedom to explore different linear and nonlinear architectures stresses the relevance of the architectural constraints. The conventional interpretation of the efficient coding hypothesis (Barlow et al., 1961) is the following: obtaining human-like results from certain statistical goal seems to suggest that the human visual system has been shaped by this goal. However, it is important to realize that the results have been obtained via the optimization of certain model. In the case of Atick et al., it was a single model (the linear filter), but in our case here we tried a range of models (architectures). Because the results for the different architectures is markedly different, the conclusion can not be about the goal, but about specific combinations of goal and architecture. Our results are a specific illustration of the fact that the computational and the algorithmic levels of analysis of visual processing systems (Marr & Poggio, 1976; Marr, 1982) are not independent (Poggio, 2021). This dependence prevents about premature conclusions about the organization principles at the computational level if sensible architectures are not adopted.

Beyond accuracy

Human-like CSFs are obtained for shallow autoencoders (two layers), or even linear networks, despite deeper architectures achieving similar or better performance in the goal. The previous literature has warned about the limitations of a single

accuracy/performance measure to identify human-like behavior. Achieving similar performance on a task does not guarantee that two models actually use the same strategy (Firestone, 2020). For instance, different strategies may become evident if performance degrades in different ways when changing the experimental setting (Wichmann et al., 2017; Geirhos et al., 2019). Therefore, additional checks different from the optimization goal have to be done in order to confirm the human-like behavior of a model. Examples include verifying additional psychophysics not included in the goal (Martinez et al., 2019), or disaggregating the results checking the consistency between model and humans in individual trials, not on averages over the data set (Geirhos et al., 2020).

In this complexity/accuracy discussion, it is important to stress that our results (shallower networks better reproduce the scale of human chromatic CSFs) is in line with the results of Gomez-Villa et al. (2020) and Flachot et al. (2020), which also show that shallower networks obtain more human-like colour representation. In a similar vein, although using higher level classification goals, Kubilius et al. (2019) show that lower performance networks may correlate better with human brain activity or psychophysics.

Final remarks

In visual neuroscience, deep models are emerging as the new standard to reproduce the activity of visual areas under natural scene stimulation. On the one hand, conventional deep models driven by object recognition goals reproduce the response from V1 (Güçlü and van Gerven, 2015; Kriegeskorte, 2015), dorsal and ventral streams (Cichy et al., 2016), and IT (Cadieu et al., 2014; Yamins et al., 2014). On the other hand, deep networks are powerful enough to fit the mappings between stimuli and measured responses (Prenger et al., 2004; Antolík et al., 2016; Batty et al., 2017). These two approaches (goal-driven and measurement-driven deep models) have been thoroughly compared in V1 and were found to be superior to linear filter-banks and simple linear–nonlinear models (Cadena et al., 2019). However, more recently, the same team has shown that linear–nonlinear models with general divisive normalization make a significant step towards the performance of state-of-the-art CNN with interpretable parameters (Burg et al., 2021).

In our low-level goal-driven case, the emergence of human-like CSFs for certain CNN autoencoders generalizes in different ways previous statistical explanations of the CSF based on linear models (Atick & Redlich, 1992; Atick et al., 1992; Atick, 2011), and is consistent with optimizations of nonlinear encoders using alternative low-level (Karklin & Simoncelli, 2011; Hepburn et al., 2022) or higher level (Lindsey et al., 2019; Akbarinia et al., 2021) goals. However, we find a strong dependence of the CSFs on the architecture with better results for shallower autoencoders (although they have similar or lower performance in the goal).

This is not in contradiction with the literature cited elsewhere in this article showing that deep networks with object recognition goals match very well higher visual areas. Note that the scope of our low-level goal is restricted to early visual stages (e.g., the retina–LGN path), and hence simpler architectures may be required there.

Beyond this difference in abstraction level, our results do illustrate the relevance of using appropriate architectures when checking a statistical goal. Following the move from conventional CNNs in Cadena et al. (2019) to more realistic divisive normalization models in Burg et al. (2021), we think that future goal-driven derivations of low-level visual psychophysics (e.g., pattern masking or perceptual distortion) should include more realistic architectures too, as opposed to conventional CNNs (although they may be flexible enough to fulfill the goal). Examples include divisive normalization with parametric interaction between features (Martinez et al., 2018, 2019) and generalizations of Wilson–Cowan interactions (Bertalmío et al., 2020). Learning frameworks with rate-distortion bottlenecks are already available (Ballé et al., 2017; Hepburn et al., 2022), and we advocate for the study of their artificial psychophysics using realistic and interpretable architectures.

Keywords: spatiotemporal and chromatic contrast sensitivity, convolutional autoencoders, modulation transfer function, noisy cones, deblurring and denoising, chromatic adaptation, natural images, statistical goals, architectures

Acknowledgments

Partially funded by these grants from GVA/AEI/FEDER/EU: MICINN DPI2017-89867-C2-2-R, MICINN PID2020-118071GB-I00, and GVA Grisolía-P/2019/035 (for JM and QL), and MICINN PGC2018-099651-B-I00 (for A.G.V. and M.B.).

Commercial relationships: none.
Corresponding author: Jesus Malo.
Email: jesus.malo@uv.es.
Address: Image Processing Lab, Building E4, Parc Cientific de la Universitat de Valencia, Carrer Catedràtic Escardino, 46980 Paterna, Valencia, Spain.

Footnotes

[1]For example, neurons in conventional CNNs have fixed nonlinearities, as opposed to the known adaptive nature of real neurons (Wilson & Cowan, 1973; Carandini & Heeger, 2012).

[2]For optical blur where the linear operator H can be obtained from the MTF (Watson, 2013), and the retinal noise is Poisson, $\boldsymbol{n}_r = F \cdot \mathbb{D}_{(|H \cdot \boldsymbol{x}|^{\frac{1}{2}})} \cdot \boldsymbol{n}$, where $\mathbb{D}_{\boldsymbol{v}}$ is a diagonal matrix with vector $\boldsymbol{v}$ in the diagonal, F is the Fano factor, and $\boldsymbol{n}$ is drawn from a unit-variance Gaussian (Esteve et al., 2020); the Jacobian in Equation 9, is $\nabla_x S_\theta(\mathbf{0}) = \nabla_x N_\theta(\mathbf{0}) \cdot (I - \frac{F}{2} \cdot \mathbb{D}_{(\boldsymbol{n} \odot |H \cdot \mathbf{0}|^{\frac{1}{2}})}) \cdot H$, where the Jacobian of the network, $\nabla_x N_\theta(\mathbf{0})$, can be obtained analytically (Martinez et al., 2018).
[3]We prepared the samples that way before actually knowing how well the networks are able to cope with this distortion.
[4]In our computer cluster typical training of the 2D models takes approximately 10 to 20 hours.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2016). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arxiv Comp. Sci*, https://arxiv.org/abs/1603.04467.

Akbarinia, A., Morgenstern, Y., & Gegenfurtner, K. R. (2021). Contrast sensitivity is formed by visual experience and task demands. *Journal of Vision, 21*(9), 1996.

Antolík, J., Hofer, S. B., Bednar, J. A., & Mrsic-Flogel, T. D. (2016). Model constrained by visual hierarchy improves prediction of neural responses to natural scenes. *PLoS Computational Biology, 12*(6), e1004927, doi:10.1371/journal.pcbi.1004927.

Atick, J., Li, Z., & Redlich, A. (1992). Understanding retinal color coding from first principles. *Neural Computation, 4*(4), 559–572.

Atick, J., & Redlich, A. (1992). What does the retina know about natural scenes? *Neural Computation, 4*(2), 196–210.

Atick, J. J. (2011). Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems, 22*, 4–44.

Ballé, J., Laparra, V., & Simoncelli, E. P. (2017). End-to-end optimized image compression. *International Conference on Learning Representations (ICLR)*, https://openreview.net/forum?id=rJxdQ3jeg.

Barlow, H. B. et al. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication, 1*, 217–234.

Batty, E., Merel, J., Brackbill, N., Heitman, A., Sher, A., Litke, A. M., … Paninski, L. (2017). Multilayer recurrent network models of primate retinal ganglion cell responses. *5th International Conference on Learning Representations (ICLR)*, https://openreview.net/forum?id=9gmuVOlKfLa.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., & Siskind, J. M. (2018). Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research, 18*(153), 1–43.

Berardino, A., Ballé, J., Laparra, V., & Simoncelli, E. (2017). Eigen-distortions of hierarchical representations. In *Proceedings of the Neural Information Processing Systems, 30*, pp. 3533–3542, https://papers.nips.cc/book/advances-in-neural-information-processing-systems-30-2017.

Bertalmío, M., Gomez-Villa, A., Martín, A., Vazquez, J., Kane, D., & Malo, J. (2020). Evidence for the intrinsically nonlinear nature of receptive fields in vision. *Scientific Reports, 10*, 16277.

Burg, M., Cadena, S., Denfield, G., Walker, E., Tolias, A., & Bethge, M. et al. (2021). Learning divisive normalization in primary visual cortex. *PLoS Computational Biology, 17*(6), e1009028.

Cadena, S., Denfield, G., Walker, E., Gatys, L., Tolias, A., & Bethge, M. et al. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Computational Biology, 15*(4), e1006897.

Cadieu, C., Hong, H., Yamins, D., Pinto, N., Ardila, D., & Solomon, E. et al. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Computational Biology, 10*(12), e1003963.

Cai, D., DeAngelis, G. C., & Freeman, R. D. (1997). Spatiotemporal receptive field organization in the lateral geniculate nucleus of cats and kittens. *Journal of Neurophysiology, 78*(2), 1045–1061.

Campbell, F., & Robson, J. (1968). Application of Fourier analysis to the visibility of gratings. *Journal of Physiology, 197*, 551–566.

Carandini, M., & Heeger, D. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience, 13*(1), 51–62.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports, 6*, 27755.

Clarke, R. (1981). Relation between the Karhunen Loève and cosine transforms. *IEE Proceedings F Communications, Radar and Signal Processing, 128*, 359–361.

Cottaris, N. P., Jiang, H., Ding, X., Wandell, B. A., & Brainard, D. H. (2019). A computational-observer model of spatial contrast sensitivity: Effects of wave-front-based optics, cone-mosaic structure, and inference engine. *Journal of Vision, 19*(4), 8.

Cottaris, N. P., Wandell, B. A., Rieke, F., & Brainard, D. H. (2020). A computational observer model of spatial contrast sensitivity: Effects of photocurrent encoding, fixational eye movements, and inference engine. *Journal of Vision, 20*, 17.

Donen, S. (1963). *Charade.* Universal Studios, CC Public Domain Mark 1.0, https://archive.org/details/Charade19631280x696.

Díez-Ajenjo, M., Capilla, P., & Luque, M. J. (2011). Red-green vs. blue-yellow spatio-temporal contrast sensitivity across the visual field. *Journal of Modern Optics, 58*, 1–13.

Enroth-Cugell, C., & Robson, J. (1966). The contrast sensitivity of retinal ganglion cells on the cat. *Journal of Physiology (London), 187*, 516–552.

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

Esteve, J., Aguilar, G., Maertens, M., Wichmann, F., & Malo, J. (2020). Psychophysical estimation of early and late noise. *Arxiv: Quantitative Biology,* 1–15, https://arxiv.org/abs/2012.06608.

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences of the United States of America, 117*(43), 26562–26571.

Flachot, A., Akbarinia, A., Schütt, H. H., Fleming, R. W., Wichmann, F. A., & Gegenfurtner, K. R. (2020). Deep neural models for color discrimination and color constancy. *CoRR*, *abs/2012.14402*.

Freleng, I. (1963). *The Pink Panther Show*. Los Angeles, CA, USA: DePatie-Freleng Enterprises (DFE Films), courtesy of Metro Goldwin Mayer.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *arXiv: 2006.16736*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations (ICLR)*, https://openreview.net/forum?id=Bygh9j09KX.

Gomez-Villa, A., Martin, A., Vazquez, J., & Bertalmio, M. (2019). Convolutional neural networks can be deceived by visual illusions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 19)* (pp. 12309–12317), doi:10.1109/CVPR.2019.01259.

Gomez-Villa, A., Martin, A., Vazquez, J., Bertalmío, M., & Malo, J. (2020). Color illusions also deceive CNNs for low-level vision tasks: Analysis and implications. *Vision Research, 176*, 156–174.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Boston, MA, USA: MIT Press, https://www.deeplearningbook.org.

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience, 35*(27), 10005–10014.

Gutiérrez, J., Ferri, F. J., & Malo, J. (2006). Regularization operators for natural images based on nonlinear perception models. *IEEE Transactions on Image Processing, 15*(1), 189–200.

Gutmann, M. U., Laparra, V., Hyvärinen, A., & Malo, J. (2014). Spatio-chromatic adaptation via higher-order canonical correlation analysis of natural images. *PloS One, 9*(2), e86481.

Hepburn, A., Laparra, V., Malo, J., & Santos, R. (2020). Perceptnet: A human visual system inspired neural network for estimating perceptual distance. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)* (pp. 121–125), doi:10.1109/ICIP40778.2020.9190691.

Hepburn, A., Laparra, V., Santos, R., Ballé, J., & Malo, J. (2022). On the relation between statistical learning and perceptual distances. *International Conference on Learning Representations, ICLR*, https://openreview.net/forum?id=zXM0b4hi5_B.

Hunt, B. R. (1975). Digital image processing. *Proceedings of the IEEE, 63*, 693–708.

Hurvich, L., & Jameson, D. (1957). An opponent-process theory of color vision. *Psychological Review, 64*(6), 384–404.

Hyvärinen, A., Hurri, J., & Hoyer, P., 2009. *Natural image statistics: A probabilistic approach to early computational vision*. Heidelberg, Germany: Springer-Verlag.

Ingling, C. R., & Martinez-Uriegas, E. (1983). The relationship between spectral sensitivity and spatial sensitivity for the primate r-g x-channel. *Vision Research, 23*, 1495–1500.

Karklin, Y., & Simoncelli, E. (2011). Efficient coding of natural images with a population of noisy linear-nonlinear neurons. In *Proceedings of the Advances in Neural Information Processing Systems, 24*, https://proceedings.neurips.cc/paper/2011/file/12e59a33dea1bf0630f46edfe13d6ea2-Paper.pdf.

Kelly, D. H. (1979). Motion and vision. ii. stabilized spatio-temporal threshold surface. *Journal of the Optical Society of America, 69*(10), 1340–1349.

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. *arXiv:1412.6980*.

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.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., ... DiCarlo, J. J. (2019). Brain-like object recognition with high-performing

shallow recurrent ANNs. In *Proceedings Advances in Neural Information Processing Systems, 32*, https://proceedings.neurips.cc/paper/2019/file/7813d1590d28a7dd372ad54b5d29d033-Paper.pdf.

Laparra, V., Muñoz, J., & Malo, J. (2010). Divisive normalization image quality metric revisited. *Journal of the Optical Society of America A, 27*(4), 852–864.

Legge, G. E. (1981). A power law for contrast discrimination. *Vision Research, 21*(4), 457–467.

Legge, G. E., & Foley, J. M. (1980). Contrast masking in human vision. *Journal of the Optical Society of America, 70*(12), 1458–1471.

LeRoy, M. (1959). *The FBI story.* Los Angeles, CA, USA: Warner Brothers/Seven Arts.

Lillicrap, T., Santoro, A., Marris, L., Akerman, C., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience, 21*(6), 335–346.

Lindsey, J., Ocko, S. A., Ganguli, S., & Deny, S. (2019). The effects of neural resource constraints on early visual representations. *International Conference on Learning Representations, ICLR*, https://openreview.net/forum?id=S1xq3oR5tQ.

Malo, J. (2020). Spatio-chromatic information available from different neural layers via gaussianization. *Journal of Mathematical Neuroscience, 10*(18), doi:10.1186/s13408-020-00095-8.

Malo, J. (2022). Paraphrasing Magritte's observation. *ArXiV: Computer Vision and Pattern Recognition,* 1–4, https://arxiv.org/abs/2202.08103.

Malo, J., & Luque, M. (2002). ColorLab: A Matlab Toolbox for color science and calibrated color image processing. *Servei de Publicacions de la Universitat de Valencia*. Valencia, Spain, https://isp.uv.es/code/visioncolor/colorlab.html.

Malo, J., Pons, A., Felipe, A., & Artigas, J. (1997). Characterization of the human visual system threshold performance by a weighting function in the gabor domain. *Journal of Modern Optics, 44*(1), 127–148.

Mannos, J. L., & Sakrison, D. J. (1974). The effects of a visual fidelity criterion of the encoding of images. *IEEE Transactions on Information Theory, 20*, 525–536.

Marr, D. (1982). *Vision: A computational approach*. San Francisco: Freeman & Co.

Marr, D., & Poggio, T. (1976). *From understanding computation to understanding neural circuitry (AI Memo No. AIM-357)*. Boston, MA, USA: MIT Libraries, http://hdl.handle.net/1721.1/5782.

Martinez, M., Bertalmío, M., & Malo, J. (2019). In praise of artifice reloaded: Caution with natural image databases in modeling vision. *Frontiers in Neuroscience*, doi:10.3389/fnins.2019.00008.

Martinez, M., Cyriac, P., Batard, T., Bertalmío, M., & Malo, J. (2018). Derivatives and inverse of cascaded linear+nonlinear neural models. *PLoS One, 13*(10), doi:10.1371/journal.pone.0201326.

Martinez-Otero, L., Molano, M., Wang, X., Sommer, F., & Hirsch, J. (2014). Statistical wiring of thalamic receptive fields optimizes spatial sampling of the retinal image. *Neuron, 81*(4), 943–956.

Martinez-Uriegas, E. (1994). Chromatic-achromatic multiplexing in human color vision. In D. H. Kelly (Ed.), (pp. 117–187). Boca Ratón, FL, USA: CRC Press.

Martinez-Uriegas, E. (1997). Color detection and color contrast discrimination thresholds. *In Proceedings of the OSA Annual Meeting ILS-XIII*, p. 81.

Morgan, M., Chubb, C., & Solomon, J. (2006). Predicting the motion after-effect from sensitivity loss. *Vision Research, 46*(15), 2412–2420.

Mullen, K. T. (1985). The CSF of human colour vision to red–green and yellow–blue chromatic gratings. *Journal of Physiology, 359*, 381–400.

van den Oord, A., & Schrauwen, B. (2014). The student-t mixture as a natural image patch prior with application to image compression. *Journal of Machine Learning Research, 15*(60), 2061–2086.

Poggio, T. (2021). *From Marr's Vision to the problem of human intelligence (CBMM Memo No. 118)*. Boston, MA, USA: MIT Libraries, https://dspace.mit.edu/handle/1721.1/131234.

Prenger, R., Wu, M., David, S., & Gallant, J. (2004). Nonlinear v1 responses to natural scenes revealed by neural network analysis. *Neural Networks, 17*(5-6), 663–79.

Reid, R. C., & Shapley, R. (1992). Spatial structure of cone inputs to receptive fields in primate lateral geniculate nucleus. *Nature, 356*, 716–718.

Reid, R. C., & Shapley, R. (2002). Space and time maps of cone photoreceptor signals in macaque lateral geniculate nucleus. *Journal of Neuroscience, 22*, 6158–6175.

Ruderman, D. L., Cronin, T. W., & Chiao, C.-C. (1998). Statistics of cone responses to natural images: Implications for visual coding. *Journal of the Optical Society of America A, 15*(8), 2036–2045.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., & Ma, S. et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*, 211–252.

Simoncelli, E., & Heeger, D. (1998). A model of neuronal reponses in visual area MT. *Vision Research, 38*(5), 743–761.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*(1), 1193–1216.

Soh, J., & Cho, N. (2021). Deep universal blind image denoising. In *Proceedings International Conference on Pattern Recognition (ICPR)* (pp. 747–754), doi:10.1109/ICPR48806.2021.9412605.

Stockman, A., & Sharpe, L. T. (2000). The spectral sensitivities of the middle- and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision Research, 40*(13), 1711–1737.

Tao, X., Gao, H., Shen, X., Wang, J., & Jia, J. (2018). Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8174–8182), doi:10.1109/CVPR.2018.00853.

Taubman, D. S., & Marcellin, M. W. (2001). *Jpeg 2000: Image compression fundamentals, standards and practice*. Norwell, MA: Kluwer Academic Publishers.

Valois, R. L. de, & Pease, P. L. (1971). Contours and contrast: Responses of monkey lateral geniculate nucleus cells to luminance and color figures. *Science, 171*, 694–696.

Wallace, G. K. (1992). The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics, 38*(1), xviii–xxxiv, doi:10.1109/30.125072.

Wang, Z., & Bovik, A. C. (2009). Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine, 26*(1), 98–117.

Watson, A. B. (2013). High frame rates and human vision: A view through the window of visibility. *SMPTE Motion Imaging Journal, 122*(2), 18–32.

Watson, A. B., & Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast. *Journal of Vision, 5*(9), 717–740.

Watson, A. B., & Ahumada, A. J. (2016). The pyramid of visibility. In *Proceedings Human Vision and Electronic Imaging HVEI16, 102*, 1–6, doi:10.2352/ISSN.2470-1173.2016.16HVEI-102.

Watson, A. B., Ahumada, A. J., & Farrell, J. E. (1986). Window of visibility: A psychophysical theory of fidelity in time-sampled visual motion displays. *Journal of The Optical Society of America A-optics Image Science and Vision, 3*, 300–307.

Watson, A. B., & Malo, J. (2002). Video quality measures based on the standard spatial observer. In *Proceedings of the IEEE International Conference on Image Processing* (Vol. *III*, pp. 41–44), doi:10.1109/ICIP.2002.1038898.

Welles, O. (1946). *The stranger*. New York, NY, USA: RKO Radio Picures.

Wichmann, F. A., Janssen, D. H. J., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., ... Bethge, M. (2017). Methods and measurements to compare men against machines. *Electronic Imaging, 10*, 36–45.

Wilson, H. R., & Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik, 13*(2), 55–80.

Wuerger, S. M., Ashraf, M., Kim, M., Martinovic, J., Pérez-Ortiz, M., & Mantiuk, R. K. (2020). Spatio-chromatic contrast sensitivity under mesopic and photopic light levels. *Journal of Vision, 20*, 23, https://doi.org/10.1167/jov.20.4.23.

Yamins, D., & DiCarlo, J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*, 356–365.

Yamins, D., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United State of America, 111*, 8619–8624.

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing, 26*(7), 3142–3155.

Appendix A: Implementation details

As stated in the main text, all the models follow the the basic toy networks studied in Gomez-Villa et al. (2019, 2020): autoencoders with convolutional layers made of eight feature maps with kernels of spatial size 5×5 and sigmoids or ReLUs as activation functions. As illustrated in Figure 4, the last reconstruction layer, has three features in every case (the three color channels) so that the input and output domains are the same. Following our purpose of using biologically plausible image representations the input to the networks and the output signals are expressed in the LMS color space (as opposed to generic RGB digital counts used in the cited references).

The spatiotemporal models follow the same spirit, in this case also including convolution in the temporal dimension: autoencoders with 3D convolutional layers made of eight feature maps with kernels of size $5 \times 5 \times 5$ and sigmoid activation functions (we did not explore ReLU in videos because in images we found no qualitative difference between the ReLU and sigmoid results). As in the image case, the last (reconstruction)

layer for every architecture (two, four, six and eight layers) only has three feature maps (the LMS channels).

Implementation and training is done in the same way as in Gomez-Villa et al. (2019, 2020): mean squared error is used as loss function in all cases and all the models are implemented using Tensorflow (Abadi et al., 2016). We train our models using ADAM stochastic gradient descent (Kingma & Ba, 2017) with a batch size of 32 examples, momentum of 0.9, and a weight decay of 0.001. In principle, a standard early stopping criterion for convergence was used based on the number of iterations with no improvement in the validation set. However, to ensure appropriate convergence, we visualized the learning curves and we let the iteration continue until train and validation error reached a common plateau. All the learning curves are explicitly shown in the Appendix C, these show that all the CSF considered in the main text come from models with the proper convergence.

Appendix B: Training stimuli and stimuli for CSF estimation

The natural stimuli to train the networks are regular photographic images from the same dataset used in Gomez-Villa et al. (2019, 2020): the Large Scale Visual Recognition Challenge, 2014 CLS-LOC validation dataset (which contains $50 \cdot 10^3$ images), leaving $10 \cdot 10^3$ images for validation purposes. This dataset is a subset of the whole ImageNet dataset (Russakovsky et al., 2015). The experiments with cartoon images were done using $25 \cdot 10^3$ frames taken from The Pink Panther Show (Freleng, 1963) reproduced with permission of the MGM. In every case we take 128×128 images and assume a sampling frequency $fs = 70$ cpd, that is, we assume that the images subtend 3.6°.

The spatiotemporal models are trained over $25 \cdot 10^3$ patches of size $32 \times 32 \times 25$ from classical Hollywood films which are in public domain: the color movies *Charade* (Donen, 1963) and *The FBI story* (LeRoy, 1959), and the achromatic movie *The Stranger* (Welles, 1946). In all the video cases we assume a spatial sampling of 30 cpd and temporal sampling of 25 Hz, that is, we assume the patches subtend 1.06° and last for 1 second. These somewhat arbitrary selections of the sampling frequencies (or extent of the stimuli) have mild consequences on the quantitative evaluation of the CSFs as discussed elsewhere in this appendix.

The transform from digital counts to LMS tristimulus values was done assuming the primaries and gamma curves of a standard CRT display (Malo & Luque, 2002).

Regarding the stimuli for the estimation of the CSFs according to Equation 3, our $\boldsymbol{b}^f$ are gratings in the classical opponent space of Hurvich and Jameson (1957). Figure 13 shows a representative subset of the gratings used to feed the networks for the estimation of the spatiochromatic CSFs. The justification of the use of these waves to probe the autoencoders follows the eigenanalysis of the linearized networks introduced in Gomez-Villa et al. (2020): the eigenfunctions of the matrices in Equation 8 were shown to be oscillating functions in space with chromatic variations in luminance and opponent red–green and yellow–blue directions. Consistently with Gomez-Villa et al. (2020) the corresponding spatiotemporal oscillations of increasing frequency for decreasing eigenvalue are

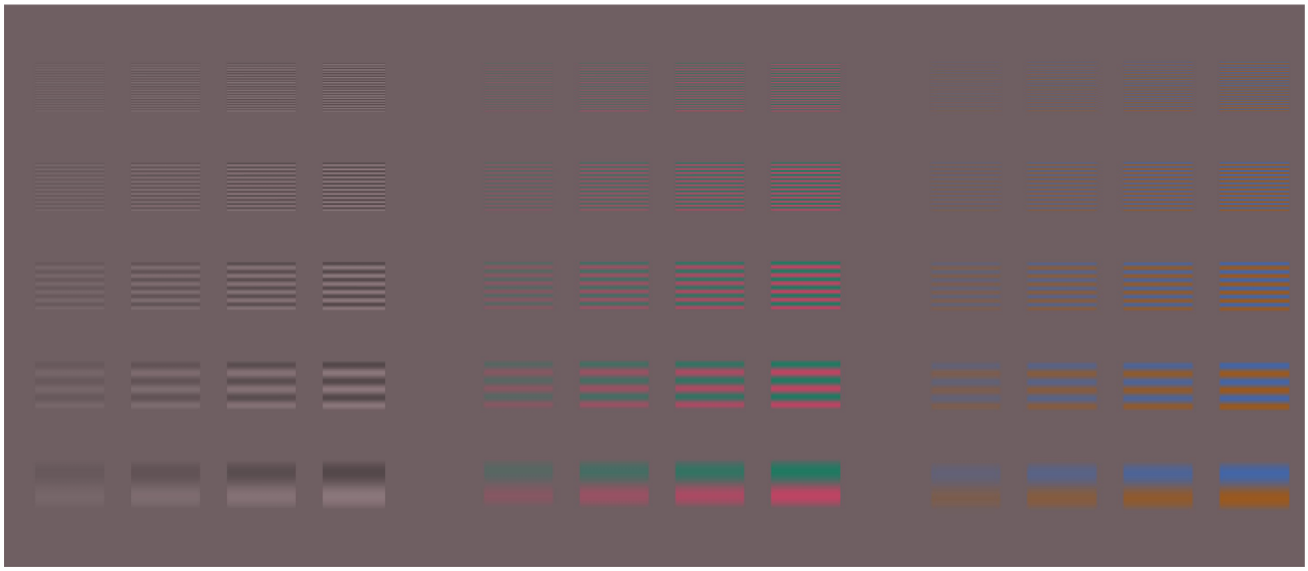


Figure 13. Representative spatiochromatic stimuli to feed the 2D networks. The 3D networks were probed with equivalent gratings mooving at different speeds (see text).

obtained when the considered Jacobian corresponds to spatiotemporal autoencoders.

The full spatiochromatic set included gratings of 60 spatial frequencies linearly spaced in the range [0.5, 35] cpd (the Nyquist region assuming $f_s = 70$ cpd), and nine contrasts linearly spaced in the range (0.07, 0.6). The average color was the white of the color system with 30 cd/m^2. The stimuli for the computation of the spatiotemporal CSFs were moving sinusoids with 16 spatial frequencies in the range of 0 to 15 cpd, 10 temporal frequencies in the range or 0 to 10 Hz, and 9 contrasts in the range of 0.07 to 0.6. The average color and luminance were the same as in the image case.

The lower limit of the explored contrast range comes from the limitation owing to noise discussed in Methods: Estimating contrast sensitivity in autoencoders. The upper contrast limit and the average luminance were selected to ensure that corrupted signals are reproducible in regular displays (which is the range of the scenes used in the training).

Appendix C: Convergence and Performance of the models

To guarantee that the presented CSF results do not come from eventual training artifacts, this appendix illustrates the proper convergence and proper performance of all the considered CNNs in all the goal/architecture scenarios. For each considered case in the Experiments, we show the learning curves and explicit examples of the responses (reconstructed signals in test). Finally, as an illustrative example, we also show one extra case (for video, experiment 6) where convergence was not complete in one of the networks and the consequences in the reconstructions and in the CSFs.

Experiment 1: Distortion compensation from a range of CNN architectures

Figure 14 shows the learning curves of all the CNN models used in experiment 1. Throughout the Appendix the gray/black curves refer to the ε_{LMS} distortion of the retinal signal in the LMS color space of Stockman and Sharpe (2000), which is constant along the learning. The cyan–blue curves show the evolution of the error in the response of the networks. The light color curve describes the error over image batches in the training phase while the dark color describes the same error in the validation set. The error of the response (solution) significantly drops below the error of the input signal (problem), thus indicating that the network is actually achieving the functional goal it has been designed for. The plateau achieved by the blue curves (not only in training but, more significantly in validation) implies that a steady convergence was achieved and the resulting model is ready to be tested. Consistency between the train and validation sets is apparent from the parallel behavior of the light and dark curves. Performance tables in the main text (Table 1 for experiment 1) and performance in the visual examples shown here (Figure 14 for experiment 1) refer to an independent test set not used in the learning (training/validation) phase. In all CNNs used in experiment 1, the training has been done in a representative set because the errors in the independent test phase (Table 1 and Figure 14) are consistent with the asymptotic behavior of the learning curves shown in Figure 14. Figure 15 shows visual examples of the performance of the linearized versions of the nonlinear models in experiment 1. It is interesting to note that the optimal linear solution (computed from the train set) has worse behavior than the linearized versions of the networks (as also seen in Table 1).

Experiment 2: Architecture trained on a range of distortion levels

Figure 16 shows the learning curves of the model trained in experiment 2 (two layer ReLU) for different levels of retinal degradation (noise/blur). Note that the specific cases where the iteration was stopped owing to the activation of the early stopping criterion (top left and bottom center), the convergence plateau was already reached. Figure 17 shows examples of the performance in test of the model considered in every training scenario considered in experiment 2.

Experiment 3: Chromatic adaptation versus distortion compensation

Figure 18 (top) demonstrates that the model trained for the five computational goals considered in experiment 3 actually achieves the goals and has proper convergence. Figure 18 (bottom) shows visual examples of the performance.

Experiment 4: Robustness under change of signal statistics

Figure 19 (top) demonstrates that the model trained for the five computational goals considered in experiment 4 actually achieves the goals and has proper convergence. Figure 19 (bottom) shows visual examples of the performance.

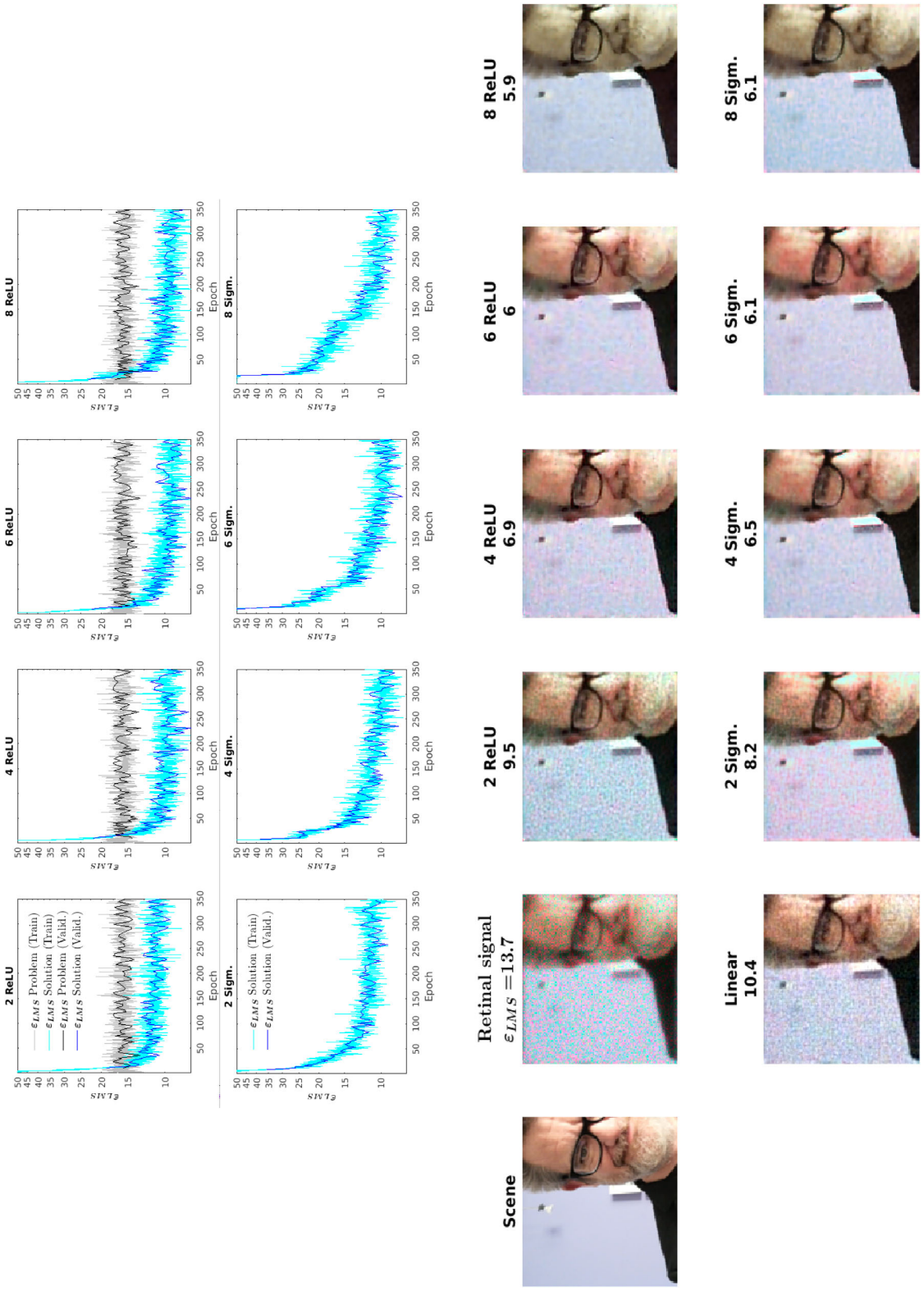


Figure 14. Experiment 1: Convergence and trained models. Learning curves (training/validation) and examples of visual performance (test) of all the models trained in experiment 1. The distortion ε_{LMS}-Problem refers to the original degradation of the images (previous to the application of the net). This distortion describes how difficult the compensation problem is. The distortion ε_{LMS}-solution refers to the degradation remaining in the signal after the application of the net. It describes how close the output is to the ideal result. Performance numbers have been truncated to the significance of the standard deviation in the test set.

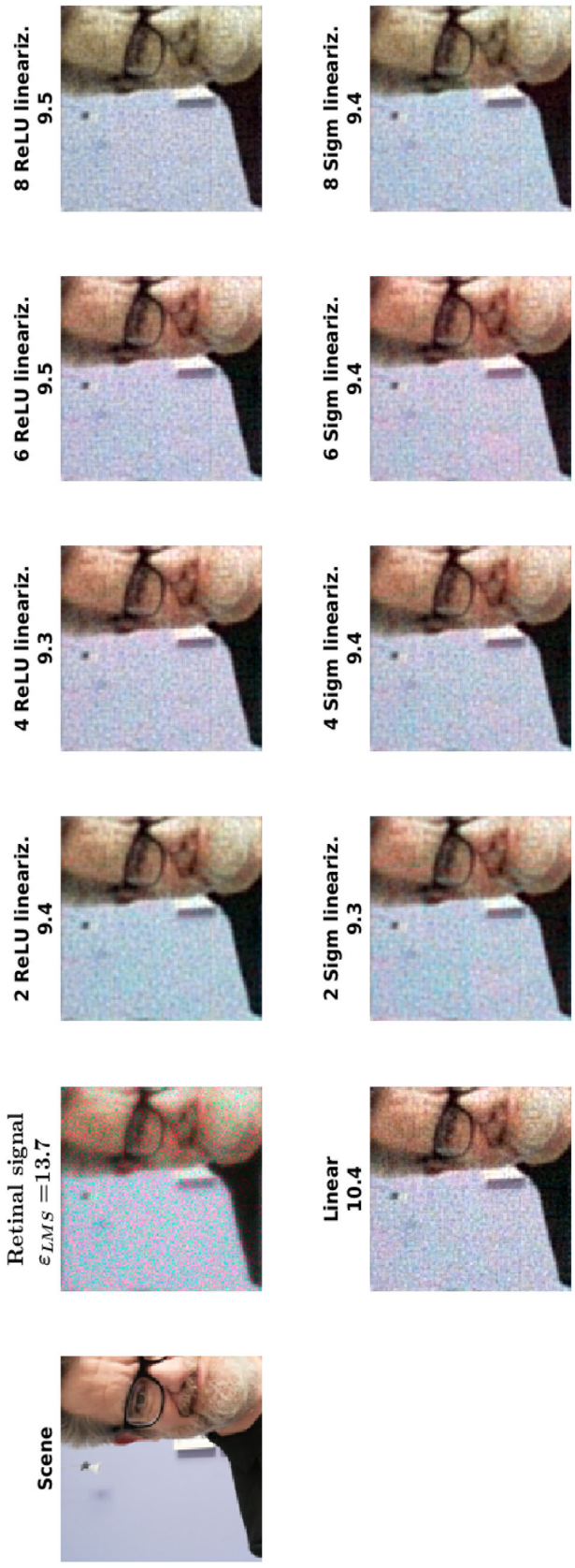


Figure 15. Experiment 1: linearized models. Examples of visual performance (in test) for the linearized CNNs of experiment 1.

Experiment 5: CSFs from bottleneck compensation and biodistortion compensation

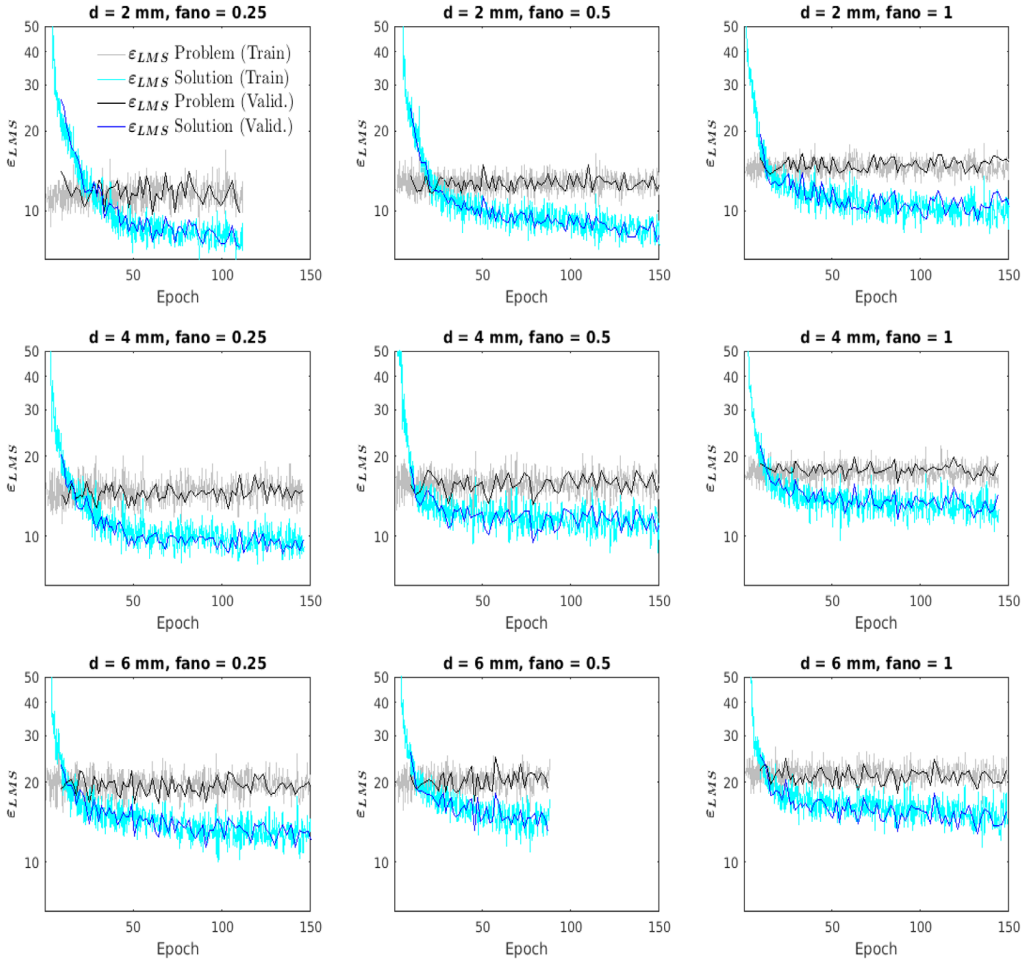
Figure 20 demonstrates that the models trained for the goals considered in experiment 5 have proper convergence and actually achieve the goals within the constraints imposed by the bottlenecks of progressive severity. Figure 21 shows visual examples of the performance.

The reconstruction error ε_{LMS} behaves quite intuitively (see Figure 20 here and Table 3 in the main text): On the one hand, in the cases that involve biodistortion all the architectures do reduce the original value of ε_{LMS} except the architecture G, which has a single feature in its bottleneck. On the other hand, the pure reconstructions cases introduce negligible distortion ε_{LMS} when the inner representation does not restrict spatial resolution nor number of features (the no-bottleneck cases A and B). And, as expected, more severe bottlenecks imply higher ε_{LMS}.

Experiment 6: Spatiotemporal temporal CSFs

The main text includes the CSF results from a range of models trained in Charade (1963). Figure 22 shows the regular evidences on convergence (top) and visual performance in test (bottom) shown for the other Experiments.

In this Appendix, we also include a replication of experiment 6 trained on a movie with higher spatial resolution (*The FBI Story*, 1959). We give the corresponding learning curves (Figure 23) and CSFs (Figure 24). This is interesting for two reasons: 1) it confirms the superiority of shallower nets even for different resolution, and 2) it shows an example of failure in convergence (see that the eight layer model in Figure 23 got stuck in a local minimum (with poor performance) and this has consequences in the complete loss of chromatic information (frame not shown because of copyright issues). Also interesting is the fact that models with four or six layers (which converged as well as the two layer model), substantially over attenuate the red–green channel with the corresponding yellowish–bluish look of the reconstruction and the corresponding impact on the relative scaling of channels in the CSFs, which is not the case for the linear and the 2-layer solutions. This is consistent in the other image/video examples in the main text.



	Degradation ε_{LMS}			2-layer ReLU ε_{LMS}			CSF RMSE		
	F=0.25	F=0.5	F=1	F=0.25	F=0.5	F=1	F=0.25	F=0.5	F=1
d=2	**10.5**±0.1	12.9±0.1	15.8±0.1	7.55±0.07	8.79±0.08	10.4±0.1	29.1	28.1	26.1
d=4	**13.5**±0.2	15.5±0.2	18.1±0.1	8.9±0.1	9.1±0.1	13.1±0.1	24.8	24.5	24.0
d=6	**17.7**±0.2	19.4±0.2	21.6±0.2	11.8±0.2	15.2±0.2	15.1±0.2	22.5	23.5	24.2

Figure 16. Experiment 2: Convergence and performance. Learning curves (training/validation) and numerical performance (in test and in the reproduction of the CSFs) of the CNN model trained in all conditions considered in experiment 2.

Figure 17. Experiment 2: Visual performance. Examples of reconstruction (in test) for the CNN in all the degradation scenarios considered in experiment 2.

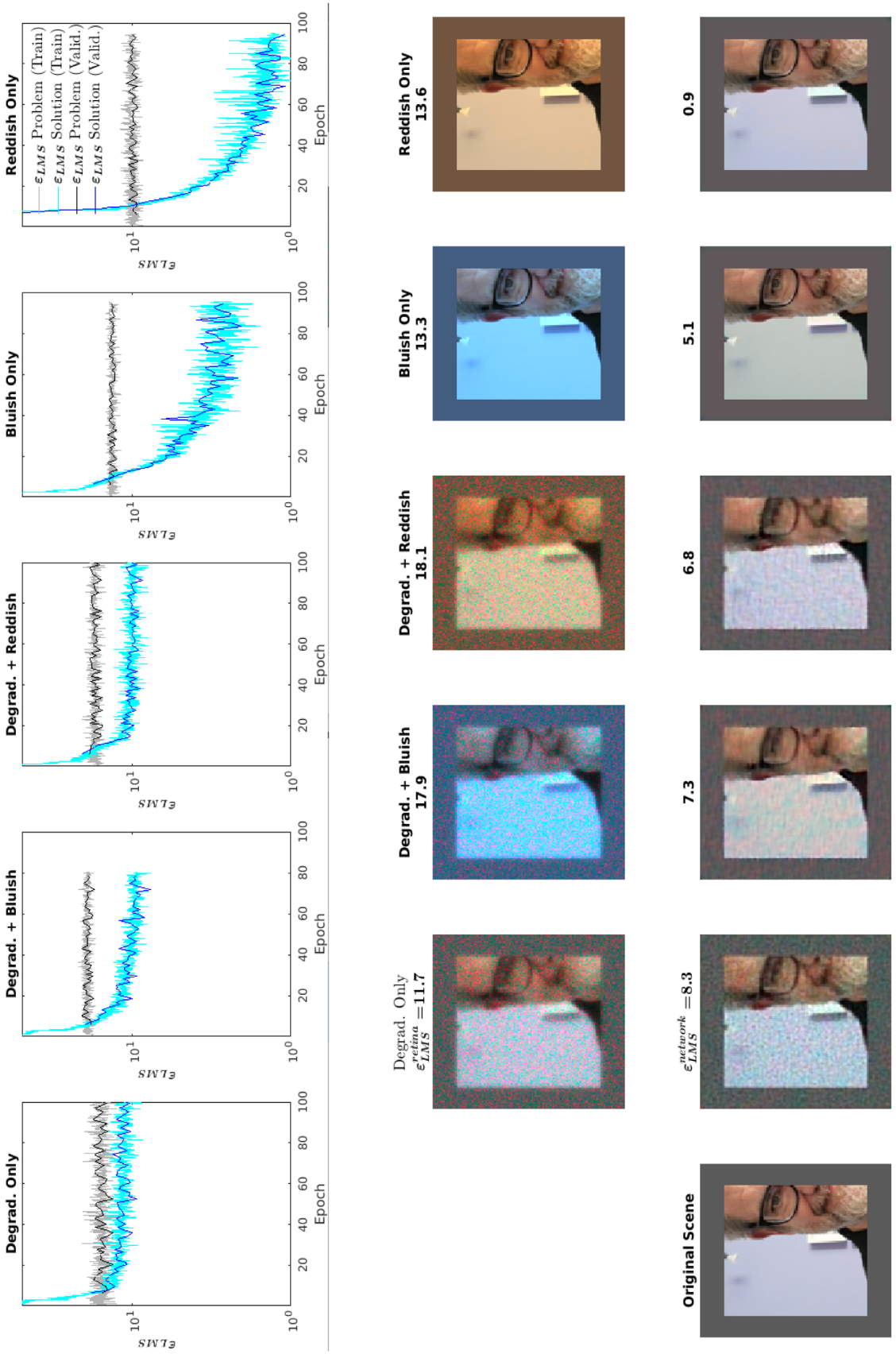


Figure 18. Experiment 3: Convergence and visual performance: Top, learning curves (train/validation) for the considered architecture in the different goals. Bottom: Visual example (test) for the CNNs in experiment 3 (natural images).

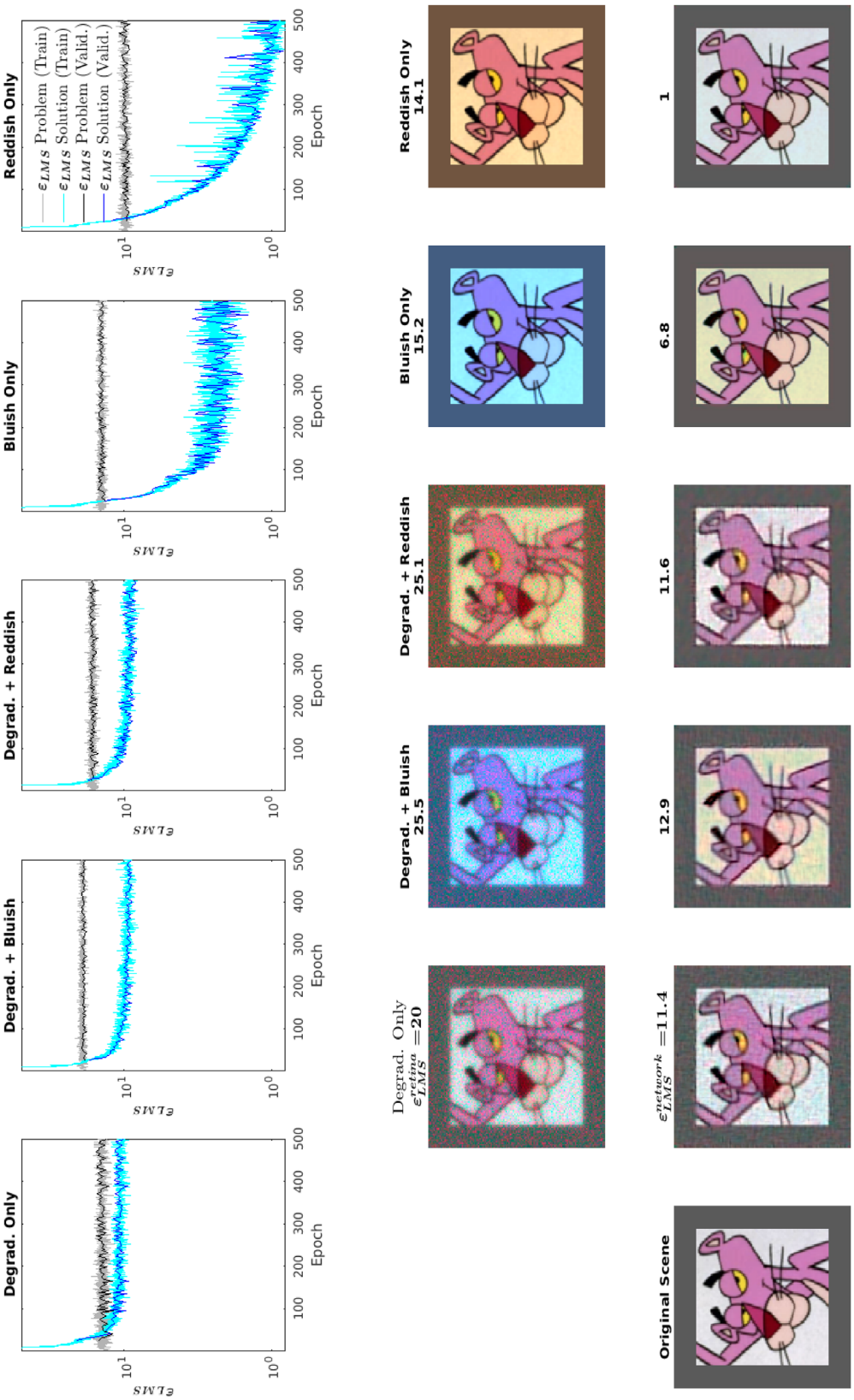


Figure 19. Experiment 4: Convergence and visual performance: Top, learning curves (train/validation) for the considered architecture in the different goals. Bottom: Visual example (test) for the CNNs in experiment 4 (cartoon images). Original image from *The Pink Panther Show* (Freleng, 1963) courtesy of MGM. Similar images (Malo, 2022) lead to equivalent performance in the networks.

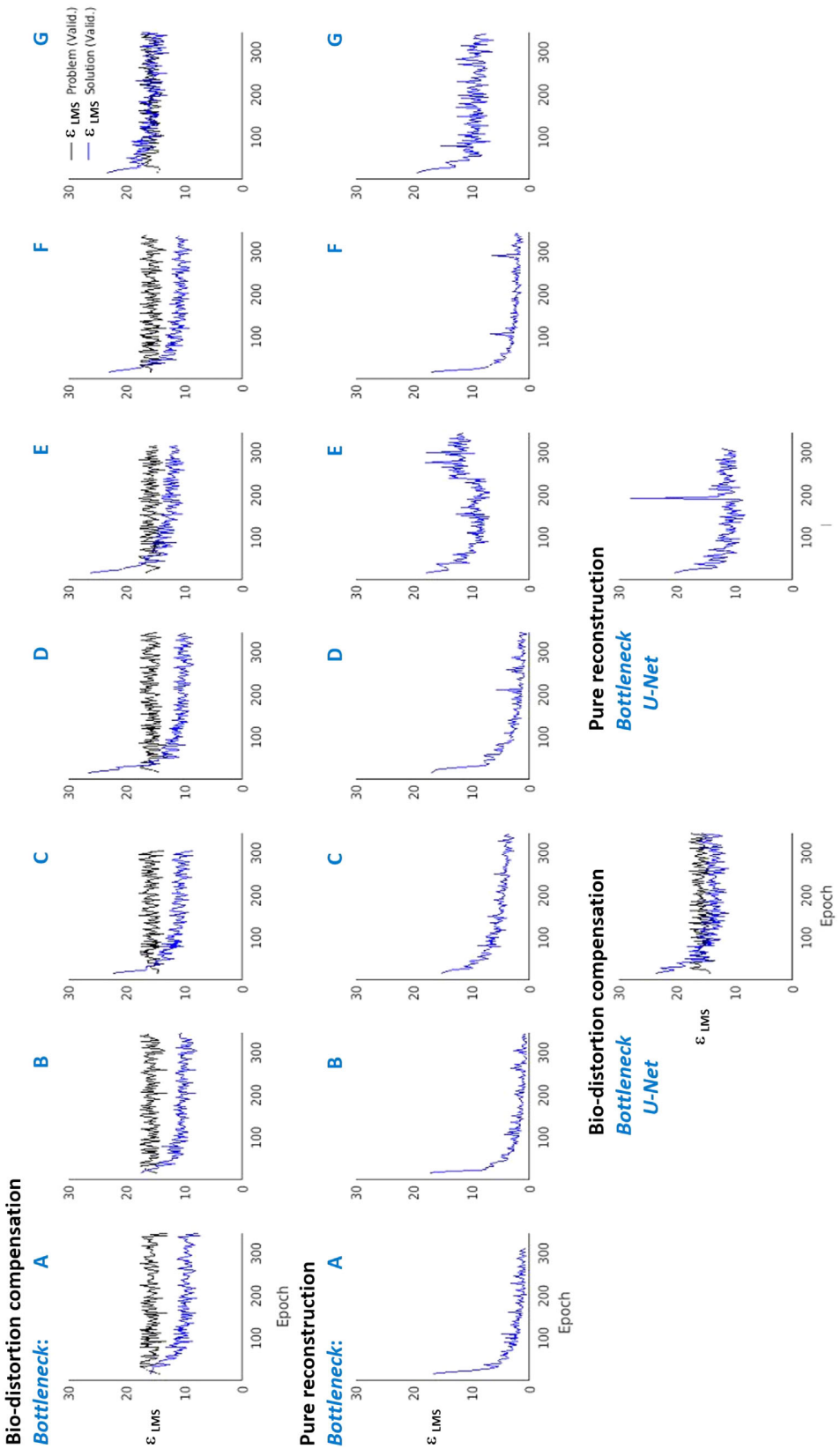


Figure 20. Experiment 5: Bottleneck compensation and biodistortion compensation. Learning curves (train/validation) for the considered architectures in reconstruction with compensation of biodistortion (top row and bottom row, left) and pure reconstruction (middle row and bottom row, right). The cases including biodistortion show the original ε_{LMS} of the problem. See Figure 4 for the structure of the architectures referred by the letters in blue.

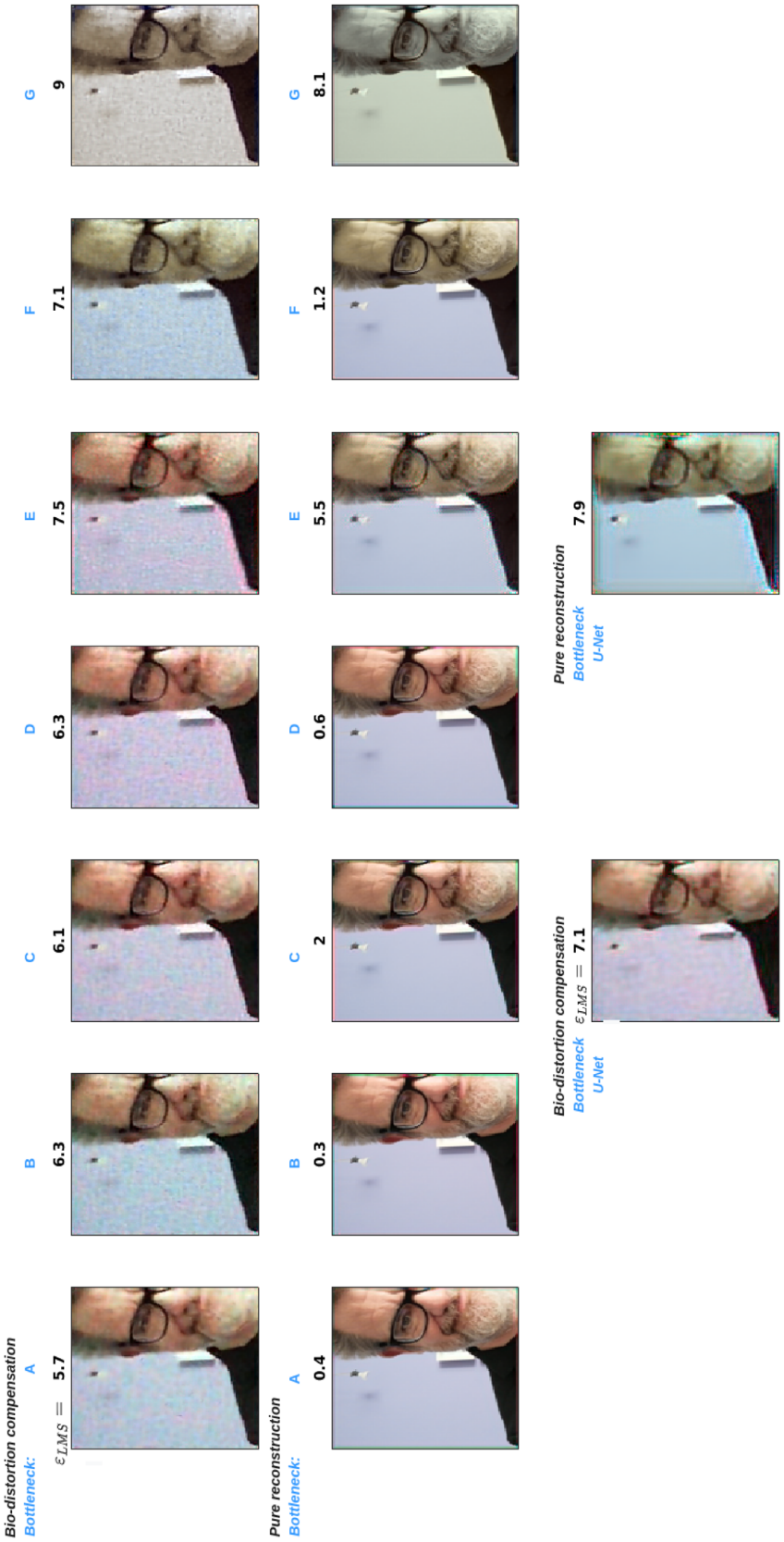


Figure 21. Experiment 5: Bottleneck compensation and biodistortion compensation. Examples of visual performance (in test) for the CNNs of experiment 5. See Figure 4 for the structure of the architectures referred by the letters in blue.

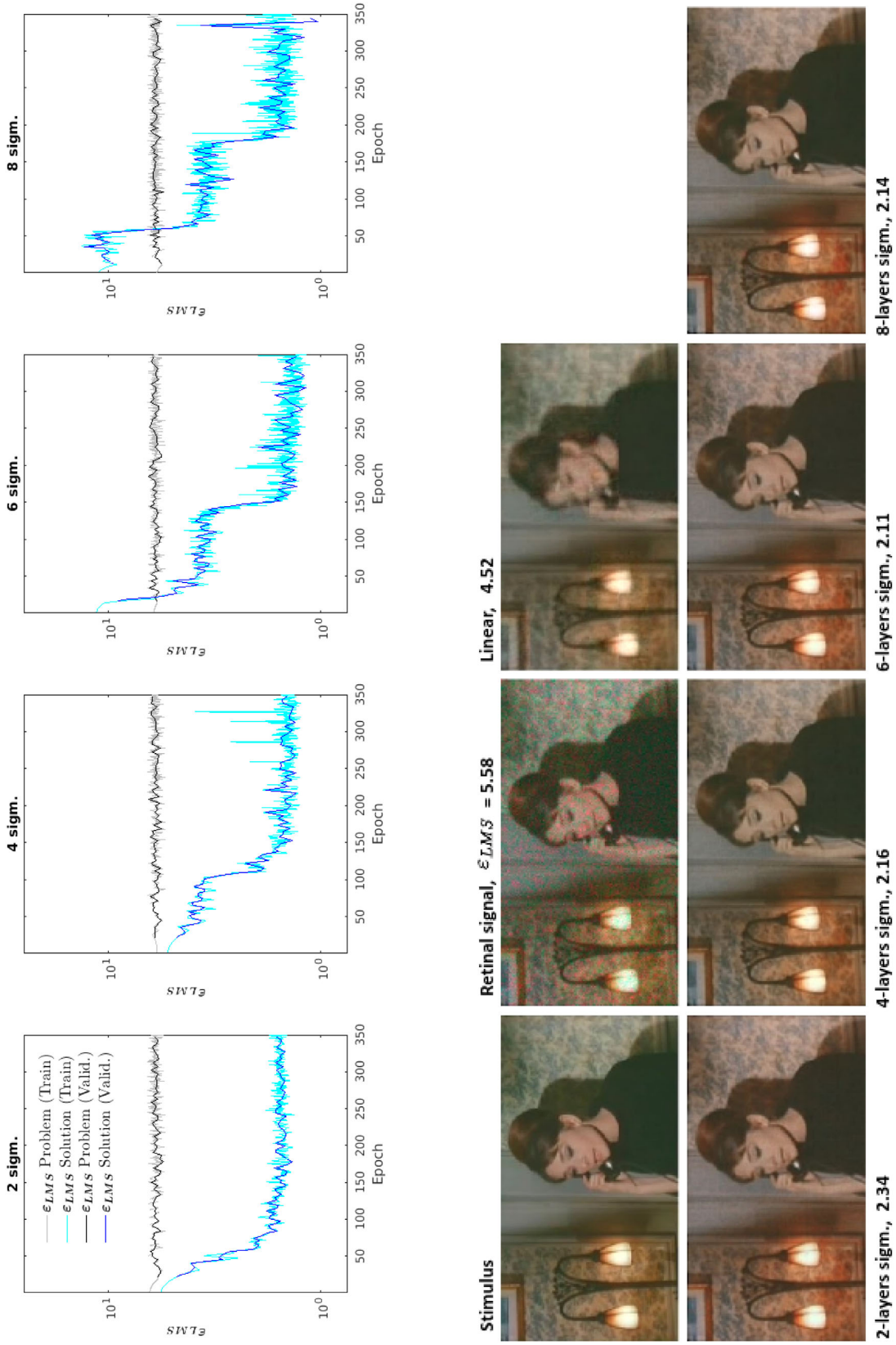


Figure 22. Experiment 6 (spatiotemporal chromatic CSFs in main text): Convergence and visual performance. Top, learning curves for the considered architectures. Visual example (test) for the linear solution and the CNNs (*Charade*, low-resolution movie). The original frame comes from the film *Charade* (Donen, 1963), which is in the public domain.

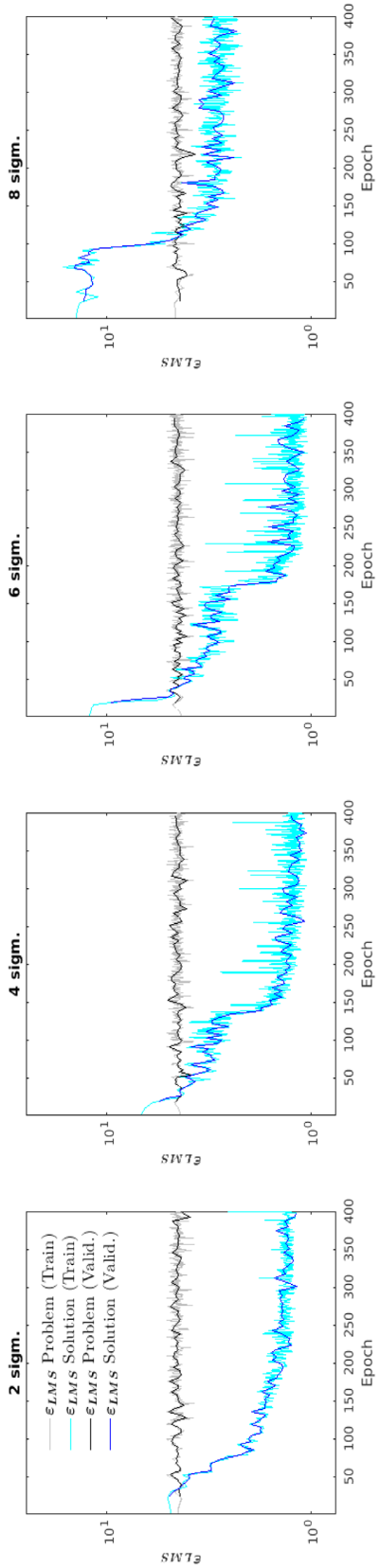


Figure 23. Extended experiment 6: Spatiotemporal chromatic CSFs with different training set (not shown in the main text): Convergence. Figure shows the learning curves for the considered architectures on a higher resolution movie (the film

←

The FBI story; LeRoy, 1959). All models converge except the 8-layer architecture. Visual results are not shown due to copyright issues. Interested readers can obtain these specific results from the authors. The local minimum in which the eight layer architecture was trapped has consequences in the complete loss of chromatic information (frame not shown because of copyright issues). Also interesting is the fact that models with four and six layers (which converged as well as the two layer model), substantially over attenuate the red–green channel with the corresponding yellowish–bluish look of the reconstruction and the corresponding impact on the relative scaling of channels in the CSFs, which is not the case for the linear and the two layer solutions.

→

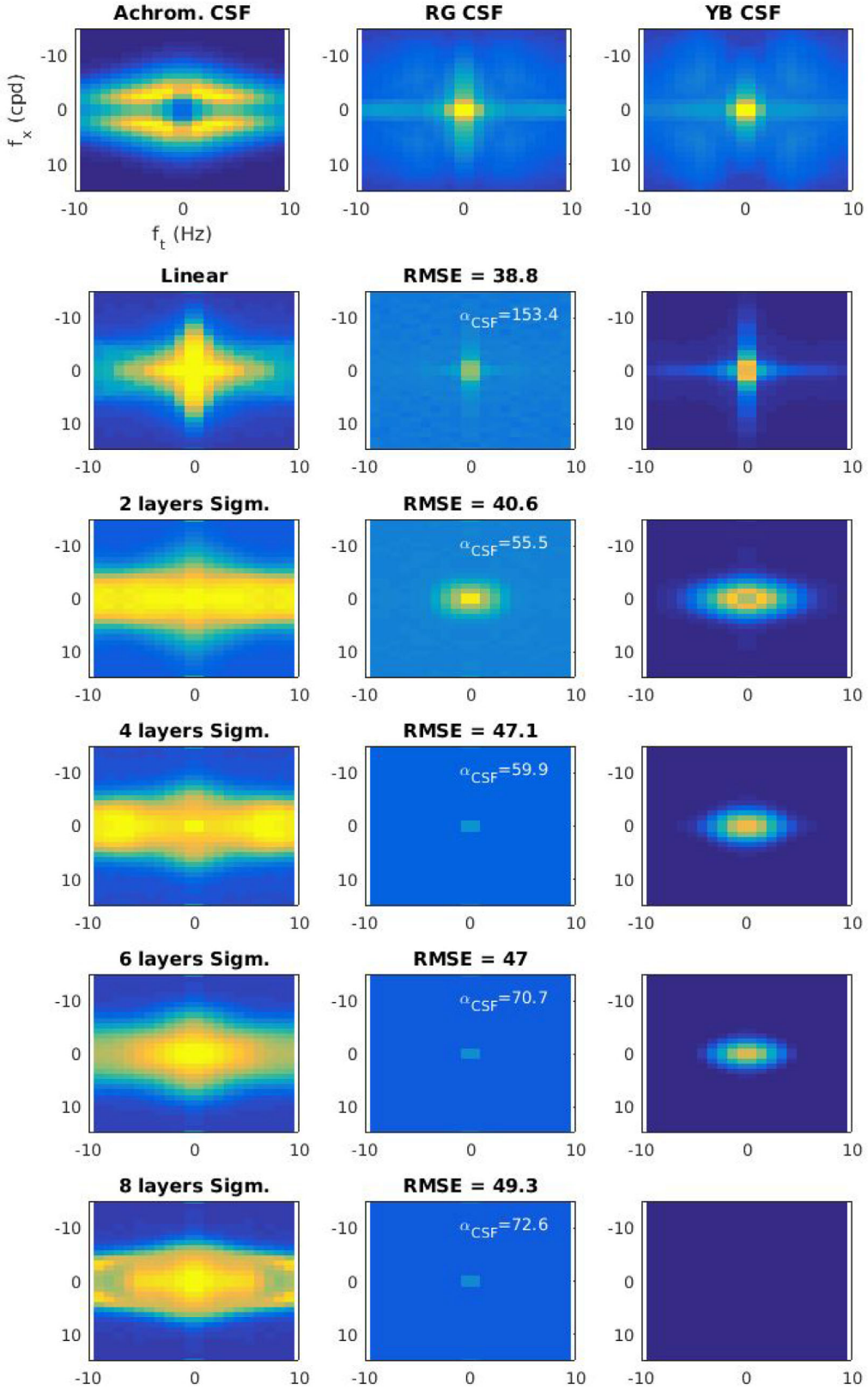


Figure 24. Extended experiment 6 (spatiotemporal chromatic CSFs of a high-resolution movie not shown in the main text). Note that in this case (see Figure 23) all models converged except the eight layer architecture, which totally removes the yellow–blue channel and almost removed the red–green channel.