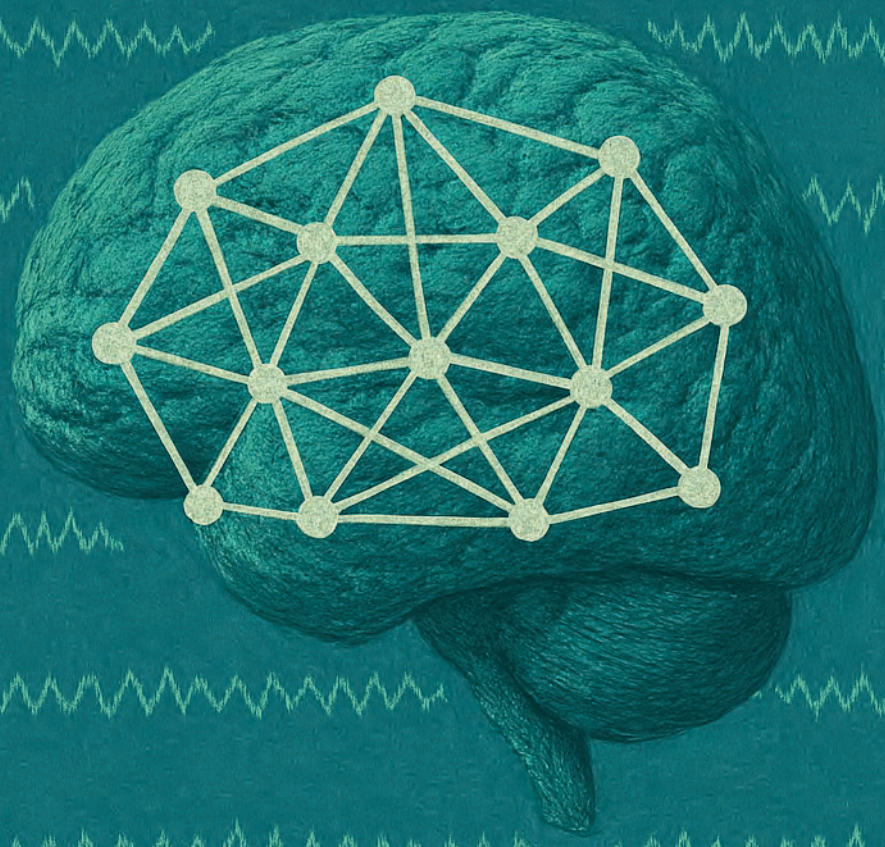
On the predictive and explanative roles of deep neural networks in neuroscience



Amr Fouad Abdelhamid Farahat

DONDERS SERIES

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

On the predictive and explanative roles of deep neural networks in neuroscience

Amr Fouad Abdelhamid Farahat

On the predictive and explanative roles of deep neural networks in neuroscience
Amr Fouad Abdelhamid Farahat

Radboud Dissertation Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Amr Fouad Abdelhamid Farahat
Cover: Amr Fouad Abdelhamid Farahat
Printing: DPN Rikken/Pumbo

ISBN: 9789465151625
DOI: 10.54195/9789465151625
Free download at: https://doi.org/10.54195/9789465151625

RADBOUD
UNIVERSITY
PRESS

On the predictive and explanative roles of deep neural networks in neuroscience

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 22 oktober 2025
om 10.30 uur precies

door

Amr Fouad Abdelhamid Farahat
geboren op 09 juli 1991
te Dakahlyia (Egypte)

Promotoren
Dr. M.A. Vinck
Prof. dr. P.H.E. Tiesinga

Manuscriptcommissie
Prof. dr. M.A.J. van Gerven
Prof. dr. G. Roig (Goethe Universität Frankfurt am Main, Duitsland)
Dr. A. Ingrosso

Table of Contents

Chapter 1
General Introduction

"You should not let your method become your theory."

— Paul Cisek

In 2012, a convolutional deep neural network (DNN) model achieved a breakthrough victory in the ImageNet object recognition competition, reducing the error rate from 26.3% to 15.1% (Krizhevsky et al., 2012). Twelve years later, DNNs have revolutionized many scientific fields, serving as powerful tools for analyzing complex, high-dimensional data (Bianchini et al., 2020; Egger et al., 2021). Their impact is particularly notable in biomedical research, where the inherent complexity of the data makes DNNs exceptionally valuable. The 2024 Nobel Prize in Chemistry, awarded for AlphaFold (a DNN that predicts protein structure from amino acid sequences (Jumper et al., 2021)), exemplifies this impact, with profound implications for drug discovery (Borkakoti & Thornton, 2023; F. Ren et al., 2023). DNNs also predict gene expression from DNA (Avsec et al., 2021) and generate novel proteins with therapeutic potential (Ferruz et al., 2022). In medical imaging, they classify skin lesions (Manole et al., 2024), screen mammograms for breast cancer (McKinney et al., 2020), and detect lung cancer (Gorenstein et al., 2023; Hroub et al., 2024) and brain tumors (Nazir et al., 2021). In neuroscience, DNNs analyze large-scale neural data (Stringer & Pachitariu, 2024) and aid in diagnosing brain disorders (Valliani et al., 2019). The ability of DNNs to handle vast amounts of data from different modalities is particularly well-suited to developing biomarkers for neurodegenerative and psychiatric conditions.

Beyond data analysis, DNNs, partially inspired by the brain (McCulloch & Pitts, 1943; Rosenblatt, 1958; Rumelhart et al., 1986), are increasingly used as computational models of brain information processing (Cichy & Kaiser, 2019; Doerig et al., 2023; Richards et al., 2019). Their performance rivals human capabilities in cognitive tasks like object recognition (He et al., 2016a), speech recognition (Graves et al., 2013), language understanding and generation (Touvron et al., 2023), navigation (Graves et al., 2016), and game playing (Mnih et al., 2015; Silver et al., 2016). They have also proven to be the best current models for predicting brain activity across cognitive domains, including vision (Cichy et al., 2016; Güçlü & Van Gerven, 2015; D. L. K. Yamins et al., 2014), audition (Kell et al., 2018), language (AlKhamissi et al., 2024; Caucheteux & King, 2022), and decision-making/control (Botvinick et al., 2020; Dabney et al., 2020). These capabilities make them attractive candidates for modeling brain function.

However, despite their success, DNNs face an "interpretability crisis" – they are often considered "black boxes" due to their complex, multi-layered architectures with potentially billions of parameters (Xua & Yang, 2024). This opacity raises concerns about their value in providing new insights into neuroscience (Bowers et al., 2023; Chirimuuta, 2021; Kay, 2018). Therefore, to effectively leverage DNNs as computational models, rigorous study design is essential. We must move beyond predictive power and investigate how these models produce behavior (Baker et al., 2018; Brendel & Bethge, 2019; Geirhos et al., 2018, 2019). Furthermore, comparing DNN and brain representations requires careful consideration of the properties of the representations (Biscione et al., 2024; Jacob et al., 2021) and the similarity metrics used (Soni et al., 2024).

Deep Neural Networks

DNNs, a subset of machine learning algorithms, are composed of interconnected "artificial neurons." Each neuron computes a weighted sum of its inputs, applying a non-linear activation function to produce an output. The weights correspond to the strength of the connections between the neurons. the neurons are organized into layers: an input layer, at least one hidden layer, and an output layer. This layered structure enables DNNs to extract hierarchical features from inputs, with complexity increasing at deeper layers. DNNs with at least one hidden layer are universal function approximators, capable of approximating any continuous function given appropriate input-output mappings (Hornik et al., 1989). Two main types of DNNs exist: feedforward networks, where neurons receive inputs only from the preceding layer, and recurrent networks, where neurons possess hidden states updated over time through recurrent connections (Goodfellow et al., 2016). A key advantage of DNNs is their architectural flexibility, allowing modification of learned inductive biases by manipulating neuron connectivity. A particularly successful example is the Convolutional Neural Network (CNN) (LeCun et al., 1998, 2015).

The architectural bias of CNNs (Fig. 1.1) contributes significantly to their success with spatially structured data (e.g., images). Each convolutional layer neuron connects only to a small input region (its "receptive field"), enabling efficient detection of local patterns – a design mimicking the visual cortex (Hubel, Wiesel, et al., 1959). Weight sharing across spatial locations creates a feature map and provides translation equivariance: translating the input correspondingly translates the output. Combined with pooling operations (which downsample feature maps by selecting the maximum or average value within a local window, inspired by complex cells in the visual cortex (Fukushima et al., 1983)), CNNs achieve transla-

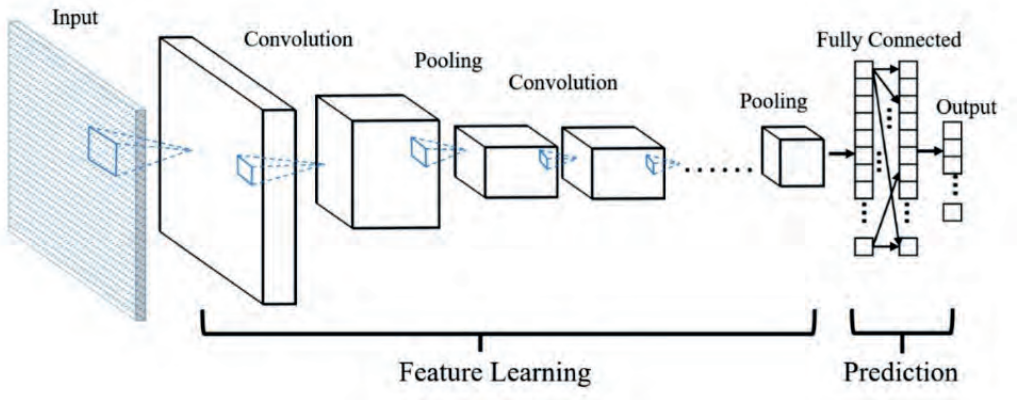


Figure 1.1. CNN architecture. A CNN typically comprises a series of convolutional layers, combined with pooling operations and non-linear activation functions. Adapted from (González-Rodríguez & Plasencia-Salgueiro, 2021).

tion invariance: the output is invariant to a feature's location. Parameter sharing also improves efficiency, reducing overfitting risk and promoting generalization.

CNNs excel at processing structured data (images, time-series) due to these inductive biases (local connectivity, translation invariance, hierarchical feature learning), enabling them to capture spatial or temporal relationships effectively. This leads to their success in tasks like object detection (S. Ren et al., 2015), image classification (He et al., 2016b; Krizhevsky et al., 2012), semantic segmentation (Guo et al., 2018), and time-series forecasting (Bai et al., 2018; Hewage et al., 2020).

While standard discriminative CNNs, optimized for classification or regression, learn mappings from input space to output labels, generative models like autoencoders learn compressed, latent representations of input data (Ballard, 1987; Kingma & Welling, 2013). An encoder maps the input to a lower-dimensional latent space, while a decoder reconstructs the original input from the latent representation. This makes autoencoders valuable for unsupervised learning, including anomaly detection (An & Cho, 2015; C. Zhou & Paffenroth, 2017) and dimensionality reduction (W. Wang et al., 2014). The learned latent representation captures the underlying data structure and salient features without explicit labels.

Prediction and Explanation: A Dual Role for DNNs in Neuroscience

As noted, DNNs serve both as tools for analyzing neural data and as computational models of brain function. It is crucial, however, to distinguish between these roles. A trade-off exists between a model's predictive power and the scientific understanding it provides (Chirimuuta, 2021). Highly accurate, complex, non-linear DNNs may lack the explanatory power of simpler, interpretable models (e.g., linear models). With millions or billions of interconnected parameters, understanding how a DNN arrives at its predictions is challenging, leading to the "black box" label. While excelling at the "what" question (prediction), they struggle with "why" and "how" questions (explanation). Neuroscientists thus seemingly face a choice between models that predict well and those that explain well. However, some researchers challenge this strict trade-off, distinguishing between understanding the model and using the model to understand a phenomenon (Kästner & Crook, 2023; Lawler & Sullivan, 2021). Even a complex, unintelligible model can provide valuable insights. For instance, post-hoc techniques like feature visualization (Olah et al., 2017) or saliency methods (Simonyan et al., 2013; Sundararajan et al., 2017; Zeiler & Fergus, 2014) can be employed to explain the models. However, these methods may not offer more causal understanding than simpler associative approaches (e.g., visualizing maximally activating examples) (Borowski et al., 2021; Zimmermann et al., 2021), and they can be unreliable (Adebayo et al., 2018; Kindermans et al., 2019; Rudin, 2019).

In some neuroscience applications, predictive power may outweigh explanation (Boon & Knuuttila, 2009). DNNs have improved decoding accuracy in brain-computer interfaces (BCIs) (Farahat et al., 2019; Lawhern et al., 2018). Recently, an RNN decoded speech from invasively recorded neural activity, enabling an ALS patient to communicate at 62 words/minute, approaching natural conversation speed (160 words/minute) (Willett et al., 2023). A similar algorithm allowed a tetraplegic patient to control three finger groups for reaching and holding targets (Willsey et al., 2025). DeepLabCut, a DNN-based method, accurately estimates animal poses, quantifying behavior in neuroscience studies (Lauer et al., 2022; Mathis et al., 2018). DNNs are also applied to neuroimaging data to discover diagnostic biomarkers for neurological and psychiatric disorders (Calhoun et al., 2021).

However, when used as computational models of brain function, DNNs are expected to provide explanatory value beyond data fitting. A critical question is whether they offer this explanatory value despite being intricate black boxes

(Chirimuuta, 2021; Cichy & Kaiser, 2019; Kästner & Crook, 2023; Kay, 2018). Simple mathematical models facilitate understanding by explicitly defining variables and interactions. This is infeasible in DNNs with millions of non-linearly interacting parameters. However, high-level, abstract parameters (architecture, objective function, training dataset, learning rules) can describe DNN models (Cichy & Kaiser, 2019; Richards et al., 2019). Although post-hoc explanation methods are still developing, they may improve with better theoretical understanding of DNN learning and generalization. It is also important to recognize that simple, interpretable models may be insufficient to capture the complexity of the brain and its supported behaviors (Wichmann & Geirhos, 2023). Therefore, models of the brain are unlikely to fulfill all desired criteria, often involving trade-offs between realism, precision, and generality (Levins, 1966; Matthewson, 2011). Modelers will often have make strategic choices, prioritizing certain properties based on research goals. Therefore, a pluralistic approach, developing multiple models with different assumptions, is also beneficial. Model assessment should be multidimensional, going beyond a single accuracy metric (e.g. neural prediciton or classification accuracy) and developing multiple metrics to address the model's strengths and limitations (Wichmann & Geirhos, 2023).

CNNs as Models of the Primate Visual System

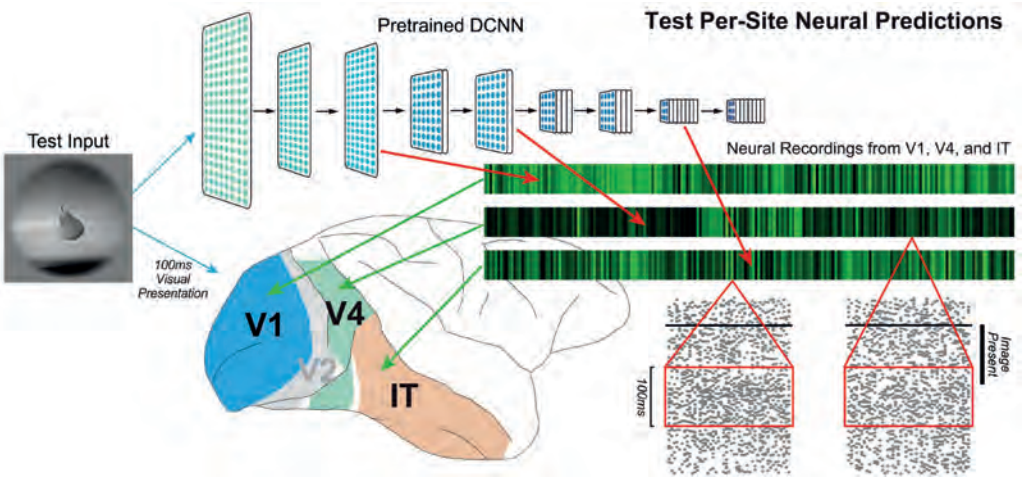


Figure 1.2. Correspondence of CNN layers and brain areas in the ventral stream of the macaque visual system. Adapted from (Zhuang et al., 2021).

Due to their biological inspiration, human or superhuman performance on vision tasks, and superior predictive power, CNNs are strong candidates for computational models of the primate visual system (Kriegeskorte, 2015; Lindsay, 2021). The

stacking of convolutional layers, non-linearities, and pooling operations mirrors the feedforward path in the primate ventral visual stream: V1 $\rightarrow$ V2 $\rightarrow$ V4 $\rightarrow$ IT (Fig. 1.2). Deeper network layers and higher brain areas have larger receptive fields and represent more complex features. When CNNs optimized for object recognition were presented with the same stimuli as macaques or humans, their layer representations successfully predicted activity in ventral stream areas (Cadena et al., 2019; Güçlü & Van Gerven, 2015; Seeliger et al., 2018; D. L. K. Yamins et al., 2014). Crucially, the hierarchy of layers best predicting different brain areas mirrored the ventral stream's hierarchy: early/intermediate layers best predicted V1; deeper layers best predicted IT. These studies used linear regression to predict brain activity as a linear combination of CNN representations, using correlation or explained variance to measure CNN-brain similarity. Representational similarity analysis (RSA) yielded similar conclusions (Khaligh-Razavi & Kriegeskorte, 2014). RSA assesses representational similarity by creating a representational dissimilarity matrix (RDM) for each source (CNN layer or brain recording), capturing how dissimilar different stimuli are in the population space (Kriegeskorte et al., 2008). RDMs are then compared using correlational measures (e.g., Kendall correlation). Because RDM dimensions are independent of the representational space's dimensionality, RSA facilitates comparing different models or recording modalities. While different similarity metrics may yield consistent hierarchical mappings, recent work highlights discrepancies in conclusions drawn from different metrics (Kornblith et al., 2019; Soni et al., 2024). Metrics employing linear regression, for example, can be influenced by its inductive biases, such as predictors dimensionality or the ratio between dependent variable dimensionality and sample size of the test set (Canatar et al., 2024; Elmoznino & Bonner, 2024; Schaeffer et al., 2024). These findings underscore the need for careful consideration of similarity metrics.

Beyond predicting neural activity, it is crucial to compare the behavioral responses of CNNs and biological brains. Although CNNs could predict object-level image classification behavior in primates, they did not account for image-level behavior within object recognition tasks (Geirhos et al., 2020; Rajalingham et al., 2018). Unlike human object recognition, which is robust to orientation changes (Biederman, 1987), DNNs exhibit substantial performance drops when classifying objects in unusual poses (Abbas & Deny, 2023; Alcorn et al., 2019; Dong et al., 2022). Compared to humans in recognizing challenging images (e.g., noise-distorted), DNNs underperformed in accuracy and error consistency (Geirhos et al., 2018, 2021). DNNs are also susceptible to adversarial attacks: small, human-imperceptible image perturbations can cause misclassification (Szegedy, 2013). These findings relate to observations that DNNs rely more on object surface characteristics, failing to recognize objects based solely on global shape (e.g., silhouettes) (Baker

& Elder, 2022; Baker et al., 2018, 2020), unlike humans (Baker & Kellman, 2018; Biederman, 1987; Biederman & Ju, 1988; Landau et al., 1988). On cue-conflict datasets (images manipulated via style transfer (Gatys et al., 2016) so an object carries another object's texture), humans reliably classified based on shape; DNNs relied more on texture (Geirhos et al., 2019). These findings challenge DNNs' capacity as computational models of the primate visual system (Bowers et al., 2023; Wichmann & Geirhos, 2023) and urge a multidimensional assessment approach, rather than relying solely on one-dimensional predictive benchmarks (Biscione et al., 2024; Jacob et al., 2021).

DNNs for Diagnosing Brain Disorders

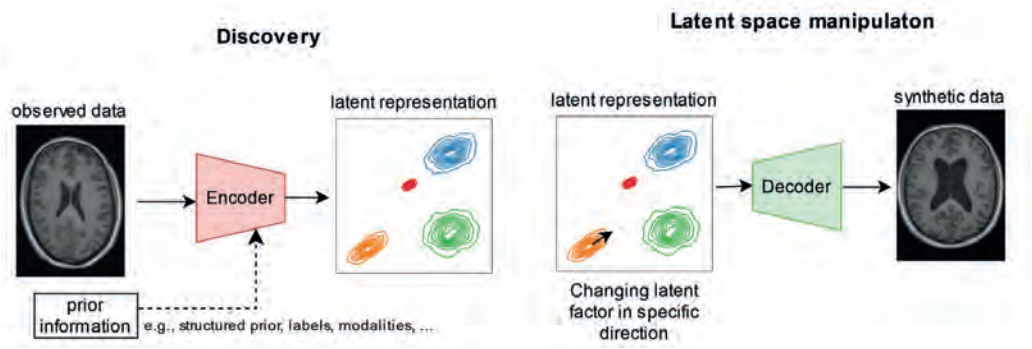


Figure 1.3. Representation learning in generative modeling. Adapted from (Seiler & Ritter, 2024).

Brain disorders (neurological and psychiatric, e.g., epilepsy, Alzheimer's, depression) accounted for over 15% of the global health burden in 2021, surpassing cancer and cardiovascular diseases (Lei & Gillespie, 2024), and are projected to increase by 22% by 2050, affecting over 4.9 billion people. In Europe, one-third of the population suffered from at least one brain disorder in 2010, costing nearly 800 billion euros annually (DiLuca & Olesen, 2014).

Early diagnosis of brain disorders is paramount for improving patient outcomes. Timely detection facilitates prompt intervention, potentially slowing progression, managing symptoms, and preventing complications. For example, early Alzheimer's diagnosis allows for lifestyle adjustments that may mitigate cognitive decline. In epilepsy, early diagnosis and treatment can prevent seizures, avoiding injuries or fatalities. Early detection can also reduce medical costs by slowing disease progression and reducing the risk of disabilities. However, early diagnosis is challenging. Early brain changes can begin years before symptomatic manifestation and can be subtle, making them difficult to detect with non-invasive techniques

like magnetic resonance imaging (MRI), computed Tomography (CT) and scalp electroencephalography (EEG). Invasive diagnostics (brain biopsy, intracranial EEG) carry risks. Symptom overlap between disorders also hinders early diagnosis. For example, 24% of Parkinson's disease (PD) diagnoses are incorrect, often overlapping with progressive supranuclear palsy (PSP), multiple system atrophy (MSA), and Alzheimer's disease (Hughes et al., 1992). American Academy of Neurology (AAN) guidelines recommend neuroimaging techniques like MRI and single-photon emission computed tomography (SPECT) to differentiate between PD, essential tremor (ET), and MSA, but this requires expert supervision (Pahwa & Lyons, 2010; Suchowersky et al., 2006). Therefore, developing multiple biomarkers for diagnosis and follow-up is an active research area (Hansson, 2021).

DNNs are poised to significantly impact early brain disorder diagnosis. Their ability to detect subtle patterns in large, complex datasets offers opportunities for timely and accurate prediction of disease-related brain changes. Various neuroimaging modalities (structural: sMRI, Diffusion Tensor Imaging (DTI); functional: fMRI, EEG, positron emission tomography (PET)) provide essential information for physicians for identifying and distinguishing disorders (Shoeibi et al., 2022, 2023). DNNs' ability to automatically learn features from raw data at different levels of abstraction makes them more suitable than conventional machine learning for fusing multi-modal data (Acosta et al., 2022; Gao et al., 2020; Stahlschmidt et al., 2022).

However, protecting patient data privacy is a top priority in healthcare limiting the availability of disease data. Even with consent, medical datasets are often imbalanced (Johnson & Khoshgoftaar, 2019), biasing learning in discriminative DNNs. Although this poses a challenge to unlocking the full potential of clinical applications of DNNs, it offers an opportunity for unsupervised training methods using only voluntarily collected data from healthy subjects. Besides predictive and discriminative tasks, DNNs excel at learning data distributions for generative modeling (Goodfellow et al., 2014; Makhzani et al., 2015). DNNs can be trained to generate synthetic data retaining the statistics of real-world healthy or disease data (Seiler & Ritter, 2024; R. Wang et al., 2023), addressing data scarcity and imbalance. Pinaya et al., 2022 used latent diffusion models (Rombach et al., 2021) to generate 100,000 high-fidelity 3D T1w MRI brain images, conditioned on covariates like age and sex. Generative adversarial networks (GANs) (Goodfellow et al., 2014) used to synthesize training data improved CNN tumor detection performance (Han et al., 2019). Lin et al., 2021 used reversible GANs (van der Ouderaa & Worrall, 2019) to translate the cheap MRI images to the more expensive PET images, using both to diagnose AD. Using synthetic PET images improved AD diagnosis over MRI images alone, matching performance with MRI and real PET

images.

Besides generating synthetic data, generative models learn low-dimensional latent representations summarizing key factors underlying data structure (Higgins et al., 2017) (Fig. 1.3). Applying these techniques to neuroimaging data could uncover disease subtypes when applied to cross-sectional data (Yang et al., 2021) or different stages when applied to longitudinal data (Couronné et al., 2021). Moreover, generative networks that accurately model brain scans cannot account for anomalous sample variability in their latent space (Schlegl et al., 2017). Thus, they can be used for unsupervised screening for deviations from the normative distribution (e.g., lesions) (Bengs et al., 2021; Nguyen et al., 2021). Thus, both discriminative and generative DNNs applied to neuroimaging data have the potential to transform early diagnostics of challenging brain disorders.

Thesis Outline

This thesis further examines the trade-off between prediction and explanation when employing complex, opaque models like DNNs in various neuroscience applications, both as scientific tools and as models of natural behavior and biological computations.

In **Chapter 2**, I develop a model explanation technique to uncover the extent to which object recognition CNNs can learn spatial relationships between features. Learning spatial relations is crucial for CNNs to develop shape representations. The nature of object representations learned by CNNs is central to discussions of CNNs as models of human object recognition and of neural activity prediction in the primate ventral stream.

In **Chapter 3**, I systematically manipulate the architecture and training of simple CNN models to predict neural activity in early and late visual areas in humans and non-human primates. Using control experiments and a multidimensional model assessment, I gain insights into what enables CNNs to successfully model different stages of the visual hierarchy, without necessarily explaining the models themselves.

In **Chapter 4**, I develop a deep anomaly detection framework for early diagnosis of epileptogenesis, the process of a healthy brain transforming into an epileptic one after injury. The framework serves as a proof-of-concept screening tool, scanning EEG signals for anomalous segments and integrating evidence over time to improve sensitivity. Importantly, the framework is designed and validated for clinical routines. It employs an opaque, non-interpretable DNN generative model. In **Chapter 5**, I discuss why a lack of interpretability should not preclude adopting similar systems in neurological applications.

In **Chapter 5**, I summarize my findings and discuss them in the context of the prediction-explanation trade-off in complex, opaque models like DNNs.

Chapter 2

A novel feature-scrambling approach reveals the capacity of convolutional neural networks to learn spatial relations

Published as: **Farahat, Amr**, Felix Effenberger, and Martin Vinck. "A novel feature-scrambling approach reveals the capacity of convolutional neural networks to learn spatial relations." Neural networks 167 (2023): 400-414.

Abstract

Convolutional neural networks (CNNs) are one of the most successful computer vision systems to solve object recognition. Furthermore, CNNs have major applications in understanding the nature of visual representations in the human brain. Yet it remains poorly understood how CNNs actually make their decisions, what the nature of their internal representations is, and how their recognition strategies differ from humans. Specifically, there is a major debate about the question of whether CNNs primarily rely on surface regularities of objects, or whether they are capable of exploiting the spatial arrangement of features, similar to humans. Here, we develop a novel feature-scrambling approach to explicitly test whether CNNs use the spatial arrangement of features (i.e. object parts) to classify objects. We combine this approach with a systematic manipulation of effective receptive field sizes of CNNs as well as minimal recognizable configurations (MIRCs) analysis. In contrast to much previous literature, we provide evidence that CNNs are in fact capable of using relatively long-range spatial relationships for object classification. Moreover, the extent to which CNNs use spatial relationships depends heavily on the dataset, e.g. texture vs. sketch. In fact, CNNs even use different strategies for different classes within heterogeneous datasets (ImageNet), suggesting CNNs have a continuous spectrum of classification strategies. Finally, we show that CNNs learn the spatial arrangement of features only up to an intermediate level of granularity, which suggests that intermediate rather than global shape features provide the optimal trade-off between sensitivity and specificity in object classification. These results provide novel insights into the nature of CNN representations and the extent to which they rely on the spatial arrangement of features for object classification.

Introduction

The development of Convolutional Neural Networks (CNNs) has led to a revolution in the field of computer vision (Krizhevsky et al., 2012; LeCun et al., 2015). Machine vision using CNNs has been able to rival human performance in object recognition tasks on large-scale datasets such as ImageNet (He et al., 2016a). Moreover, a series of recent works have shown that CNN activations can be used to predict neural activity in the ventral stream of the primate visual system known to be responsible for object recognition (Cadieu et al., 2014; D. L. K. Yamins et al., 2014; D. L. Yamins & DiCarlo, 2016). Therefore, there has been a growing interest in developing behavioral benchmarks that evaluate similarities and differences between CNN models and human vision (Geirhos et al., 2018, 2021; Rajalingham et al., 2018). Crucial to the behavior of these artificial and biological vision systems is their internal representation of objects. The ability of humans to recognize objects based on their abstract shapes (Baker & Kellman, 2018; Biederman & Ju, 1988; Landau et al., 1988) suggests that the internal representations of objects in the brain must reflect the global structure of objects (Barenholtz & Tarr, 2006; Biederman, 1987). An abstract representation of the global shape of an object requires the encoding of the spatial relations between the set of its local features or parts (Barenholtz & Tarr, 2006; Biederman, 1987). Accordingly, in order to understand the biases that govern the strategies of CNNs performing object recognition, it is central to determine the spatial extent of the diagnostic features CNNs use for object recognition. Moreover, it is equally important to investigate the role that spatial relations play in the construction of these diagnostic features.

Recent studies have shown inconsistent conclusions regarding the reliance of CNNs trained for object recognition on sets of local features or a global representation of objects (Baker & Elder, 2022; Baker et al., 2018, 2020; Brendel & Bethge, 2019; Geirhos et al., 2019; Jo & Bengio, 2017; Kubilius et al., 2016; Ritter et al., 2017; Tartaglini et al., 2022). Some studies have shown that CNNs trained for object recognition are biased towards surface statistical regularities (*texture*) (Baker & Elder, 2022; Baker et al., 2018, 2020; Geirhos et al., 2019; Jo & Bengio, 2017). In these studies, CNNs were tested on image datasets that included, for example, low-frequency filtered images (Jo & Bengio, 2017), shape-texture cue conflict stimuli using style transfer (Gatys et al., 2016; Geirhos et al., 2019), deformed silhouettes and other abstract shape images (Baker & Elder, 2022; Baker et al., 2018) and simple geometric shapes (Baker et al., 2020). However, other studies reached different conclusions using other image manipulations or different evaluation methods (Kubilius et al., 2016; Ritter et al., 2017; Tartaglini et al., 2022). We reckoned that these different conclusions may be due to the hypothesis-driven

approach resulting from the choice of the nature of the stimulus datasets and the object classes represented in them. For this reason, we developed a framework for training and testing CNNs that enables us to inspect the shape representations of CNNs by separately controlling the granularity of CNN features (local vs. global) and the spatial relations between them. This approach allows us to take on the question of to what extent the CNN architecture constrains their capacity to learn shape representations and whether CNNs use the spatial relations among features for object recognition.

Previous work has shown that grid-based image scrambling can be used to identify brain areas sensitive to global configurations of objects (Grill-spector et al., 1998), expressing characteristic decreases in neural activity with the degree of image scrambling (Grill-spector et al., 1998; Rainer et al., 2002; Vogels, 1999). Image scrambling, however, disrupts not only the spatial relations between object parts but also the shape of the parts themselves (Margalit et al., 2017). To disentangle these two effects, we developed a feature-scrambling approach that allows us to spatially scramble the pretrained features of CNNs with restricted effective receptive fields (ERFs) (Brendel & Bethge, 2019) without introducing the confounding factors of an image-based scrambling approach. The ERF of a CNN is defined as the set of all pixels that can influence the activity of a unit in its last convolutional layer (Le & Borji, 2017). These features represent diagnostic parts of the objects at the ERF level of granularity. After that, we feed these scrambled features to a follow-up CNN that spatially integrates these features and is trained to recognize the class of objects. Recent work suggests that CNNs with restricted ERF sizes can achieve a performance similar to regular CNNs on ImageNet (Brendel & Bethge, 2019). However, it remains unclear whether these models use the same strategies as regular CNNs to solve the task. Notably, the approximation of regular CNNs performance on ImageNet with CNNs with restricted ERFs implies that CNNs rely on a classification strategy that pools local evidence from separate locations in the image without learning the spatial relations between them. This observation would predict, for instance, that training a follow-up CNN on the pretrained features of a CNN with restricted ERFs should minimally affect performance. It would also predict that spatially scrambling the pretrained input features to the follow-up CNN would not lead to a significant difference in performance to training with the right spatial arrangement of the features. In this work, we tested these predictions on different datasets that comprise texture-rich and texture-less images to examine whether CNNs employ different classification strategies for different datasets. Furthermore, we examined to what extent CNNs with smaller ERFs develop representations similar to CNNs with larger ERFs. Finally, we performed a minimal recognizable configuration (MIRC) analysis (Ullman et al.,

Blocks	Residual Units	Feature Maps	Stride	Filter Sizes				
				ERF11	ERF23	ERF47	ERF95	ERF227
Block 1	2	128	2	3,3	3,5	3,5	3,5	5,5
Block 2	3	256	2	1,1,1	3,1,1	3,3,5	3,3,5	5,5,5
Block 3	3	512	2	1,1,1	1,1,1	1,1,1	3,3,3	5,5,5
Block 4	2	1024	1	1,1	1,1	1,1	1,1	5,5

Table 2.1: Architecture details for our ResNets of different ERFs.

2016) to quantify the minimal image patch sizes required by CNNs to achieve correct classification.

Methods

Datasets

We trained CNNs on three datasets with different feature characteristics: the Sketchy, Animals, and ImageNet datasets. The Sketchy dataset contains 75,471 human-drawn sketches spanning 125 classes (Sangkloy et al., 2016). Each sketch is a textureless, black-and-white bitmap graphic that only contains information about the contours of objects without any surface proprieties, and sketches have a high degree of intra-class variability (Fig. 2.1c). The Animals dataset consists of 37,322 color images spanning 50 classes (Xian et al., 2019) (Fig. 2.1b). The well-known ImageNet dataset contains 1.2M color images across 1000 classes (Deng et al., 2009) that span different animals and man-made artifacts.

Models

We created residual CNNs (He et al., 2016a) with ERFs of variable sizes (Table 2.1) by changing the size of the filters of different residual units across layers (Brendel & Bethge, 2019). The residual CNNs consist of 4 blocks that contain 2, 3, 3, and 2 residual units, respectively. Each residual unit consists of 3 convolutional layers: The first and last layers always have filters of size 1×1 and the filter size of the middle layer varies according to Table 2.1. Adjusting the filter size of the residual units results in models with ERFs of either 11, 23, 47, 95, or 227 pixels squared in the last layer. We refer to these models by their ERF sizes, writing ERF23 for a network with an ERF of size 23×23 pixels. Note that since our input images are always of size 224×224 pixels, only the model ERF227 has units in the last convolutional layer with ERFs covering the entire image, before features are globally averaged across spatial locations in the penultimate layer.

Feature-scrambling approach

For the feature-scrambling approach, we build CNN models that are composed of two sub-networks, a *base network* and a *follow-up network* (Fig. 2.1d). The base network transforms the image to high-level feature maps of a given size by being trained on image classification in a standalone way. These pretrained features are then fed into a follow-up network. The follow-up network then further transforms these feature maps in a series of convolutional layers. Finally, features are pooled in a location-discarding way in a global average pooling layer and then a Softmax classification layer. This approach allows us to independently examine the granularity of features used by CNNs for object recognition and to determine to what extent the spatial relations among them contribute to their performance.

We used networks with different ERFs as base networks. We trained them separately for image classification and then detached the fully connected classification layer and the global average pooling layer of the trained network and used it with frozen weights as the base network in our feature-scrambling approach. Subsequently, we attached the follow-up network such that it receives the features of the pretrained base networks as inputs in either a scrambled or unscrambled way. Specifically, for the unscrambled case, we passed the feature maps unchanged to the follow-up network. For the scrambled case, we generated random indices once and used them to permute the feature vectors across spatial locations. The follow-up network is a residual block formed of four residual units. We differentiated between two types of follow-up networks: *with* or *without* spatial aggregation: (1) A follow-up network *with* spatial aggregation has a stride of 2 for its first two residual units and filter size 3×3 for all its residual units. (2) A follow-up network *without* spatial aggregation is formed exclusively of convolutional layers with filter size 1×1 and no down-sampling. In summary, for each of our base networks (ERF11, ERF23, ERF47, ERF95, and ERF227), we trained 3 models depending upon: (1) the type of the follow-up network (with or without spatial aggregation); (2) scrambling the features between the two sub-networks or not.

The considered models can be summarized as follows:

- **Base:** only the base network trained in a standalone way.
- **Base + Follow-up without scrambling:** the model is formed of the pretrained base network plus the follow-up network with spatial aggregation and without feature-scrambling.
- **Base +** 1×1 **Follow-up without scrambling:** the model is formed of the pre-

trained base network plus the follow-up network without spatial aggregation and without feature-scrambling. This model serves as a control for the significance of increasing the ERF of the model by adding the follow-up network.

- **Base + Follow-up with scrambling:** the model is formed of the pretrained base network and the follow-up network with spatial aggregation and with global feature-scrambling during training.

Additionally, the **Base + Follow-up without scrambling** models were tested while the input features to the follow-up network were randomly scrambled either globally or locally.

Training

All simulations were performed using the TensorFlow library (Martín Abadi et al., 2015). We used stochastic gradient descent with momentum $= 0.9$ to update the weights with initial learning rate $= 0.01$ for the first 10 epochs followed by exponential decay for the rest of training. For the ImageNet dataset, we trained for 50 epochs, and for the animals and Sketchy datasets, we trained for 75 epochs.

During training, for non-square images, we first cropped the central square portion of the image with the shortest dimension of the image to keep the aspect ratio of the objects constant before resizing the images to 256×256 pixels. We then applied minimal data augmentation in the form of random right and left horizontal flipping of the images, followed by random cropping of 224×224-pixel patches used for training. During testing, after centrally cropping the images, we resized them to 256×256 pixels, and then we cropped the central 224×224-pixel patch.

Representational similarity analysis (RSA)

We used RSA to investigate the representations of the CNNs of different ERFs (Nili et al., 2014). To avoid the results being biased to the number of classes in each dataset, we sampled 50 random classes from each dataset (the lowest number of classes in the three datasets). Then we sampled 8 random images from each class for a total of 400 images, ran them through all the models of different ERFs, and extracted the activations of the last convolutional layer of each residual unit ($n = 10$), the global average pooling layer (GAP) and the Softmax layer. For the Sketchy and Animals datasets, we averaged the layers' RDMs across 5 repetitions of random initialization. We created the representation dissimilarity matrix (RDM) for each layer by computing the pairwise correlation distance for its activations (400×400 matrix). Next, we computed a second-order RDM for all the layers

of the models (60×60 matrix) by computing the correlation distance between the upper triangle of the layers' RDMs. For visualization purposes, we used multi-dimensional scaling (MDS) to reduce the dimensionality of the second-order RDM to two dimensions.

Minimal recognizable configurations analysis (MIRC)

We adopted the MIRC analysis (Ullman et al., 2016) previously used for humans for CNNs. MIRC analysis is a recursive process that search for the smallest image patches that still yield a correct classification result. MIRC analysis starts with a given, correctly classified image of class c. Starting from the whole image as one patch, four descendant patches are created from each patch. Each descendant batch spans 75% of the height and width of the patch at the previous level starting from one of the four corners (Fig. 2.4a). Each patch is then upsampled using bilinear interpolation to 224×224 pixels to match the input size of the models. The recursive subdivision process continues for each patch as long as the patch is still correctly classified as belonging to class c. Subdivision stops once the classification of a patch is no longer correct. This process defines a tree structure and the leaves of the tree are the MIRCs. The level of a leaf node in the tree is referred to as the level of the MIRC it represents. By construction, the higher the MIRC level, the smaller the patch of the image used for classification.

Results

Feature scrambling during training and testing

We trained CNNs of different ERF sizes on three different datasets: the Sketchy (Sangkloy et al., 2016), the Animals (Xian et al., 2019) (Fig. 2.1a-b), and the ImageNet (Deng et al., 2009) dataset. Example ERFs of five models are shown in Fig. 2.1g. We note that the ERF is a theoretical upper limit on the set of pixels that can activate a given unit, and that not all pixels of the ERF necessarily activate the corresponding deep unit, depending on connection weights. We found that CNN performance increased with ERF size for both the Sketchy and Animal datasets, with a visible saturation for larger ERFs. However, CNN performance depended more strongly on the ERF size for the Sketchy dataset than for the Animals dataset (Fig. 2.1c). Because changing the filter sizes across models will also induce changes in the number of trainable parameters in the models and consequentially their expressive capacity, we performed a control experiment in which we created wide networks with small ERFs but matched the number of parameters of the network

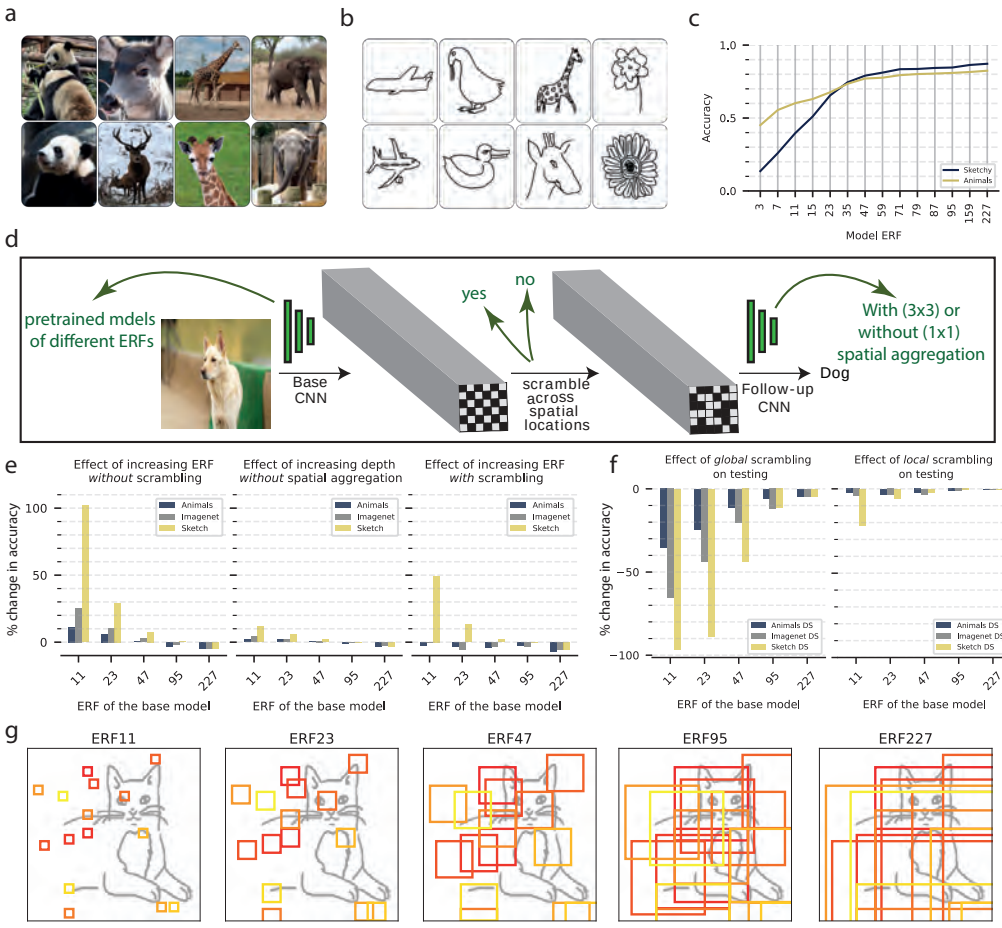


Figure 2.1. Feature scrambling during training and testing. (a, b) Example images for the Animals and Sketchy datasets, respectively. **(c)** CNN performance as a function of the ERF, separately for the Sketchy and Animals datasets. **(d)** A schematic for the feature-scrambling approach. **(e)** Effects of adding the follow-up network to the pretrained base networks either with spatial aggregation without scrambling (left), with spatial aggregation with scrambling (right), or without spatial aggregation (middle). **(f)** Effect of global and local feature-scrambling on the testing performance of the base + follow-up models with spatial aggregation without scrambling. **(g)** A schematic depicting the ERF of random artificial neurons in the last convolutional layer of models of different ERFs.

with the largest ERF (ERF227). We found a slight increase in accuracy, but the models still showed a substantial reduction in performance compared to the corresponding network with a large ERF (Fig. A.S2.1).

The dependence of the classification performance on ERF size suggests that the network's ERF has a major impact on object recognition, especially for textureless

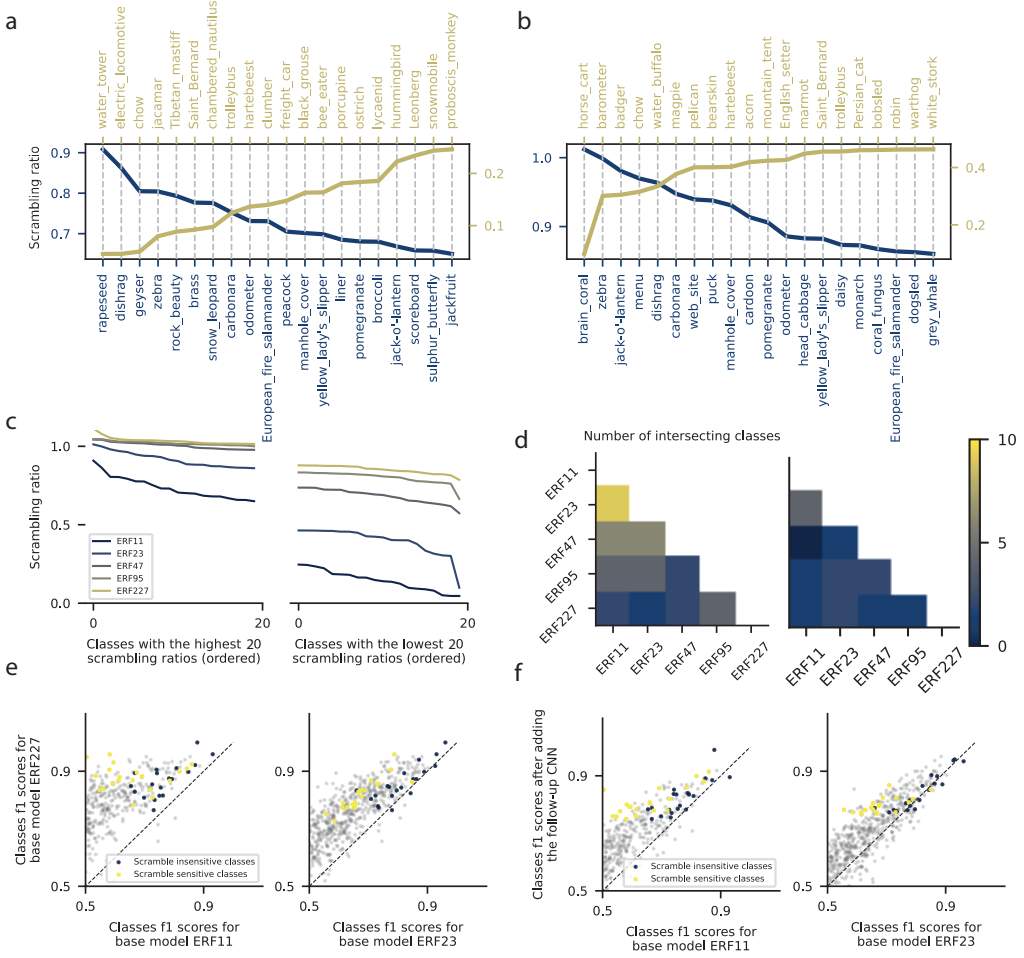


Figure 2.2. (a, b) The 20 least (in blue) and most (in yellow) scrambling-sensitive ImageNet classes for the models ERF11 (a) and ERF23 (b). **(c)** Scrambling ratios of the 20 least (left) and most (right) scrambling-sensitive ImageNet classes for models of different ERFs. High and low values of the scrambling ratio indicate that feature-scrambling has minor and major effects on class performance, respectively. **(d)** Number of the intersecting classes for the 20 least (left) and most (right) scrambling-sensitive ImageNet classes among models of different ERFs. **(e)** $f1$ performance scores of ImageNet classes for ERF11 and ERF23 models against ERF227 model. In blue and yellow are respectively the least and most scrambling-sensitive classes. **(f)** $f1$ performance scores of ImageNet classes for the base model vs. base model after adding the follow-up network. In blue and yellow are respectively the least and most scrambling-sensitive classes.

datasets such as the Sketchy dataset. One explanation for the observed performance increase could be that CNNs with large ERFs can learn to exploit relatively large-scale features, which are especially important for texture-less datasets. However, the comparison between networks with large ERFs and small ERFs does not

yet provide direct evidence that CNNs with large ERFs rely on large-scale shape features. For example, it is possible that the pooling in large ERFs does not take into account the spatial configuration among the features. Instead, the network might just accumulate local evidence in a different manner than networks with smaller ERFs. This reasoning suggests that in order to investigate the network's sensitivity to the spatial configuration of features, it is necessary to distort (i.e. scramble) the spatial arrangement of features and then test the impact of this distortion. Importantly, this scrambling should be done at the level of the network features rather than at the image level, as the latter often leads to confounding high-contrast image artifacts. Specifically, we took the following approach:

1) We trained a network with a small ERF size on an object recognition task. We call this the base CNN, which was not further modified.

2) We then trained a follow-up network, which received input from the last convolutional layer of the pretrained base CNN. These pretrained input features represent diagnostic features of certain granularity depending on the ERF of the base CNN i.e. object parts at different scales. The follow-up network has an ERF that covers the entire image. We observed that adding the follow-up network led to an increase in performance compared to the base network. Consistent with the ERF survey experiment (Fig. 2.1c), the increase in performance was relatively small for the ImageNet and Animals datasets but was large for the Sketchy dataset for base networks with smaller ERFs (Fig. 2.1e, left panel). Absolute performances are shown in Fig. A.S2.2.

3) To rule out the possibility that the observed performance increase for such stacked networks was just caused by increasing the depth of the model by appending the follow-up network, we trained a follow-up network that consisted only of 1×1 convolutions without strides to prevent spatial aggregation. We observed only a slight increase in accuracy for all datasets (Fig. 2.1e middle), which shows that spatial aggregation of inputs was crucial for the observed performance boost (Fig. 2.1e left).

4) To examine whether the spatial configuration of features mattered, we trained the same follow-up networks after spatially scrambling the features in the last convolutional layer of the base network. We used a fixed spatial permutation (i.e. scrambling) of these features that was constant during training. We observed a smaller increase in performance for the Sketchy dataset for ERFs 11 and 23 (Fig. 2.1e right). Furthermore, no further increase in accuracy could be observed for the Animals and ImageNet datasets in this case (Fig. 2.1e right). Taken together, these findings suggest that CNNs can learn to utilize the configuration of

spatially distant features when constructing more complex features in subsequent layers, especially for datasets in which shape is expected to be critical for object classification.

5) As a complementary approach to the fixed scrambling during training, we also performed random feature scrambling during testing. As before, the scrambling was again done at the last convolutional layer of the base network. As predicted, we observed a general decrease in the accuracy of the models with spatial aggregation (base + follow-up) when the features were globally scrambled during testing (Fig. 2.1f left). This effect depended strongly on the dataset, with a relatively weak effect for the Animals dataset and a very strong effect for the Sketchy dataset. Moreover, the performance reduction was particularly pronounced for models with small ERFs that exclusively encode local features of fine granularity before the scrambling is done. It is worth noting that this effect cannot be simply explained by the type of the dataset (sketches versus natural images) since the reduction in performance varied substantially between the Animals and ImageNet datasets, even though both consist of natural images.

6) As a control, we also performed a "local" scrambling, in which the features were scrambled only at neighboring locations. The reduction in performance with local scrambling was much weaker compared to global scrambling, indicating that the loss of performance with global scrambling is due to the distortion of the global configuration of the features, not the confounding effects of the scrambling process itself.

Together, these results highlight the importance of the granularity of features and their spatial configurations for object recognition, especially for datasets in which texture is less informative. In other words, models with larger ERF can extract more coarse-grained features, which are more diagnostic for the object class, i.e. have higher accuracy and are less susceptible to scrambling. These coarse-grained features are diagnostic on their own and do not need to be spatially integrated to construct more complex features in subsequent layers (the follow-up network). However, the granularity of these features differs between datasets.

Variability of classification strategies between classes in ImageNet

Depending on the dataset, we observed different effects of ERF sizes and feature scrambling on network classification performance. Changing the ERF size had the weakest effect on performance for the Animals dataset and the strongest effect

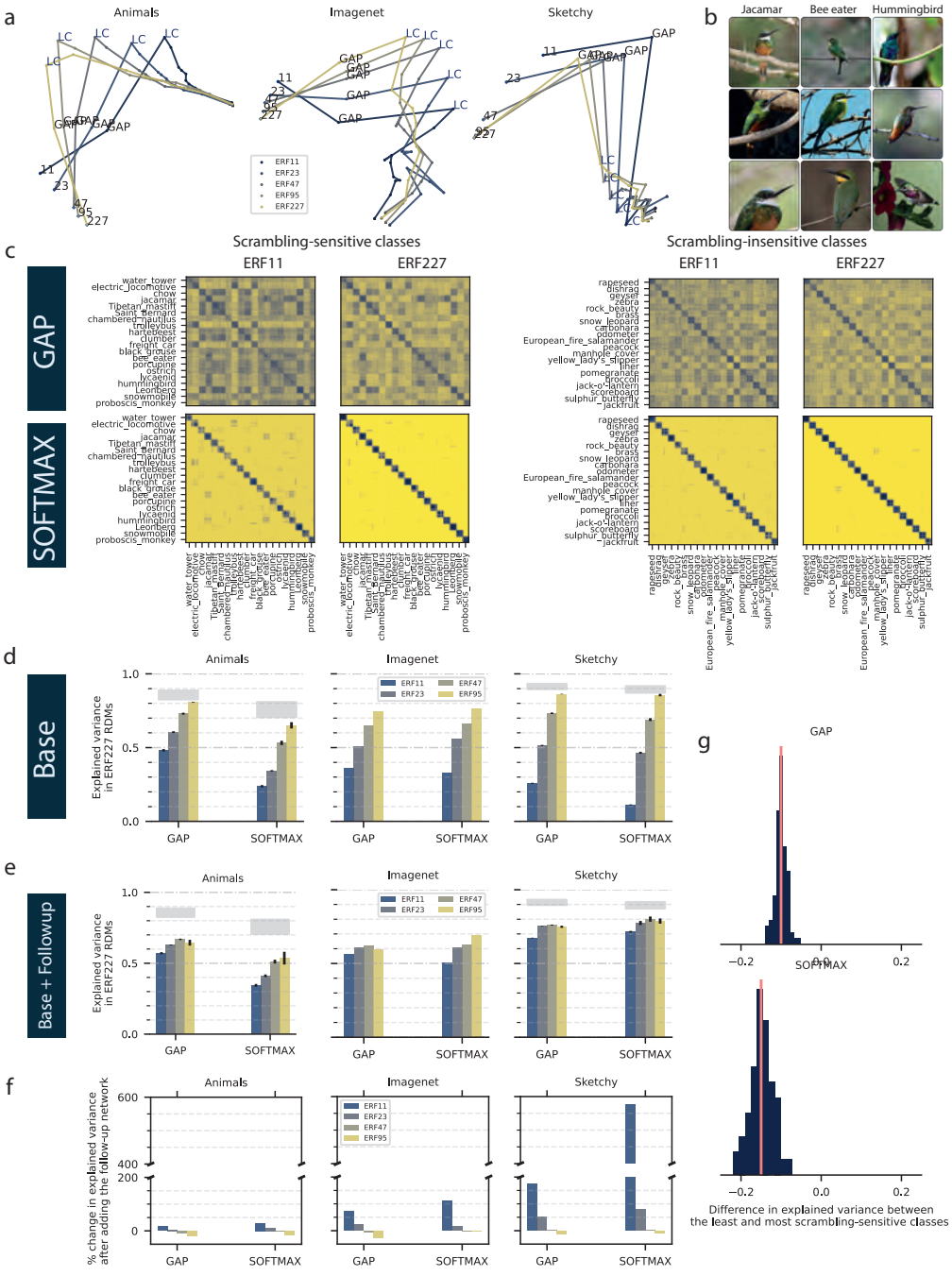


Figure 2.3. See next page

Figure 2.3. (a) Representation trajectories for five CNNs with different ERFs trained on 3 different datasets. For the Sketchy and Animals datasets, we averaged the layers' RDMs (Representational Dissimilarity Matrices) of 5 training iterations of each model of a certain ERF size before computing the second-order RDM of all layers. LC: last convolutional layer. GAP: Global Average Pooling layer. ERF number indicates the classification layer of the corresponding model. **(b)** Each column shows three examples from the ImageNet dataset for three bird classes. **(c)** RDMs of the global average pooling (GAP) and Softmax layers for the models ERF11 and ERF227 computed separately on the 20 least and most feature-scrambling sensitive ImageNet classes as estimated using the ERF11 model and the feature-scrambling approach. We sampled 20 images randomly from each class so each RDM is 400×400 (better viewed digitally). **(d)** The amount of explained variance (R^2) by the GAP and Softmax layers' RDMs of models with different restricted ERFs in the RDMs of the ERF227 model. **(e)** The amount of explained variance (R^2) by the GAP and Softmax layers' RDMs of models with different restricted ERFs after adding the follow-up network in the RDMs of the ERF227 model. **(f)** Percentage change in the amount of explained variance by RDMs of models with different restricted ERFs in the RDMs of the ERF227 model after adding the follow-up network that increases the ERF of the models to cover the whole image. **(g)** The distributions of the difference in explained variance by ERF11 model RDMs in ERF227 model RDMs between scrambling-sensitive and scrambling-insensitive classes. RDMs were computed by randomly sampling images separately from the scrambling-sensitive and scrambling-insensitive classes. The number of repetitions is 100.

for the sketches dataset, with ImageNet in between (Fig. 2.1e left). The strongest effect of feature scrambling was observed on the Sketchy dataset, followed by ImageNet and then the Animals dataset, which was least affected by feature scrambling (Fig. 2.1f left). These findings can be explained by the image statistics in the different datasets. Two extremes are given by the Animals and Sketchy dataset: While images in the Animals dataset can already be classified using local textural features, pictures in the Sketchy dataset require the integration of spatially distant features for classifications. For ImageNet, the classification may allow for different class-specific strategies (e.g., animals vs. man-made artifacts). To test the hypothesis that CNNs use different classification strategies for different ImageNet classes, we used the feature-scrambling approach described above. As a measure of how CNN classification performance is affected by global feature scrambling, we consider the scrambling ratio as the ratio of class $f1$ scores before and after scrambling. A high scrambling ratio indicates that a class is not sensitive to feature scrambling (which we call scrambling-insensitive), and a low value indicates sensitivity to scrambling (which we call scrambling-sensitive). This ranks the classes according to their sensitivity to the global spatial feature configuration in the last CNN layer of the base network (Fig. 2.2a-c). For this analysis, we only

considered classes that the model reliably classified before scrambling ($f1 > 0.75$).

As hypothesized, the least scrambling-sensitive classes predominantly express characteristic surface patterns (texture) such as the rapeseed, brain coral, and zebra classes (Fig. 2.2a and b in blue for base models ERF11 and ERF23 respectively). Scrambling-sensitive classes, on the other hand, were not found to express such characteristics textures, such as the water tower, electric locomotive, and horse cart classes (Fig. 2.2a and b in yellow for base models ERF11 and ERF23 respectively). We hypothesized that the variability in scrambling sensitivity was due to the intrinsic properties of the classes and their performance at low ERFs, rather than to the scrambling operation itself. In fact, we found that classes with high scrambling sensitivity only exhibited this high sensitivity for models with small ERFs (Fig. 2.2c right). However, the scrambling sensitivity of classes was found to be mostly independent of ERF size (Fig. 2.2c left). To confirm that this effect is a consequence of the heterogeneity of the ImageNet dataset and not the ordering process, we repeated the same analysis for the Animals dataset and did not observe such substantial variability in the scrambling ratios among classes, e.g., for the base model ERF11, scrambling ratios ranged from 0.05 to 0.91 and from 0.61 to 0.96 for ImageNet and Animals datasets respectively. We furthermore found that the set of the least scrambling-sensitive classes is mostly consistent across models (Fig. 2.2d left). This is in contrast to the set of the most scrambling-sensitive classes (Fig. 2.2d right). Thus, the performance of scrambling-sensitive classes depends more on the models' ERFs and, therefore, relies on features of coarser granularity.

Therefore, we hypothesized that the scrambling ratio should predict the performance increase from the ERF11 to the ERF227 network (Fig. A.S2.2), as well as the performance increase obtained by adding the follow-up network to the pretrained base network (Fig. A.S2.2 and Fig. 2.1e). Indeed, the performance increase for ERF227 compared to ERF11 and ERF23 was greater for scrambling-sensitive than for scrambling-insensitive classes (Fig. 2.2e). Specifically, the performance ($f1$ score) of the model ERF227 on the 20 most scrambling-sensitive classes was higher than that of all other models (ERF11, ERF23, ERF47, and ERF95) in a statistically significant way according to the Wilcoxon signed-rank test. In contrast, for the 20 least scrambling-sensitive classes, the performance of the ERF227 model was only significantly higher than the models ERF11, ERF23, and ERF47, but not ERF95. Similarly, the performance increase caused by the addition of a follow-up network was larger for scrambling-sensitive classes than for scrambling-insensitive classes (Fig. 2.2f). In particular, increasing the ERF of the models by adding the follow-up network led to a statistically significant increase in the performance of the 20 most scrambling-sensitive classes for the models ERF11, ERF23, ERF47, and ERF95. For the

20 least scrambling-sensitive classes, it only led to a statistically significant increase in performance for the models ERF11 and ERF23.

Representation Similarity Analysis

Next, we investigated the role of ERF size on the classification strategies used by CNNs. We used representation similarity analysis (RSA) (Nili et al., 2014) to test whether CNNs of different ERF sizes develop comparable representations, reflecting similar or different classification strategies (Fig. 2.3a-c; see section). For each layer, we computed a representation dissimilarity matrix (RDM) by computing the pairwise correlation distance on the activations resulting from different images. We then computed the dissimilarity (using the pairwise correlation distance) of the RDMs between all layers of all models, thus indicating the similarity of the representations between different layers of different models. To facilitate visualization, we employed multi-dimensional scaling to reduce the dimensionality of the RDM so that each point in the 2-d space represents a layer in a model and connected the layers of each model with a solid line of a different color (Fig. 2.3a). We observed that models with comparable ERFs are closer in the low dimensional space (Fig. 2.3a), indicating that the distances among the corresponding layers of the models depend on the models' ERF.

To further investigate whether CNNs with small ERFs use classification strategies similar to those of standard CNNs with large ERFs, we correlated the RDMs for all models with the RDM for the ERF 227 model. Specifically, we computed the variance explained (R^2) between the RDMs of the ERF227 model and the RDMs of the models with smaller ERFs (Fig. 2.3d). This was done separately for the Global Average Pooling (GAP) and Softmax layers for the three datasets. For both GAP and Softmax, we observed a gradual increase in the amount of explained variance with ERF size, i.e., models with small ERFs are more dissimilar to the ERF227 model. Moreover, the amount of explained variance depended on the dataset: The Sketchy dataset had the lowest amount of explained variance for models with small ERF, followed by ImageNet and the Animals datasets. This result agrees with the differences between datasets in terms of the models' classification performance (Fig. 2.1e). We repeated the same analysis after adding the follow-up networks to the base models, which in each case increased the ERF to cover the whole image (e.g. $235 pixels^2$ for ERF11 base model) (Fig. 2.3e). We noticed an increase in the amount of explained variance after adding the follow-up network, especially on the Sketchy dataset and for the models with small ERFs (Fig. 2.3f). Again, there was only a minor and intermediate increase for the animals and ImageNet databases, respectively. This supports the notion that CNNs can deploy

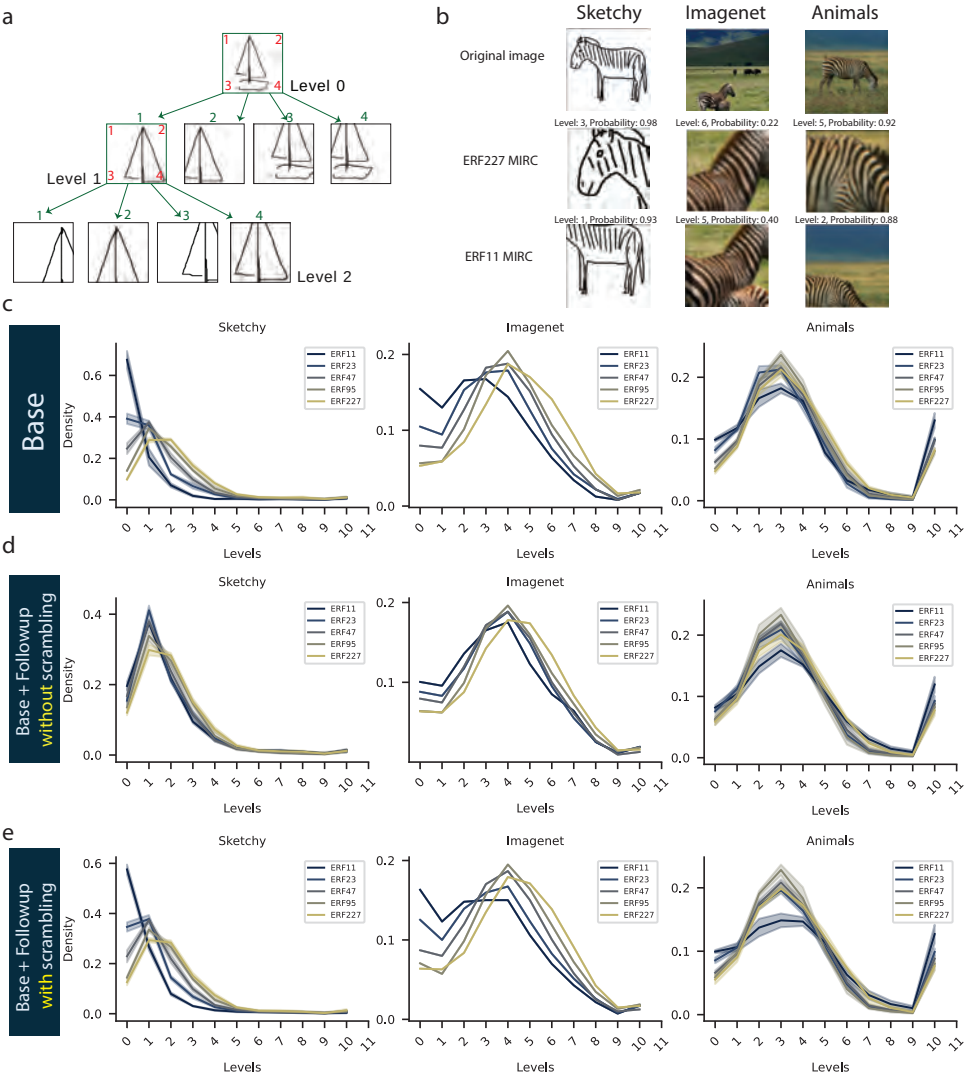


Figure 2.4. **(a)** Illustration of the MIRC procedure. Each image patch yields four descendants. Each descendant is a 75% crop from one of the four image patch corners. Each green numbered descendant patch corresponds to the red equivalently numbered corner of the parent patch (See section). **(b)** Example MIRCs for three different images (first row) of the zebra class from three datasets (Sketchy, ImageNet, and Animals) for the models ERF227 (second row) and model ERF11 (third row). The MIRCs shown are the MIRCs with the highest probability among the MIRCs of the highest level of that image. **(c, d, e)** Distribution of the maximum MIRC level for each correctly classified image in the test dataset for the Sketchy, ImageNet, and Animals datasets, respectively for the base networks of different ERF sizes (c), after adding the follow-up network without scrambling (d), and after adding the follow-up network with the spatial scrambling of its input features (e). For the Sketchy and Animals datasets, the histograms are averaged over 5 training iterations. The shaded area represents the standard deviation. The high frequency of images with MIRCs of level 10 in the Animals dataset is because of the images that belong to the classes that the models usually predict when the correct class cannot be identified.

different classification strategies depending on their ERF.

Furthermore, according to our feature-scrambling analysis, CNN classification strategies should also differ among object classes even within the same model. Therefore, we hypothesized that the explained variance between ERF11 and ERF227 should differ between the scrambling-sensitive and scrambling-insensitive classes. In particular, we expected that the explained variance should be smaller for scrambling-sensitive classes because, for those classes, one expects more spatial integration. For that purpose, we selected the 20 most and least scrambling-sensitive classes of the ImageNet dataset as determined by our feature-scrambling approach for the base model ERF11 (Fig. 2.2a), randomly selected 20 images from each class, passed them through the models ERF11 and ERF227, computed the RDMs of the GAP and Softmax layers for each model (ERF11, ERF227) and condition (scrambling-sensitive, scrambling-insensitive) separately (Fig. 2.3c). We repeated the process 100 times and each time we calculated the variance explained in model ERF227 RDMs by model ERF11 RDMs for both conditions. We subtracted the variance explained in the condition of the scrambling-insensitive classes from the variance explained in the condition of the scrambling-sensitive classes to create a distribution of the difference in the variance explained by model ERF11 in model ERF227 between the scrambling-sensitive and scrambling-insensitive classes (Fig. 2.3g). Indeed, we observed the expected difference (Fig. 2.3g). Additionally, by visual inspection, the difference between the RDMs of the models ERF11 and ERF227 calculated on the scrambling-sensitive classes is especially pronounced in the off-diagonal part of the matrix, which represents the similarity among the inter-class pairs of images (Fig. 2.3c left two columns). We hypothesized that the reason behind this difference is that the ERF11 model extracts lower-level features that are not indicative of a specific class, but rather shared among multiple classes. For example, we observe these blocks of low dissimilarity in Fig. 2.3c (most lower left panel) between the class jacamar and the classes bee-eater and hummingbird, which have shared color and local features (Example samples of each of the three classes are shown in Fig. 2.3b). Together, these results further support our conclusion that the granularity of features used by CNNs (which in terms are determined by their ERF sizes) plays a crucial role in their ability to perform object recognition. Moreover, the granularity of the CNN features is determined not only by its ERF but also by the statistics of the images in the datasets, separately for each class. Although more coarse-grained features can be more reliable for object recognition, they are only exploited by CNNs when needed e.g. the Sketchy dataset and scrambling-sensitive classes in ImageNet. This agrees with the simplicity bias in CNNs (and more generally all neural networks) when trained with a gradient-based learning rule: Networks tend

to become selective to the easiest (and most local) features that allows them to solve the classification task at hand.

Minimal recognizable configurations (MIRCs) analysis

The results so far suggest that CNNs recognize objects based on features that vary in their granularity depending on the dataset and the object class. For datasets and object classes that have relatively little or no texture information, CNNs can learn to construct diagnostic features of coarser granularity from more fine-grained features by exploiting the spatial relations between them. This raises the following questions: 1) What is the spatial extent of these coarse features and spatial relations learned by CNNs? 2) What is the advantage of more coarse-grained features over more fine-grained features for object recognition? The feature-scrambling results shown above indicate that even for the Sketchy dataset, increasing the ERF of the base models beyond 47×47 had a limited effect on performance. This result suggests that the features required for reliably recognizing objects are still predominantly local, i.e., they span maximally about 4-5% of the image.

To further test the reliability of the features utilized by models of different ERFs and visualize them in the image space, we performed a MIRC analysis. MIRC analysis tests the ability of the models to categorize images based on localized image patches by searching for the minimal (i.e. smallest) feature configurations in the image that are still correctly recognizable by the models. We searched for the MIRCs of each image in the test dataset of the Sketchy and Animals datasets, and randomly sampled one-third of the images in the test dataset of the ImageNet dataset. For each image, we cropped 75% of the image starting from each corner so that each image yields 4 descendants (Fig. 2.4a). We then upsampled each descendent crop to the original image size (224×224) and used the model to predict its object class. We repeated the process for each descendant that was correctly classified by the model until we reached the image that was correctly identified by the model but had no correctly classified descendants. This image was declared a MIRC and its level in the search tree defines its size, i.e. the deeper (higher) the level, the smaller the image patch.

In Fig. 2.4b, we show examples of MIRCs generated from three different images for the zebra class from the three datasets and their deepest MIRCs that have the highest classification probabilities using the ERF227 and ERF11 models. These examples show that on the one hand, the ERF227 model was able to classify the image with high classification probability by relying exclusively on relatively local features, i.e. the zebra's face or stripes. On the other hand, the ERF11 model

required larger image patches for successful classification, especially on the Sketchy dataset. This seems to indicate that the model with the larger ERF actually requires a much smaller part of the image to reach the correct classification as compared to the model with the smaller ERF.

To verify whether this finding holds in general, we computed the histograms of the deepest MIRC levels for each image for all datasets and models (Fig. 2.4c-e). We observed for the base models a dependence between ERF size and maximal MIRC levels, i.e., the larger the ERF size of the CNN, the higher its maximal MIRC levels (i.e. a smaller part of the image was sufficient to classify) (Fig. 2.4c). By contrast, networks with smaller ERFs typically cover a larger part of the image or the entire image for classification. We found this dependence to be dataset-specific. The difference between ERF227 and ERF11 was largest for the Sketchy dataset and smallest for the Animals dataset. The difference between ERF227 and models with smaller ERFs was reduced after adding the follow-up network without spatially scrambling the features (Fig. 2.4d). However, the difference was not affected when a follow-up network was added after spatially scrambling the features during training (Fig. 2.4e). The effect of feature-scrambling on the distribution of the levels of MIRCs demonstrates the different strategies CNNs can employ for object recognition. On the one hand, spatial integration of features without scrambling led the follow-up networks to be able to construct and be selective to more reliable coarse-grained features than the base models. Subsequently, these models (base + follow-up) had smaller MIRCs than their base models. On the other hand, spatially scrambling the features before feeding them to the follow-up networks prevented them from exploiting the spatial relations between the features to construct more reliable coarse-grained features. Therefore, the follow-up networks were only able to learn the set of more fine-grained features that correlates with the target class. Subsequently, these models retained the relatively large-sized MIRCs of their base models.

To visualize the features required for recognizing a certain class, we obtained latent representations for all MIRCs of all images of a given class using the model. We then used the k-means algorithm to group the latent representations into 5 clusters. In Fig. A.S2.3, we show examples for the horse and eyeglasses classes of the Sketchy dataset for the model ERF227. For each cluster, we show the eight MIRCs that are the closest to the cluster center and originate from distinct images. We observe that each cluster is composed of MIRCs that represent visually similar features. For example, we observe clusters representing hair, the side view of the head, and leg features for the horse class S2.3. For the eyeglasses class, we can identify a cluster containing double-lined frames, one for thin frames, and one for

reflective glass features.

Discussion

Despite the exceptional performance of CNNs in object recognition tasks (He et al., 2016a; Krizhevsky et al., 2012), the nature of their representations is still poorly understood. One aspect of the learned object representations in CNNs is whether they are capable of encoding the global shape of objects. Global shape representations describe objects in the form of both their diagnostic features and the spatial arrangement of these features (Barenholtz & Tarr, 2006; Biederman, 1987) in contrast to models in which the presence of these features can serve alone as evidence for object identity without encoding the spatial relations between them (Edelman, 1993; Wallis & Rolls, 1997). There exists a wide range of visual features e.g. contours, textures, colors, or object parts. We used features of pretrained CNNs of restricted ERFs to represent the diagnostic local features (Brendel & Bethge, 2019). By comparing the two conditions of training a follow-up network on top of these local features either *with* or *without* scrambling of the spatial locations of the features, we could assess the amount of additional information that CNNs can extract by exploiting the spatial relations between features. Moreover, by examining the MIRCs of CNNs, we were able to evaluate the spatial extent of spatial relations learned by CNNs for object recognition.

It has recently been reported that CNN representations may be mostly local (Baker et al., 2020; Brendel & Bethge, 2019) and consequently more biased toward object surface regularities (Geirhos et al., 2019; Jo & Bengio, 2017) than the global form of objects. This led to the hypothesis that they might not be capable of representing spatial relationships among features (Baker & Elder, 2022; Baker et al., 2018). In contrast to conclusions drawn in other works, our analysis allows us to provide the following more nuanced view: (1) We provide evidence that CNNs are capable of using relatively long-range spatial relationships for object classification, especially for textureless datasets (such as sketches). This finding is supported by several analyses, including a new scrambling approach in which we perturbed spatial relations between features within the CNN, and a systematic investigation of how CNN performance is impacted by different effective receptive field sizes. (2) We show that CNNs use different strategies for different datasets, rather than one unified strategy (e.g. pooling evidence based on local texture). Notably, we found that classification strategies can vary even between classes within the same dataset. These strategies differ in the granularity of the features used and in the degree of reliance on the spatial relations between them. This suggests that there is a continuous spectrum of CNN strategies, ranging from exclusive

reliance on local features (insensitive to spatial relations, found for example for the Animals dataset and the scrambling-insensitive classes in ImageNet) to a stronger reliance on spatial relations (for example for the Sketchy dataset and the scrambling-sensitive classes in ImageNet). (3) We furthermore show to what extent spatial relations among features are used by CNNs to perform object recognition tasks. In particular, we provide evidence that the spatial arrangement of features is used only to construct features up to an intermediate level of granularity. That is, we did not find evidence of spatial integration in CNNs that allows them to capture the global shape of the objects in the datasets tested.

One possible explanation for a bias towards local features is the locality of the convolution operation (Baker et al., 2020). However, our finding that CNNs learned features of intermediate granularity for classification agrees with another possible explanation, namely that a bias to local features and not to global shape is a consequence of the optimization process for object classification (G. Malhotra et al., 2022). Specifically, from an information-theoretic perspective, features of intermediate granularity are the most informative for image classification tasks (Ullman et al., 2002). The idea is that, on the one hand, very complex features could be highly diagnostic because their presence gives high confidence about the class identity. However, on the other hand, these complex features may not be sufficiently sensitive (i.e. they do not exist in each exemplar) to be generalizable across exemplars of an object class. By contrast, very simple features would generalize better, but in addition would also lead to more false positives (i.e. lower specificity). Thus, features of intermediate complexity can provide an optimal trade-off between sensitivity and specificity (Ullman et al., 2002). Interestingly, it has been shown that randomly initialized CNNs display an increase in the representational structure similarity from early to late layers between different levels of abstraction of visual stimuli (photos, drawings, and sketches) (Singer et al., 2022). However, after training CNNs on ImageNet, they showed a drop in the representational structure similarity in later layers after peaking in the intermediate layers which consequently led to lower classification performance on drawings and sketches. These results show that the optimization process and not the CNN architecture steers the representations to be biased to more local features that are optimal for solving its objective function. Along the same lines, prepending regular CNNs with a fixed non-trainable bank of Gabor filters led to better out-of-distribution generalization to line drawings, silhouettes, robustness to noise corruptions (Evans et al., 2022) and adversarial attacks (Dapello et al., 2020). These findings further suggest that similar to the sketchy dataset, limiting the amount of surface information through the Gabor filters led the CNNs to depend on more coarse-grained features that were more robust to pixel corruptions and more generalizable to

different visual domains.

The bias towards local features can also be related to the idea of simplicity bias of neural networks, which states that neural networks preferentially extract the simplest features needed to solve a given task (G. Malhotra et al., 2020; Shah et al., 2020). Consistent with this explanation, our MIRC analysis showed that models with small ERFs that by design are only capable of extracting simpler fine-grained features require larger patches of images for correct object recognition (because they have lower specificity). In contrast, models with larger ERFs that are capable of extracting more coarse-grained and more specific features were able to assign objects to their corresponding correct classes based on smaller image patches. Therefore, our results suggest that optimization for object recognition is unlikely to yield bias to the global shape of objects, even if the models have the capacity to learn it. A similar principle may hold for human vision, as it has been shown that in humans shape bias can be task- and context-dependent (Cimpian & Markman, 2005; Diesendruck & Bloom, 2003; Yoshida & Smith, 2003).

Our results have major implications for the ongoing discussion concerning shape and texture representation in CNNs, and whether certain biases exist. There is little consensus about the extent to which CNNs are texture- or shape-biased. Some studies have suggested that CNNs are shape-biased (Kubilius et al., 2016; Ritter et al., 2017; Tartaglini et al., 2022), whereas others have suggested that CNNs are strongly texture-biased (Baker & Elder, 2022; Baker et al., 2018, 2020; Geirhos et al., 2019). Here, instead of using a shape-texture dichotomy to understand the nature of CNN representations, we have used the dichotomy of local vs. global features. We argue that this dichotomy is useful for two reasons: 1) It can be quantified without specific interpretations of what constitutes texture or shape, as we showed with our feature-scrambling approach. In fact, our approach does not test specific assumptions about the nature of the representations because we do not perform specific image manipulations to provide evidence for either texture or shape bias. Rather, we manipulate the network architecture and the spatial arrangement of the representations to determine the locality of the features. 2) It is flexible in that it allows local features to be both shape-like or texture-like. This means that the shape-texture dichotomy only maps partially to the global-local dichotomy. For example, this dichotomy is able to account for the existence of highly diagnostic shape features of fine granularity that are highly specific and sensitive (e.g., the nose of a dog). Indeed, when ranking ImageNet classes according to their scrambling sensitivity, it is not always obvious that the scrambling-sensitive classes would map to shape classes as would be intuitively expected. A possible explanation for previous inconsistent findings with respect to

shape and texture is that the respective studies made very specific manipulations that did not generalize beyond these examples. For example, the texture bias observed in CNNs trained on ImageNet when tested on shape-texture cue conflict stimuli (Geirhos et al., 2019) was significantly reduced when the background of the images was removed (Tartaglini et al., 2022). Our findings suggest an explanation for these observations, in that the fine-grained (texture) features are less reliable than the more coarse-grained (shape) features, and therefore need to cover a large portion of the image to be diagnostic. Removing them from the background reduced their predictive power and led CNNs to be more shape-biased (Tartaglini et al., 2022) (on this specific test set). Another example is that many studies used silhouette stimuli to test shape bias in CNNs (Baker & Elder, 2022; Geirhos et al., 2019; Kubilius et al., 2016) and reached different conclusions. However, they used different datasets containing different classes. According to our results, this is expected since CNNs employ different classification strategies per object class and consequently will lead to variable classification performances on silhouette stimuli if the classes are different.

Given that CNN models are currently used as models of brain activity, specifically for the ventral stream of the visual system, which is believed to be responsible for object recognition (Cadieu et al., 2014; Cichy et al., 2016; D. L. K. Yamins et al., 2014; D. L. Yamins & DiCarlo, 2016), it is important to understand the representations they develop and how they deviate from the brain. Although humans rely mostly on complex shape cues for object recognition (Landau et al., 1988), recent evidence has shown that the categorical organization of the entire ventral stream can be explained by mid-level features that do not include intact objects and do not convey any semantic information (Ayzenberg & Behrmann, 2022a; Henderson et al., 2022; Jagadeesh & Gardner, 2022; Long et al., 2018). Moreover, albeit it is likely that humans rely on more than one mechanism to object recognition (Peissig & Tarr, 2007; Smith, 2009), some of these mechanisms might only depend on patchy diagnostic local features (Ullman et al., 2001) especially given the fact that humans are capable of recognizing familiar objects from local image patches (Ullman et al., 2016) and these image patches evoke responses in higher-order category-selective visual areas (Holzinger et al., 2019). Furthermore, it has been reported that human children's ability to recognize objects based on their global shape begins to develop only at 18-24 months of age (Pereira & Smith, 2009; Yee et al., 2012). Before that, they are capable of recognizing objects based solely on their local features. In general, it has been shown that categorization of objects in humans relies on combinations of different perceptual and high-level semantic mental object representations constructed to model human similarity judgments (Hebart et al., 2020). These results bear a resemblance to our findings

of the heterogeneity of CNNs classification strategies across different datasets and different classes in ImageNet. The heterogeneity of CNN classification strategies across datasets also agrees with the observations in the literature that CNNs trained for object recognition rely on higher and wider distributions of spatial frequencies than CNNs trained on face recognition and consequently exhibited less robustness to blurring (Jang & Tong, 2021) and it is believed that humans recognize faces holistically as a whole in contrast to objects that can be recognized as a set of independent features (Grand et al., 2004; Tanaka & Simonyi, 2016). Our results, therefore, provide additional evidence for the hypothesis that features of intermediate granularity which are optimal for object recognition (Ullman et al., 2001, 2002) could be shared between CNNs and the ventral stream of the visual cortex (Henderson et al., 2022; Jagadeesh & Gardner, 2022; Long et al., 2018).

In summary, we showed here that although CNNs do not exploit global shape representations to perform object recognition, they can learn to utilize distributed feature constellations if this is required for solving the object classification task at hand. Looking ahead, we hypothesize that developing new tasks and objective functions to train CNNs instead of object recognition might lead to biases more aligned with humans. Reinforcement learning (RL) is a candidate objective function because it has been suggested that manual exploration may be a key factor in the development of shape bias in children (Pereira et al., 2010; Soska & Johnson, 2008) and it has been shown that action planning using RL leads to divergent representation than supervised and unsupervised learning (Lindsay et al., 2021). Moreover, neural agents that are trained to communicate efficiently i.e. be optimal on the trade-off between informativity and complexity of the messages used were shown to exhibit shape bias (Portelance et al., 2021). Future investigations of such novel objective functions can not only lead to more effective biases and representations in such networks but also shed more light on how the observed human biases emerge.

Conclusions

We provide evidence that CNNs have the capacity to learn the spatial relations between features for object recognition. Specifically, the spatial arrangement of features is exploited by CNNs to build more coarse-grained features that are more reliable for object classification. Notably, the capacity of CNNs to learn the spatial arrangement of features varies according to the dataset and according to the class within the same dataset. We noticed, however, that CNNs employ the spatial configuration of features to build more coarse-grained features only up to an intermediate degree of granularity and do not exploit the global shape of objects.

The reason for this is that features of intermediate granularity are more likely to be optimal in the trade-off between sensitivity and specificity i.e. generalizable and yet reliable.

Declaration of competing interest

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

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix

Controlling for the number of parameters of the models. We performed a control analysis to verify that the performance differences observed in our study among CNNs of different ERFs can be attributed indeed to their ERFs and not the number of model parameters. We trained a wider model of small ERF (11×11 pixels) but *with* matched the number of parameters to the model with the largest ERF (227×227 pixels). For both the Animals and Sketchy datasets, we observed a slight increase in the classification performance of the models by increasing the number of parameters. However, a small ERF model with a large number of parameters did not reach the performance of the model with the largest ERF, indicating the importance of the ERF to the models' performance. Furthermore, for the Sketchy dataset, the performance of the wider model with ERF = 11×11 did not even reach the performance of the regular model with ERF = 15×15 pixels. This is in line with our other results showing the reliance of the performance of CNNs on their ERF size, especially for the Sketchy dataset. Note that in the manuscript, we included several additional controls, e.g. scrambling during training, a 1×1 follow-up network, and local scrambling, which further show the importance of ERF size.

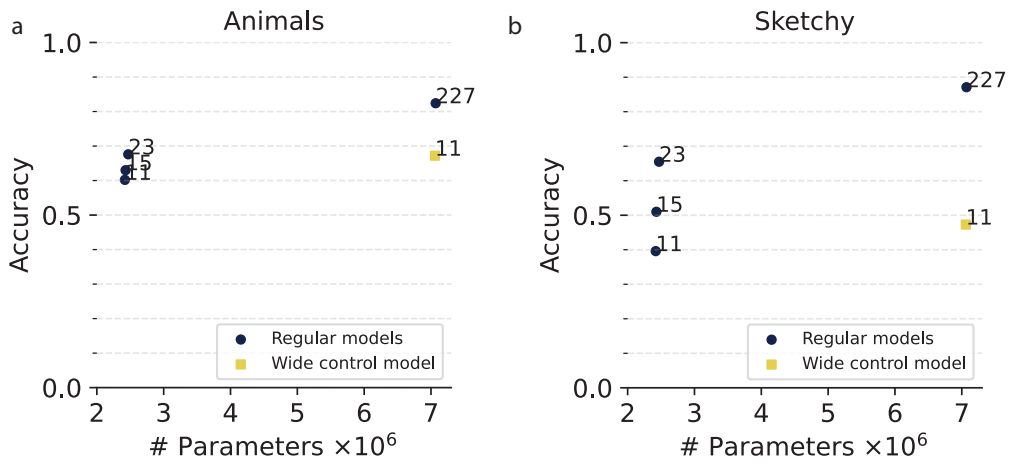


Figure S2.1. A control experiment in which we trained a wide model that has a small ERF (11 pixels), while matching the number of parameters of the model with the largest ERF (227 pixels). The numbers shown in the figure are the ERF of the corresponding models in pixels.

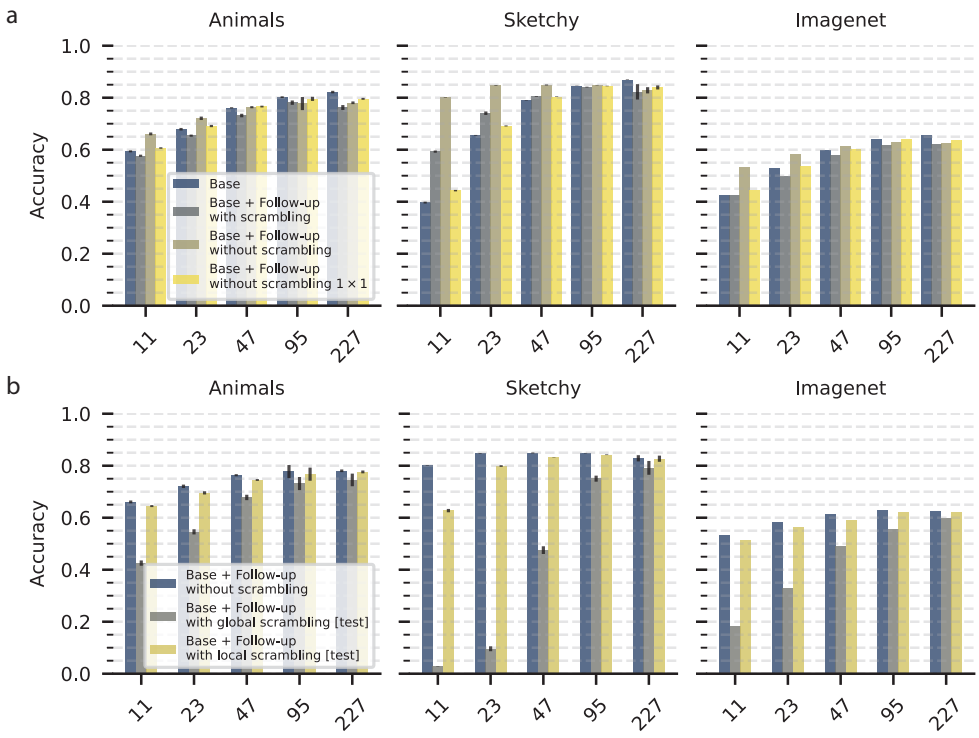


Figure S2.2. (a): Classification accuracy for CNN models of different ERFs under different training conditions of the feature-scrambling approach (Fig. 2.1c). **(b)**: Classification accuracy of the base models with spatial aggregation without scrambling under different testing conditions (global and local scrambling).

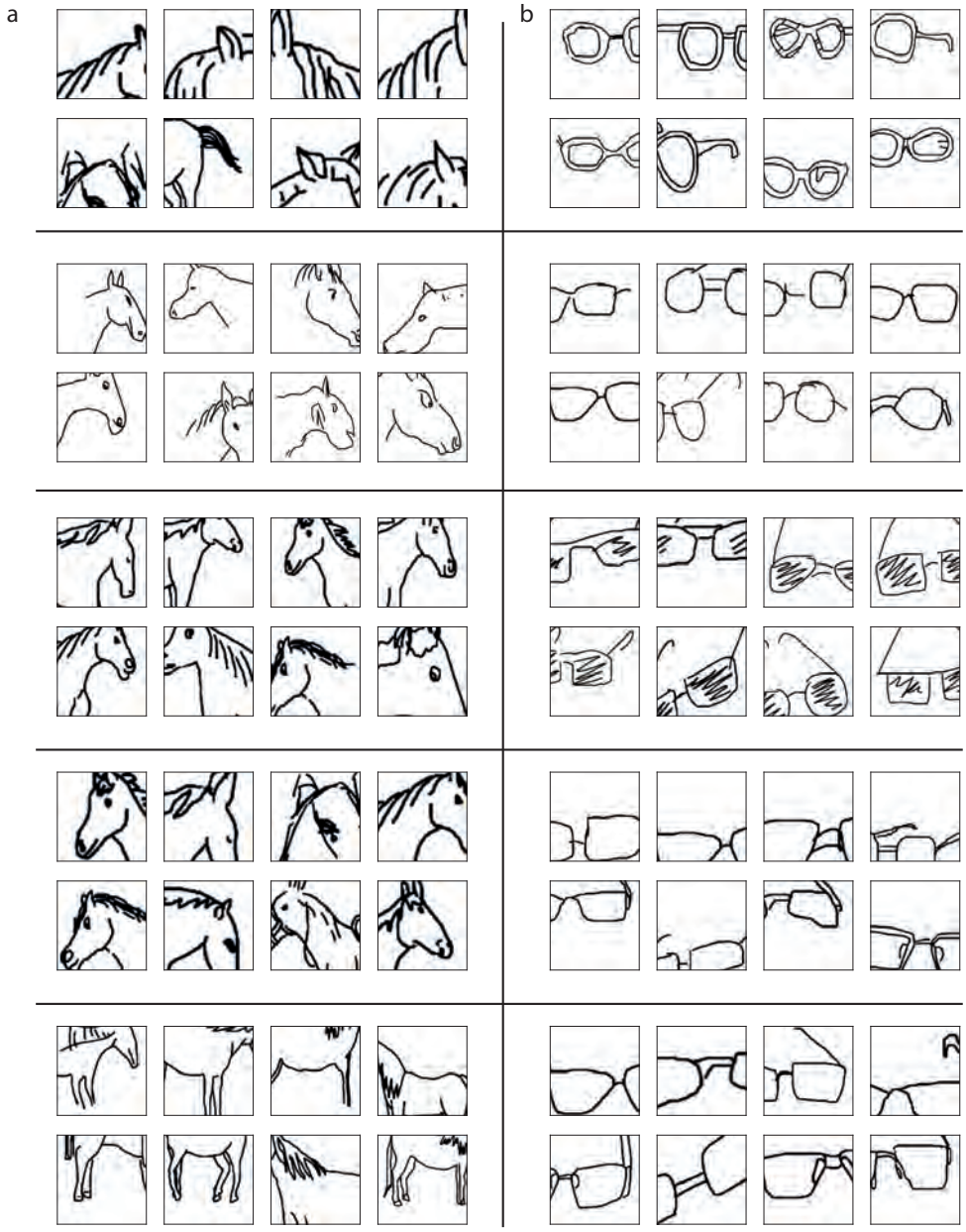


Figure S2.3. Clustering of all the MIRCs of the horse (a) and eyeglasses (b) classes (Sketchy dataset) in the representational space of the model ERF227 Each panel shows the eight closest MIRCs, generated from unique test images, in the representational space to the center of one cluster.

Chapter 3

Neural responses in early, but not late, visual cortex are well predicted by random-weight CNNs with sufficient model complexity

In revision for Communications Biology and preprinted as: **Farahat, Amr**, and Martin Vinck. "Neural responses in early, but not late, visual cortex are well predicted by random-weight CNNs with sufficient model complexity." bioRxiv (2025): 2025-02.

Abstract

Convolutional neural networks (CNNs) were inspired by the organization of the primate visual system, and in turn have become effective models of the visual cortex, allowing for accurate predictions of neural stimulus responses. While training CNNs on brain-relevant object-recognition tasks may be an important pre-requisite to predict brain activity, the CNN's brain-like architecture alone may already allow for accurate prediction of neural activity. Here, we evaluated the performance of both task-optimized and brain-optimized convolutional neural networks (CNNs) in predicting neural responses across visual cortex, and performed systematic architectural manipulations and comparisons between trained and untrained feature extractors to reveal key structural components influencing model performance. For human and monkey area V1, random-weight CNNs employing the ReLU activation function, combined with either average or max pooling, significantly outperformed other activation functions. Random-weight CNNs matched their trained counterparts in predicting V1 responses. The extent to which V1 responses can be predicted correlated strongly with the neural network's complexity, which reflects the non-linearity of neural activation functions and pooling operations. However, this correlation between encoding performance and complexity was significantly weaker for higher visual areas that are classically associated with object recognition, such as monkey IT. To test whether this difference between visual areas reflects functional differences, we trained neural network models on both texture discrimination and object recognition tasks. Consistent with our hypothesis, model complexity correlated more strongly with performance on texture discrimination than object recognition. Our findings indicate that random-weight CNNs with sufficient model complexity allow for comparable prediction of V1 activity as trained CNNs, while higher visual areas require precise weight configurations acquired through training via gradient descent.

Introduction

The development of convolutional neural networks (CNNs) was originally inspired by features of the primate visual system, such as its hierarchical organization (Felleman & Van Essen, 1991; Vezoli et al., 2021) and localized receptive fields (RFs) with repeated feature kernels across space (Fukushima et al., 1983; Hubel, Wiesel, et al., 1959; LeCun et al., 1989). CNNs, and deep neural networks (DNNs) in general, have in turn become effective models of the primate visual ventral stream, allowing for relatively accurate prediction of neural responses to novel, natural stimuli (Cadena et al., 2019; Güçlü & Van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; D. L. K. Yamins et al., 2014; Zhuang et al., 2021). The efficacy of task-optimized CNNs in predicting a given brain area's neural responses (henceforth referred to as "encoding performance") is thought to depend on several factors, such as the network architecture, the objective function, training dataset and learning rules used for training(Cichy & Kaiser, 2019; Doerig et al., 2023; Richards et al., 2019).

Although earlier work postulated that CNNs can effectively predict neural activity because they are trained on ecologically relevant object recognition tasks (Cadieu et al., 2014; Mehrer et al., 2021), the effect of training may differ substantially between hierarchical levels of the primate ventral stream. It stands to reason that training networks on a diet of natural images with an objective function probing for invariant image classification (e.g. supervised object recognition (D. L. K. Yamins et al., 2014) or contrastive self-supervised learning (Zhuang et al., 2021)) is important for predicting neural activity in higher regions of the primate ventral stream, given their high degree of functional specialization (Cadieu et al., 2014; DiCarlo et al., 2012; Hung et al., 2005; Rust & DiCarlo, 2010). Indeed, several studies highlight the importance of training objectives and datasets in predicting the activity in higher primate ventral stream regions, while suggesting that differences in architecture between trained CNNs have a smaller influence on explaining responses throughout the primate ventral stream (Conwell et al., 2024; Storrs et al., 2021; Zhuang et al., 2021). Yet, it is less clear to what extent training CNNs is essential for predicting activity in early visual areas of the primate cortex, which show much less functional specialization and may be involved in a wider range of functions beyond object recognition including scene segmentation (Self et al., 2013), motion processing (Gur & Snodderly, 2007), and salience detection (Li, 2002). While CNNs trained for object recognition outperform traditional models like linear-nonlinear Poisson models and Gabor filters in predicting macaque V1 responses to natural images (Cadena et al., 2019; Simoncelli et al., 2004; Willmore et al., 2008), this superior prediction may be either due to the CNN's architecture

or the used training objective. A recent study showed that the RF size of neurons in object-recognition trained CNNs was an important determinant of encoding performance, suggesting that the CNN's architecture does play an important role for predicting V1 activity (Miao & Tong, 2024).

Furthermore, it is plausible that differences between trained and untrained CNNs in predicting neural activity depend strongly on the initial architecture of the CNN with random weights. Recent studies showed that the generalization capacity of DNNs can be attributed to their loss landscapes upon initialization dictated by their architectural design (Chiang et al., 2022; Ramasinghe et al., 2022). In particular, random-weight DNNs can show strong differences in the complexity of input/output functions dependent on e.g. the non-linear activation function used in the network (Teney et al., 2024). It is possible that training CNNs steers them towards a certain non-linear complexity matching neural complexity, thereby masking initial differences in architecture, but that a random-weight CNN with an appropriate RF size and non-linear complexity may already allow for accurate prediction of brain activity.

Here, we systematically test whether certain architectural components contribute to CNNs' ability to encode neural data of early and high visual areas in primates' brains. Specifically, we constructed CNN models with a linear readout to predict neural data and investigated when training the convolutional filters is necessary for good encoding performance versus only training the linear readouts for an otherwise random-weight CNN.

Results

Neural encoding performance of VGG16 model

We analyzed three neural datasets: (1) Firing rates of 166 V1 neurons from two macaques, recorded while the animals passively viewed natural and texture images (Cadena et al., 2019); (2) activity from 168 multi-unit sites in the IT cortex of two macaque monkeys, passively viewing 3200 grayscale images (Cadieu et al., 2014); and (3) the fMRI Natural Scenes Dataset (NSD) (Allen et al., 2022), comprising fMRI responses from 8 human subjects viewing thousands of color natural images. Neural activity was predicted by linearly transforming the three-dimensional activation maps of each convolutional layer in the VGG16 model into a one-dimensional vector representing neural activity (either firing rates for macaque datasets or voxel activations for the human fMRI dataset). See supplementary Fig. S3.1 For a visual illustration of the models. The weights of this

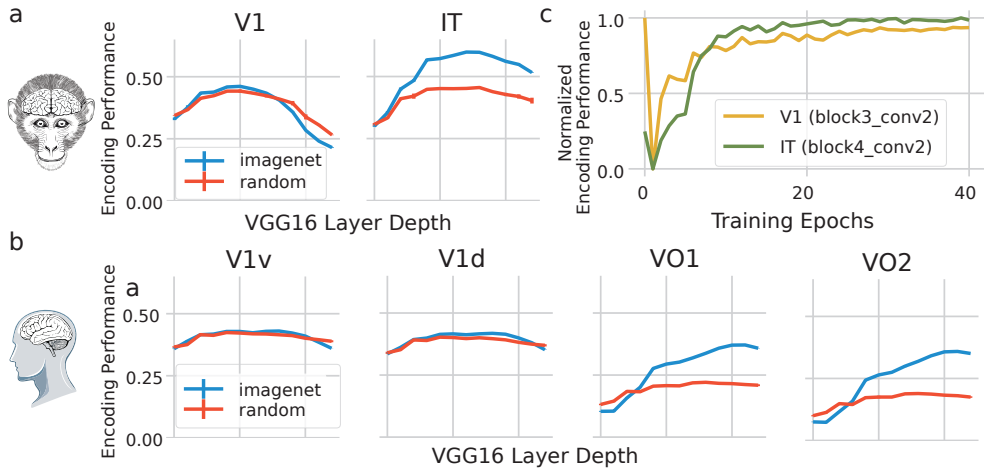


Figure 3.1. Encoding performance of VGG16 model. **(a)** Encoding performance of a linear readout optimized on top of the representations of the convolutional layers of VGG16 model either upon random initialization (in red) or pretraiend on ImageNet dataset for object recognition (in blue) for two neural datasets recorded from the early visual cortex V1 (left) or the higher visual area IT (right) in macaques. **(b)** same as **a** but for two early visual cortex ROIs (V1v and V1d) or higher visual areas (VO1 and VO2) in huamn fMRI data. **(c)** The normalized encoding performance of the best performing layer for V1 and IT brain areas in macaques tracked over training time on ImageNet dataset from random initialization till full convergence. Each line is the average of 3 training iterations.

linear transformation were fitted on 80% of the dataset and tested on the remaining held-out test set. The "encoding performance" for each VGG16 layer was quantified as the median Pearson correlation between the predicted and actual neural responses across all neuronal sites. Consistent with previous findings, the layers of the ImageNet-trained VGG16 model could predict a substantial amount of variance in neural responses to unseen (i.e. test-set) stimuli in both early and higher visual areas of the macaque and human brain (Fig. 3.1).

Next, we investigated the influence of object recognition training on the encoding performance. When the VGG16 model was initialized with random weights, thereby omitting task training, its ability to predict primate V1 responses showed only minor differences compared to the ImageNet-trained VGG16 model (Fig. 3.1; difference trained vs. random-weight CNNs: 0.009 for macaque V1, 0.008 for human V1v, 0.012 for human V1d). By contrast, a substantial decrease in encoding performance was observed when predicting responses in higher visual areas using random-weight CNNs (inferotemporal cortex (IT) in macaque and ventral occipital areas (VO1 and VO2) in humans (Fig. 3.1; difference trained vs. random-weight CNNs: 0.132 for macaque IT, 0.138 for human VO1, and 0.176 for human VO2).

The loss in encoding performance for IT was statistically much larger than for V1 (Mann-Whitney U rank test $p \ll 0.001$). We found qualitatively similar results using other popular convolutional models such as Resnet50 (He et al., 2016b), Inception (Szegedy et al., 2015), and DenseNet (Huang et al., 2016) (see supplementary Fig. S3.3).

To examine the impact of ImageNet training on the encoding performance of VGG16, we quantified the encoding performance across each training epoch, starting with randomly initialized weights and progressing to full convergence (Fig. 3.1c). Although the network's performance on object recognition improved monotonically with training, the encoding performance showed a markedly different profile: Starting from the first epoch, the encoding performance for V1 declined notably after the initial training epoch compared to the randomly initialized weights. Hence, training on the object recognition initially decreases the encoding performance, i.e. the ability to predict V1 activity. The encoding performance for V1 only recovered upon reaching full convergence. For IT, however, the encoding performance mostly showed a monotonic increase from the first towards the last training epochs.

Together, these findings indicate that training a CNN architecture (VGG16) on object recognition is not essential for predicting primate V1 activity, as random-weight CNNs demonstrate comparable performance to trained CNNs.

Simple convolutional models for encoding early and higher visual areas

To further investigate the efficacy of randomly initialized networks in predicting neural responses, we constructed simpler CNN models, systematically changed their architecture and training, and then evaluated their neural encoding performance across various brain regions in macaques and humans. We varied network depth between shallow (2 layers) and deeper (4 layers) architectures, while adjusting convolutional kernel sizes to maintain consistent receptive field sizes across models (see Methods and supplementary Fig. S3.2). Additionally, we evaluated average and maximum pooling operations and four distinct activation functions (ReLU, ELU, Tanh, and Linear). In this case, we optimized the neural network weights and a linear readout to directly predict neural activity, rather than training on an object recognition task (see Methods), similar to a previous study that developed a shallow neural-network model to predict V1 activity (Du et al., 2024).

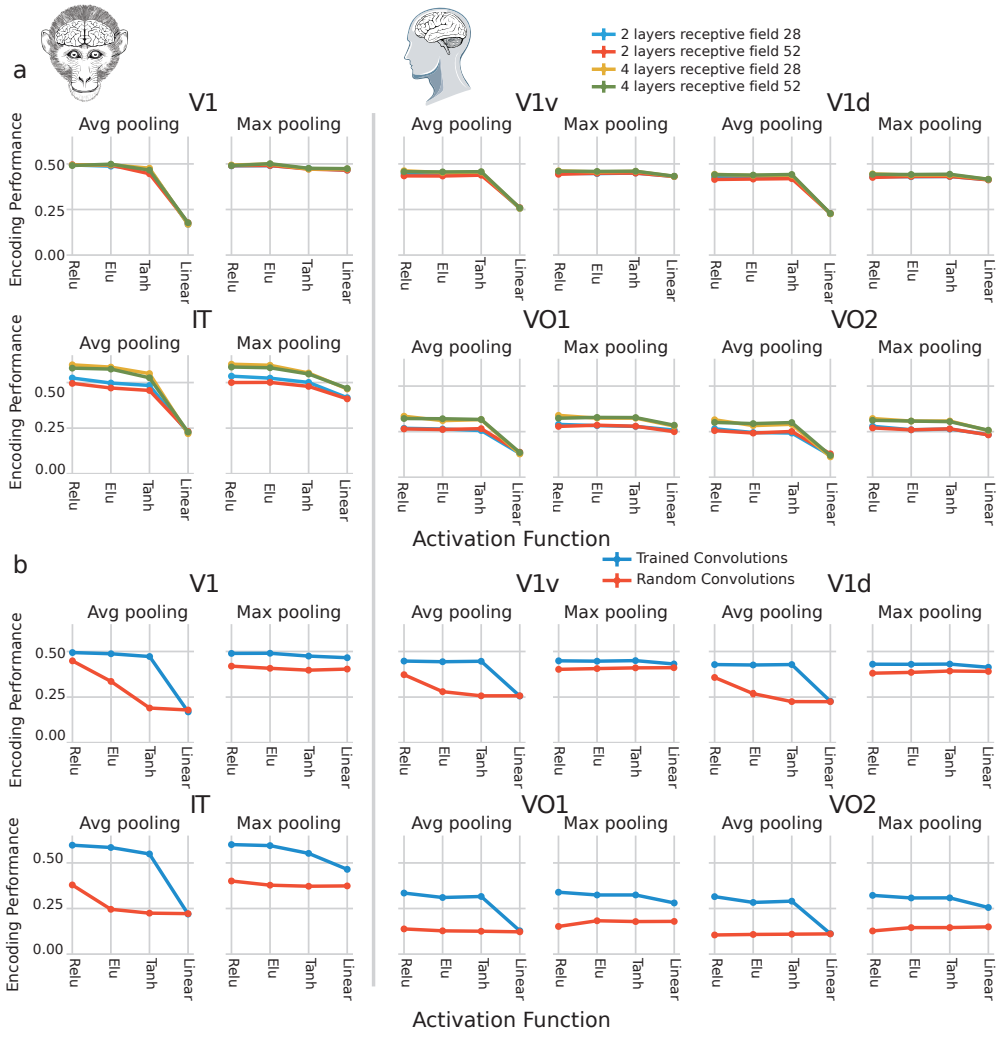


Figure 3.2. Encoding performance of simple convolutional models on early and higher visual areas of macaques and human brains. (a) Encoding performance of shallow (2 layers) and deeper (4 layers) convolutional models each manipulated to have effective receptive field of either 28 or 52 pixels^2 fully trained on neural data from early (V1, V1v, and V1d) and higher (IT, VO1, and VO2) visual brain areas of macaques and humans. Each model has 8 variants: 2 different pooling strategies (average pooling and maximum pooling) and 4 different activation functions (ReLU, ELU, Tanh, and Linear). **(b)** Encoding performance of the best performing fully trained models from **a** (in blue) – the 2-layer 28×28 models for V1, V1v, and V1d brain areas and the 28×28 4-layer models for IT, VO1, and VO2 brain areas – and their randomly initialized counterparts (in red). In the latter case, only the linear readout was trained on top of the randomly initialized weights. Each line is the average of 5 training iterations.

For the trained models, we observed that, with the exception of linear networks,

all models achieved comparable performance for the V1 area in both macaques and humans (Fig. 3.2a). Specifically, linear networks employing average pooling exhibited poor encoding performance, while different non-linear activation functions or linear networks utilizing max pooling operations yielded higher and comparable V1 encoding performance. The V1 encoding performance was comparable between shallow and deep CNNs. The performance of the models in predicting responses from higher visual areas (IT, VO1, and VO2) was comparable across non-linear activation functions but improved for deep compared to shallow networks. In sum, when convolutional filters are optimized, the architectural bias of the models is subtle, i.e., the performance difference between different activation functions and pooling mechanisms is almost negligible (except for linear networks with average pooling).

We then evaluated the performance of the networks when only the linear readout was optimized, while the convolutional filters were frozen at their randomly initialized weights. For these and subsequent analyses, unless otherwise specified, we focused on shallow 2-layer networks for the early visual cortex and the deeper 4-layer networks for higher visual areas. The comparison of the encoding performance of models with trained convolutional layers to those with randomly initialized weights showed that random ReLU networks approached the performance of their trained counterparts in predicting V1 responses for both macaques and humans (Fig. 3.2b). The differences between random and trained ReLU networks were 0.045 for the average pooling models and 0.068 for the maximum pooling models. Compared to V1, there was a much greater difference in encoding performance in higher visual areas (IT, VO1, VO2; Fig. 3.2b). The differences between random and trained ReLU networks were 0.214 for the average pooling models and 0.196 for the maximum pooling models. In contrast to fully trained networks, networks with randomly initialized convolutional weights showed substantial differences in V1 encoding performance across the activation functions, especially those using average pooling operations (Fig. 3.2b; $s = 0.106$ for random nonlinear models and $s = 0.009$ for trained nonlinear models).

In summary, random ReLU networks achieved significantly higher encoding performance than other random networks for both early and higher visual cortices. Furthermore, random networks with max pooling operations exhibited substantially higher performance than their counterparts with average pooling, except for ReLU networks, which achieved comparable encoding performance in both scenarios. However, the difference between the best performing random architecture and its trained counterpart was substantially smaller for early visual cortex in comparison to higher visual areas. In conclusion, we identified the

ReLU activation function and maximum pooling as key architectural components that significantly contribute to the V1 encoding performance of CNNs. This is evidenced by the comparable performance achieved by randomly initialized models that incorporate these components as compared to their fully trained counterparts.

Complexity of deep neural networks explains their V1 encoding performance

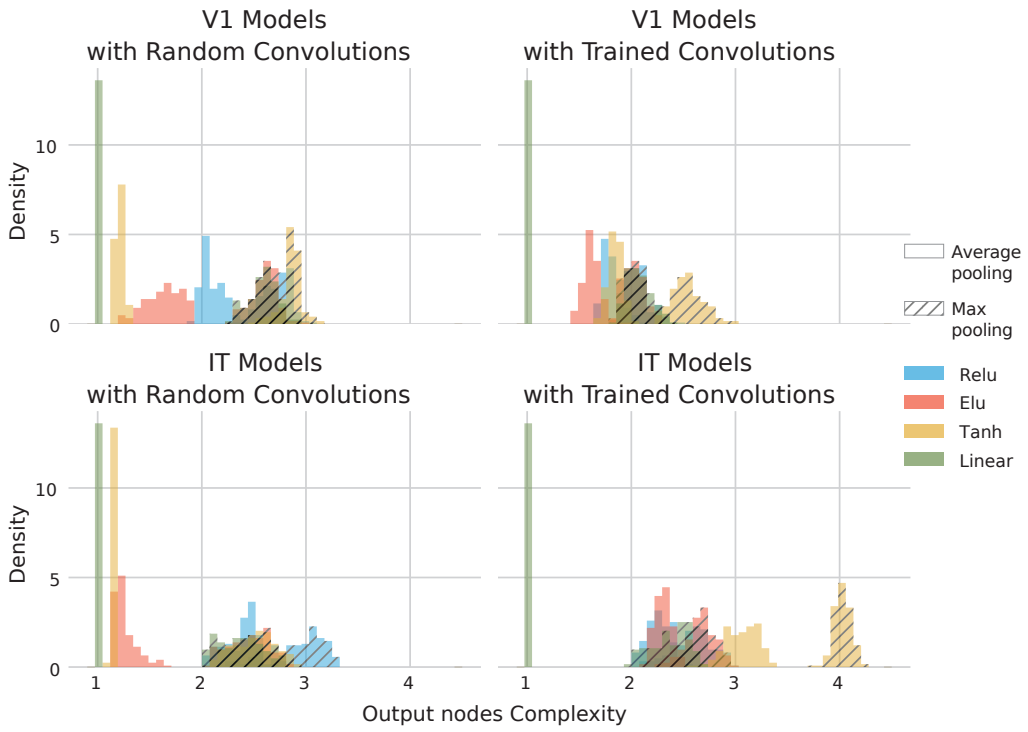


Figure 3.3. Distributions of encoding models' complexity across neurons. Each distribution across neurons represent the complexity (see methods) of a certain model configuration (with respect to the pooling strategy and activation function) trained on V1 (upper row) and IT (lower row) data. Only linear readout was trained on top of random convolutional features (left column) or the model was fully trained (right column). Each distribution is the average of 5 training iterations. The corresponding distributions for the human fMRI models are in supplementary Figure S3.2.

We showed that random-weight CNNs varied substantially in their encoding performance of brain responses as a function of the non-linear activation function and pooling strategies, especially for early visual cortex (Fig. 3.2b). However, after having trained the CNNs to predict neural activity, they showed compara-

ble performance across the nonlinear activation functions. Recent studies have shown that random-weight neural networks represent functions of different complexities depending on their architectural components such as their activation functions (Chiang et al., 2022; Ramasinghe et al., 2022; Teney et al., 2024). We thus wondered whether complexity may explain the encoding performance of random-weight CNNs.

We approximated the function implemented by each output node in each model using a set of Chebyshev polynomials (see Methods). Complexity was then quantified as the average of the polynomial orders, weighted by their coefficients (Teney et al., 2024). Fig. 3.3 shows the complexity distribution of output nodes in our models of V1 and IT responses, with both random-weight and trained convolutional kernels (analyses of fMRI models, see supplementary Fig. S3.4).

The random-weight CNNs with a low V1 and IT encoding performance (linear, ELU, Tanh with average pooling) generally showed low complexities as compared to their trained counterparts. By contrast, the random-weight CNNs with high V1 and IT encoding performance, such as ReLU and models with max pooling, had higher complexities than the random-weight CNNs with a low encoding performance. Finally, there was a greater overlap among the complexity distributions of models with trained convolutions than those with random convolutions, which is a consequence of the fact that the networks were optimized for the same target function. That is, training CNNs with a different architecture makes their complexity more homogeneous. For V1, random-weight CNNs with a comparable complexity to the trained counterparts also have similar encoding performance, suggesting that model complexity is a main driver of encoding performance. By contrast, for IT, there are major differences in encoding performance between random-weight and trained models despite similar complexity, suggesting that the specific configuration of weights is an additional important factor for IT.

We explored the relationship between model complexities and their encoding performance by plotting the median of the complexity distribution of each model's output nodes against its encoding performance (Fig. 3.4). To quantify the relationship between complexity and encoding performance, we fitted a quadratic function. We found a systematic relationship between the median complexity of the models and their encoding performance for V1 in both humans and macaques (Explained variance was 86%, 81%, and 84% for the areas V1, V1v, and V1d respectively). For higher visual areas, the relationship was substantially weaker (Explained variance was 63%, 57%, and 55% for the areas IT, VO1, and VO2 respectively). Specifically, while random-weight models with similar model complexity as trained counterparts had comparable encoding performance for V1, there was

major increase in encoding performance for IT. Together, these results suggests that for V1, complexity alone explains encoding performance, while for higher visual areas, the precise configurations of connection weights discovered through gradient descent are crucial for strong encoding performance. To further test the dependence of encoding performance on the precise configuration of weights, we shuffled the weights of all convolutional kernels of all layers of the trained models across all dimensions (space, input channels and output channels), freezing the weights, and subsequently retraining the linear readout. For a fair comparison, the deeper models (4 convolutional layers) were used for both early and higher visual areas. As anticipated, a much stronger decrease in encoding performance was observed for the shuffled models in higher visual areas compared to area V1 for both macaque (Fig. 3.6a) and human (Fig. 3.6b) brains.

Precise configuration of convolutional weights is critical for object recognition but not texture discrimination

Next, we investigated what kind of visual computations / tasks can be performed by random-weight CNNs, and which tasks are strongly dependent on training. To this end, we created a Texture-MNIST dataset for which two different tasks can be defined (Fig. 3.5b). Texture-MNIST is a dataset in which every sample is an MNIST digit filled with a texture batch (see Methods). Texture batches are randomly sampled from 10 high-quality texture images. We trained the 4-layer models to predict either the object (digit) identity or the texture patch identity. Similar to the neural data, we either trained only the readout, leaving the convolutional layers frozen at their randomly initialized state, or we trained both the readout and the convolutional layers. We observed that random-weight ReLU networks, with either average or maximum pooling, outperformed all other random-weight networks in predicting the correct identity of the texture class and the digit class (Fig. 3.5a). However, random-weight ReLU networks achieved almost the same performance as the trained networks on the texture discrimination task, while there was a major difference in performance for digit recognition task.

Similar to the neural data, we also investigated the dependence of task performance on the complexity of the models. We found that the complexity of the models showed a very strong relationship with texture discrimination accuracy (explained variance 95%) but not for digit recognition accuracy (explained variance 32%). These findings demonstrate that object recognition performance requires a precise optimization of convolutional kernels, while texture discrimination can already be subserved by random-weight neural networks. We further confirmed this observation by shuffling the weights of the optimized convolutional kernels

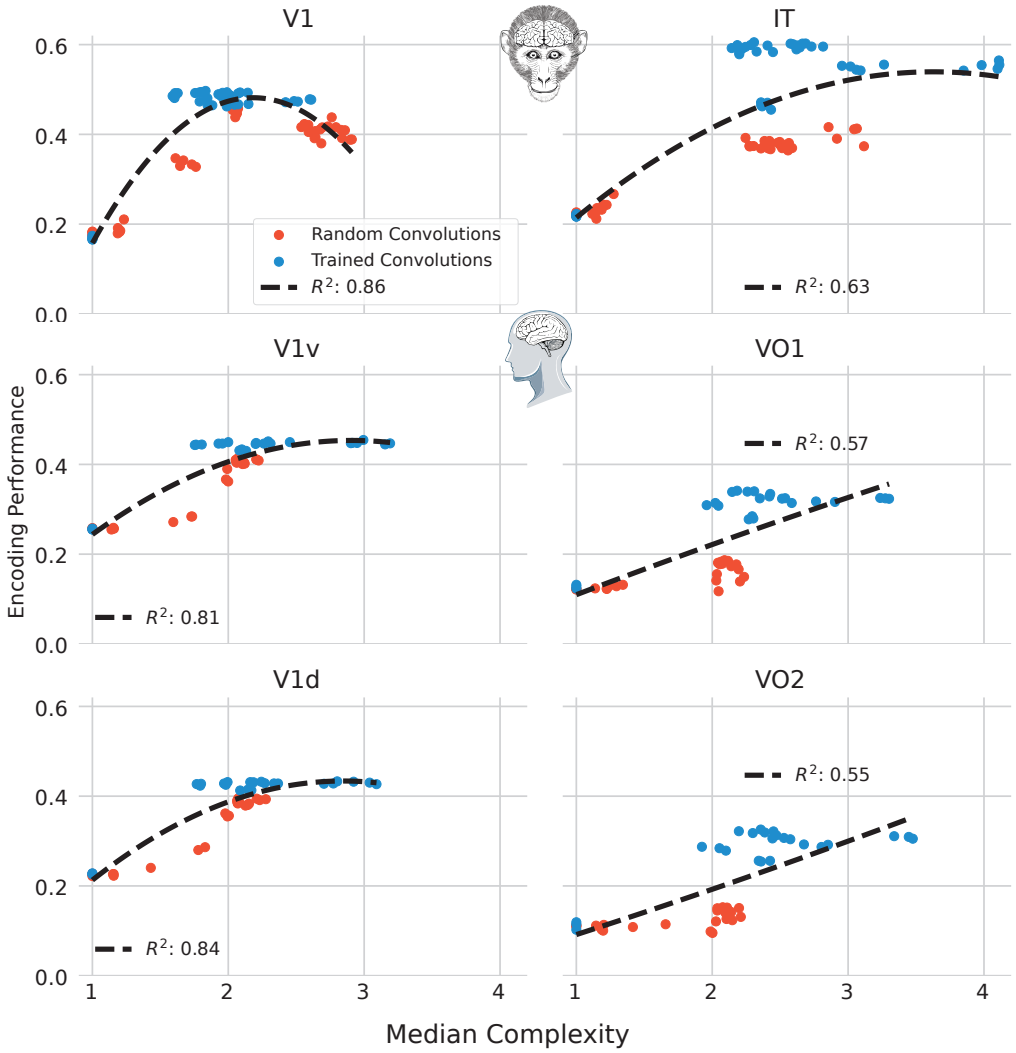


Figure 3.4. The relationship between the median complexity of the models and their encoding performance. Each panel shows the relationship between the median complexity of the models and their encoding performance for a certain brain area in macaques and humans. Models vary in their activation functions (4), pooling strategy (2) and wether only the readout was trained (in red) or the full model was trained (in blue). Each model configuration is represented 5 times that vary in their random initializations. The dashed black curve represent the fitted quadratic function to the data whose goodness of fit is quantified through R^2 printed in the legend.

then retraining the readout. Object recognition performance showed a stronger decrease than the texture discrimination performance (Fig. 3.6c).

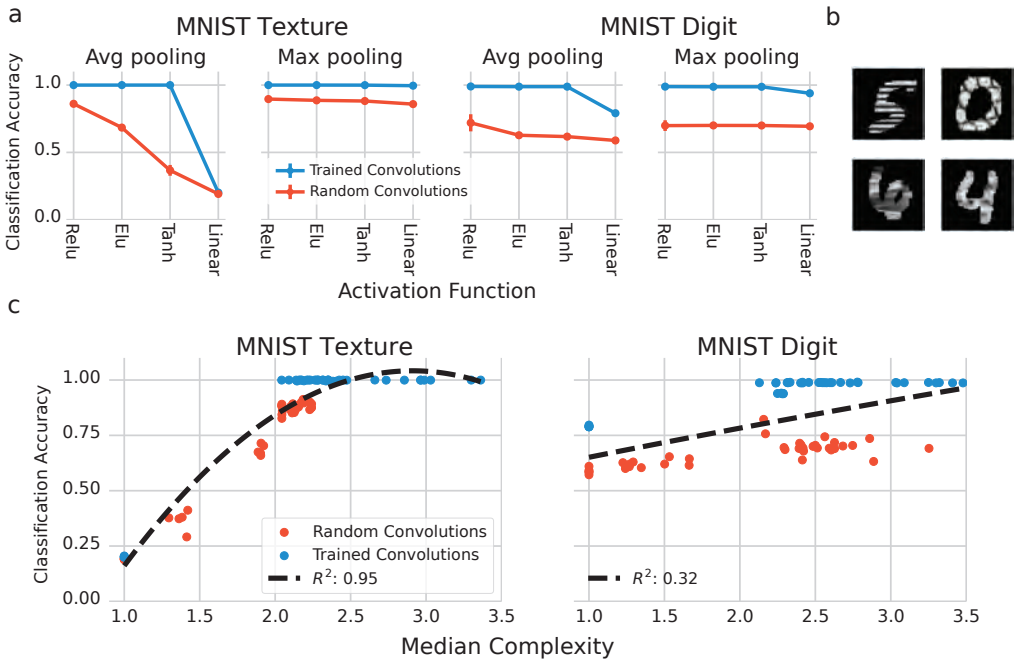


Figure 3.5. Texture discrimination and digit recognition performance on the Texture-MNIST dataset. (a) Performance of our 4-layer models trained on the Texture-MNIST dataset to either predict the texture identity (MNIST Texture) or the digit identity (MNIST Digit). Only the linear classification layer was trained (in red) or the full convolutional model was trained (in blue). Each line is the average of 5 training iterations. **(b)** Examples of the Texture-MNIST dataset. **(c)** The relationship between the median complexity of the mdoels and their texture discrimination accuracy (left) and digit recogntion accuracy (right).

Trained networks develop similar orientation selectivity to V1

We demonstrated that the representations of random-weight CNNs, with an architectural bias that entails sufficient model complexity, suffice for encoding the responses of early visual cortex in macaques and humans. However, it is well-established that V1 neurons exhibit selectivity for certain features, such as the orientation of a bar or grating stimulus (Hubel, Wiesel, et al., 1959). We sought to determine whether random-weight CNNs also possess such feature tuning. To test this, we generated Gabor patches of varying orientations and phases and presented them to the neural networks to assess their orientation selectivity, and then compared the selectivity distributions to experimental V1 data (Fig. 3.7a). We analyzed the central neurons in the last convolutional layer (i.e. those with a receptive field in the center of the image) of the random-weight neural networks

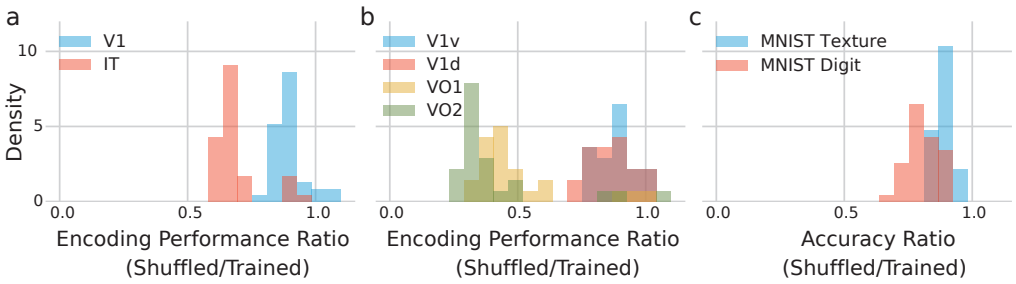


Figure 3.6. Effect of shuffling the convolutional weights on the neural encoding performance and classification accuracy. For each fully trained model, we shuffled the convolutional weights in the model and retrained the linear readout to predict the neural responses and to classify the texture/digit identity. Then we calculated the ratio between the performance of the model after shuffling and its performance before shuffling. **(a)** The distribution of the encoding performance ratio for all models across different configurations for the macaque brain areas. **(b)** same as **(a)** but for the fMRI human brain areas. **(c)** same as **(a)** but for the models trained for texture discrimination and digit recognition.

and the networks trained to predict V1 activity. To compute the activation for each orientation, we averaged across all the different phases of the Gabor stimuli. In Fig. 3.7b, we show the four most orientation-selective neurons in each trained model (for random models, see supplementary Fig. S3.5a). We then quantified the neurons' orientation selectivity by calculating the circular variance of their tuning curves. Circular variance is a measure used to quantify the dispersion of data points around a circle, with a lower value indicating more concentrated responses, and thus, higher selectivity (Mazurek et al., 2014) (see Methods). We compared the circular variance distribution of the artificial neurons in the models (Fig. 3.7c for trained models and supplementary Fig. S3.5b for random-weight models) with the circular variance distribution of V1 neurons recorded from alert macaque monkeys (Gur et al., 2005). To this end, we quantified the difference between distributions using the Wasserstein (i.e. Earth Mover) distance, which we term the "V1 deviation score" (lower scores indicate more similarity). Random-weight ReLU networks exhibited the lowest median circular variance among random-weight models (Fig. 3.7d), i.e. they were the most orientation-selective. Moreover, random-weight ReLU models also demonstrated the lowest V1 deviation score among all random-weight networks (Fig. 3.7e). Furthermore, training the models on V1 data led to stronger orientation selectivity (i.e. a lower median circular variance) for all the models except the linear ones (Fig. 3.7d). Moreover, training the models on V1 data also led to lower V1 deviation scores for all of the models except the linear ones (Fig. 3.7e), with trained ReLU models being the most similar to the V1 orientation selectivity distribution. To examine whether the V1 deviation

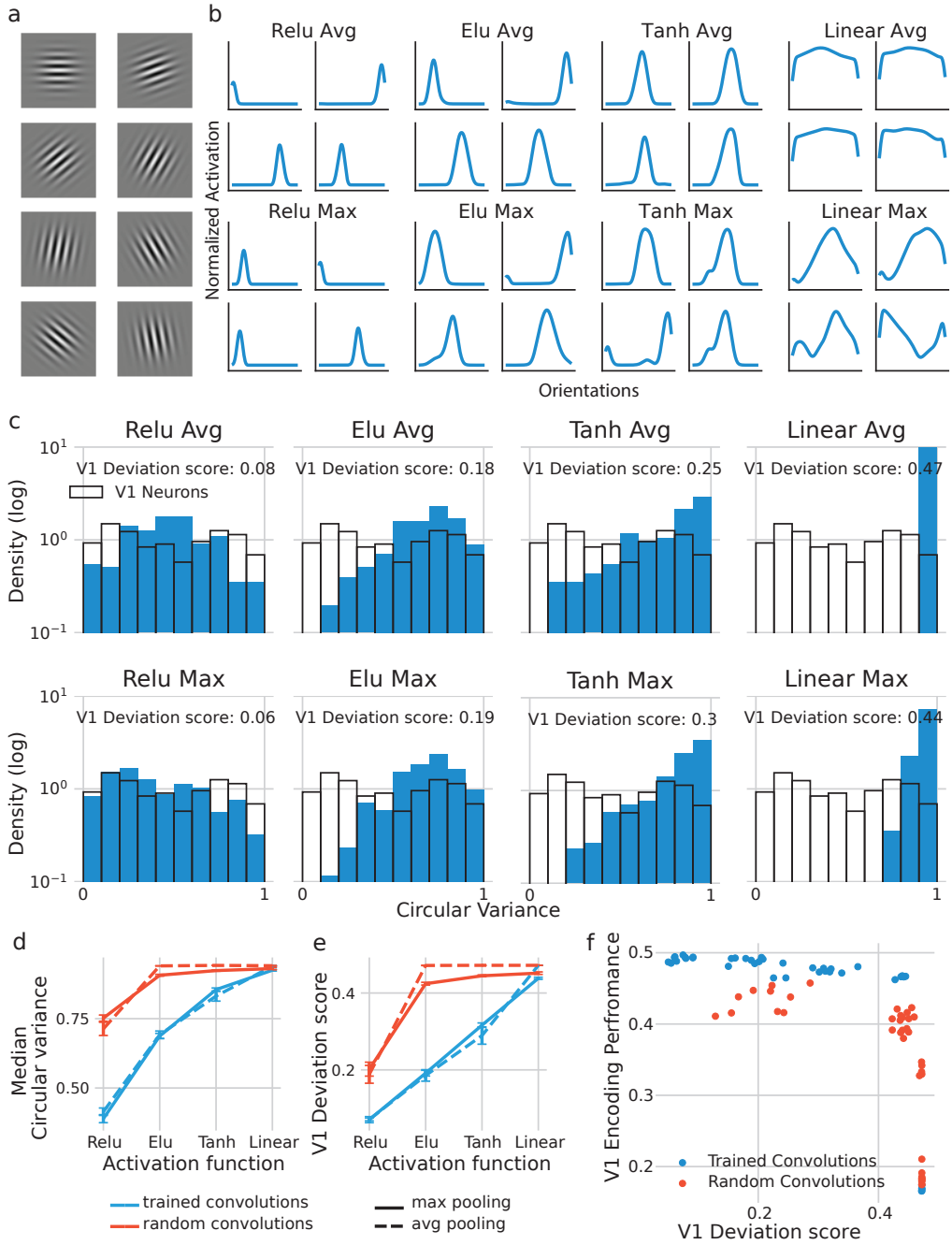


Figure 3.7. See next page

scores decreased because the networks were specifically trained to predict V1, we also tested the orientation selectivity of the models trained on IT data. We found that IT trained models were less orientation-selective than V1 trained models

Figure 3.7. Orientation selectivity of random and trained models (a) Examples of the Gabor patches that were represented to the models. **(b)** Tuning curves of the most orientation selective artificial neuron in the last convolutional layer of each model configuration trained on the V1 data. For the randomly initialized models see supplementary figure S3.2a. **(c)** Distributions of the circular variance of the artificial neurons of the last convolutional layer of each model configuration trained on the V1 data. For the randomly initialized models see supplementary figure S3.2b. **(d)** Median circular variance of the artificial neurons of the last convolutional layer of random and V1 trained models of different configurations averaged across 5 iterations. Error bars are the standard deviation. **(e)** V1 deviation scores of the random and V1 trained models averaged across 5 iterations. Error bars are the standard deviation. **(f)** The relationship between V1 deviation score and the V1 encoding performance for all the models that varied in their activation functions, pooling strategies, and their random initializations. Spearman correlation $= -0.87$, $p \ll 0.001$.

(supplementary Fig. S3.6a), and had higher V1 deviation scores (supplementary Fig. S3.6b).

To examine if networks with more similar orientation tuning to V1 also better predicted V1 activity (on another dataset), we examined the relationship between both variables across random and trained models (Fig. 3.7f). We found a strong negative monotonic relationship (Spearman correlation = -0.87) between the V1 deviation score and V1 encoding performance, suggesting that the more similar the orientation selectivity of the models is to that of V1 neurons, the better the models are at predicting V1 responses (on another dataset). However, it is also worth noting that although trained networks demonstrated small variation in their V1 encoding performance, they showed high variability in their V1 deviation score (Fig. 3.7f), suggesting that having similar orientation selecting to V1 neurons is not a strong requirement for the model to be able to effectively encode V1 responses.

Discussion

We investigated the factors contributing to the success of CNNs in predicting responses within the visual cortex in both macaques and humans. We demonstrate that, unlike higher visual areas, accurate prediction of early visual cortex responses does not necessarily depend on the optimization of convolutional kernels via training. Instead, architectural components, such as pooling strategies and activation functions, play a crucial role. Specifically, we found that the ReLU activation function and maximum pooling were the most critical components for achieving high encoding performance for early visual cortex, even in the

absence of optimization of the convolutional kernels based on a task or neural data. Furthermore, we observed that random-weight CNNs exhibited substantial variability in the complexity of the functions they represented, depending on their architectural components. Notably, the CNN's complexity explained significantly more variance in the encoding performance in early visual cortex compared to higher visual areas. These findings held true across electrophysiological data from macaques and fMRI data from humans. Additionally, model complexity explained significantly more variance in performance on a texture discrimination task than on a digit (object) recognition task using the same dataset. This suggests that precise configuration of convolutional kernels is more essential for object recognition and, consequently, more critical for predicting responses in higher visual areas. Importantly, our results indicate that training the full convolutional models masked the effect of architectural bias, as the variability in encoding performance across different model configurations decreased significantly after training. However, when employing an alternative metric for aligning models with experimental V1 data – namely, the circular variance distribution of the model's artificial neurons as a proxy for their orientation selectivity – a different picture emerged: Trained models exhibited significant variability in their alignment with V1 orientation selectivity, despite displaying low variability in their V1 encoding performance. Overall, our findings indicate that random-weight CNNs with sufficient model complexity allow for comparable prediction of V1 activity as trained CNNs, while higher visual areas require precise weight configurations acquired through training via gradient descent.

Our work highlights the importance of random-weight controls to reveal architectural bias. According to the deep learning framework for neuroscience (Richards et al., 2019), various deep neural network models, trained under specific constraints regarding their architecture, objective function, and training data, can serve as tests for specific hypotheses about brain function (Cichy & Kaiser, 2019; Doerig et al., 2023; Dwivedi et al., 2021; Zhuang et al., 2021). The degree to which these models' representations of complex natural stimuli align with brain responses can provide evidence for certain hypotheses, such as the necessity of recurrent connections for cortical information processing (Kietzmann et al., 2019; Kubilius et al., 2019). However, it has been shown that multiple models with different architectures, trained similarly, or multiple models with the same architecture, trained differently, can achieve similar performance in predicting neural data (Conwell et al., 2024; Storrs et al., 2021). Further investigations revealed that task optimization increased the effective dimensionality (Del Giudice, 2021) of CNNs representations. This increase in dimensionality correlated with their encoding performance of higher visual brain areas (Elmoznino & Bonner, 2024). A recent

study showed that substantially scaling the dimensionality of random CNNs, but not transformers or fully connected networks, led to a comparable encoding performance of higher visual areas to ImageNet-trained AlexNet model suggesting a strong contribution of architecture bias to DNN encoding performance of neural data (Kazemian et al., 2024). In the present work we showed that scaling the complexity of random CNNs, in the low-dimensional regime, leads to a comparable encoding performance for lower, but not higher visual areas. Moreover, it has been shown that layers of task-optimized VGG19 and Alexnet models with matching receptive field sizes yielded similar V1 encoding performance despite having different depths i.e. different numbers of nonlinear transformations suggesting the significance of receptive field size in predicting V1 data (Miao & Tong, 2024). These findings underscore the necessity of implementing appropriate controls when employing task-optimized or brain-optimized DNNs for predicting neural activity. Specifically, the use of random-weight models with variable architectural components is crucial to reveal the architectural biases of neural encoding models.

Beyond the hypothesized neural factors influencing the neural encoding performance of DNNs, their success in predicting brain data could originate from non-neural, biologically implausible design choices implemented by researchers. For instance, the emergence of grid-like representations in DNNs optimized for path integration has been shown to depend on specifically designed readout mechanisms (Schaeffer et al., 2022). Moreover, numerous studies have demonstrated that the measured similarity between models and the brain can be highly dependent on the chosen similarity metric (Davari et al., 2022; Soni et al., 2024). Specifically, neural predictability scores based on linear regression can be heavily influenced by the inductive biases of linear regression, the dimensionality of the model representations, or the ratio of the number of stimuli in the benchmarking dataset to the dimensionality of the model representations (Bowers et al., 2023; Canatar et al., 2024; Elmoznino & Bonner, 2024; Schaeffer et al., 2024). When we assessed the brain alignment of V1 models using two independent scores: encoding performance on one dataset and V1 deviation score on another dataset, we found large difference in the variability of both scores despite correlating with each other. Therefore, consistent with the existing literature, our findings highlight the importance of moving beyond single metrics of model-brain alignment. Instead, our findings highlight the importance of a multidimensional model assessment approach that enables the dissection of the similarities and differences between computational models and the brain (Biscione et al., 2024; Jacob et al., 2021; Rajesh et al., 2024; Wichmann & Geirhos, 2023). This holds true in particular considering that many of these similarity measures are correlational (Bowers et al.,

2023), and models with high prediction scores can still operate in qualitatively different ways than the brain (Baker et al., 2018; Farahat et al., 2023; Geirhos et al., 2018, 2019; Wichmann & Geirhos, 2023). This multidimensional approach will facilitate more targeted model improvement and informed hypothesis generation in future research.

By systematically manipulating the architectural components of the models and comparing the performance of randomly initialized models with their fully trained counterparts, we identified the essential components that underpin neural encoding, raising the question to what extent they mimic the architecture of visual cortex. The ReLU activation function emerged as a key factor in generating visual representations that supported the most efficient encoding performance, considering the number of trainable parameters. The ReLU activation function was introduced to DNNs as a more biologically plausible alternative to Sigmoid and Tanh functions, given its one-sided nature (outputting zero for negative inputs) and its promotion of sparse representations (Attwell & Laughlin, 2001; Douglas et al., 1995; Glorot et al., 2011). Indeed, ReLU networks, even with random convolutions, not only exhibited the best encoding performance for V1 but also displayed the smallest distance to the orientation selectivity distribution of V1 neurons. Importantly, while models fully fitted to predict V1 data, with ReLU, ELU, or Tanh activation functions, exhibited similar encoding performance, they still displayed substantial variability in their similarity to V1 orientation selectivity, with ReLU networks being the most V1-like. This result demonstrates that the combined application of multiple model assessment metrics and systematic architectural manipulations enables the identification of key, potentially biologically plausible architectural components that contribute significantly to neural encoding performance. Furthermore, the fact that non-linearities with sufficient model complexity are a major factor in predicting neural activity fits with the general idea that non-linearities are a central component of cortical inter-areal interactions beyond mere linear information transmission (DiCarlo et al., 2012; Vinck et al., 2023).

Beyond examining the encoding performance and tuning properties of the models' representations, it is imperative to understand the computational advantages of models' representations in supporting visual tasks. We demonstrated that random-weight CNN representations, which were sufficient for predicting early visual cortex responses, performed well in discriminating between texture families compared to their trained counterparts. Conversely, these random-weight representations were considerably worse than fully trained models in invariantly classifying the identity of digits within images. Studies have shown that V1 activity, in particular superficial cortex, exhibits selectivity for texture statistics, albeit less

pronounced than in a higher visual area, LM (Bolaños et al., 2024; Ziemba et al., 2019). In humans, texture discrimination task-learning has been shown to induce local changes within the early visual cortex without requiring the recruitment of higher visual areas (Schwartz et al., 2002). Additionally, a decoder trained on macaque V1 population activity elicited by texture samples could discriminate between 15 different texture families (Ziemba et al., 2016). In contrast, several studies have demonstrated that IT neurons possess the tolerance to identity-preserving transformations that is essential for object recognition (Hung et al., 2005; Rust & DiCarlo, 2010).

The efficacy of random features in machine learning has been well-documented, often rivaling hand-crafted or even learned features across various learning tasks (Gallicchio & Scardapane, 2020; Rahimi & Recht, 2008a; Scardapane & Wang, 2017). Random-weight CNNs were shown to be frequency-selective and translation-invariant which explains their superior performance over random non-convolutional networks on image classification tasks (Saxe et al., 2011). Moreover, only training a small fraction of the convolutional weights or only training the batch normalization layers in random-weight CNNs led to object recognition performance competitive with their trained counterparts (Frankle et al., 2021; Rosenfeld & Tsotsos, 2019). Furthermore, the structure of a random generator CNN can capture significant low-level image statistics even without any learning. This inherent structure acts as a prior, making random-weight CNNs useful for various image processing tasks such as image restoration, denoising, inpainting and super-resolution (Ulyanov et al., 2018). These findings emphasize the significant contribution of convolution and pooling operations, independent of learning, in visual processing tasks. Consequently, it is plausible that random features with the right convolutional architectural bias could effectively model the representations found in V1, considering the diverse range of visual tasks that V1 supports. One hypothesis is that V1 comprises an array of neurons representing high-dimensional, non-linear random basis functions, capable of supporting a diverse set of downstream functions (Rahimi & Recht, 2007, 2008a, 2008b). In addition to our results that showed that random-weight ReLU networks exhibit orientation-selective neurons, previous research on biologically plausible recurrent models of mouse V1 demonstrated the emergence of orientation selectivity even when both feedforward and recurrent connections are randomly initialized (Hansel & van Vreeswijk, 2012; Pattadkal et al., 2018).

Our findings contribute to the growing body of literature that emphasizes the importance of conducting controlled experiments to systematically investigate the architectural and training components that contribute to the neural encod-

ing performance of computational models. Moreover, our results underscore the necessity of developing comprehensive batteries of neural and perceptual metrics to facilitate more informed conclusions about the similarities between computational models and the brain (Biscione et al., 2024; Jacob et al., 2021). Finally, considering the computational benefits of the models' representations that support the prediction of brain responses is valuable, as it helps formulate hypotheses regarding the functional roles of different brain areas (Cichy & Kaiser, 2019; Dwivedi et al., 2021).

Methods

Datasets

V1 monkey dataset

We used a public dataset that consists of neural activity recordings from 166 neurons across different layers of V1 brain area in two monkeys (Cadena et al., 2019). The monkeys were shown 7,250 images, each presented 1-4 times for a duration of 60 milliseconds. Each image was displayed within a circular window spanning 2 degrees of visual angle, with the edges gradually fading out to blend with the surroundings.

IT monkey dataset

We used a publicly available IT monkey dataset which consists of neural recordings from 168 multiunit sites within the inferotemporal (IT) cortex of two macaque monkeys (Cadieu et al., 2014). The monkeys were presented with 3,200 unique grayscale images, each showing one of 64 objects from eight categories. These images were designed to mimic real-world visual scenes by placing the cropped object images onto various natural image backgrounds at different positions, orientations, and sizes.

fMRI human dataset

The Natural Scenes Dataset (NSD) is a publicly available fMRI dataset that captures the brain activity of eight human participants as they viewed thousands of natural images (9,000–10,000 distinct color natural images for each subject repeated up to 3 trials) (Allen et al., 2022). The images were taken from the Microsoft Common Objects in Context (COCO) database square-cropped and presented at a size of 8.4° x 8.4°. We used the regions of interest (ROI) V1v and V1d manually drawn

based on the results of a population receptive field (pRF) experiment, and the higher-order ROIs VO1 and VO2, defined using a visual probabilistic atlas.

Texture-MNIST dataset

We created the Texture-MNIST dataset to probe different models' texture and shape discrimination abilities. We created binary masks from the MNIST dataset, resized and overlaid them over 64×64 patches of texture randomly copped from a high-resolution texture image unique for each digit class. Using this dataset, we can train our models to either predict the class of the object (digit) or the class of the texture of each image.

Models

We used simple DNN models consisting of a convolutional block and a linear readout. The convolutional block included two and four convolutional layers for the shallow and deeper models respectively. Each convolutional layer is followed by a batch normalization layer and an activation function. To maintain an efficient number of trainable parameters we used depthwise separable convolutions in all convolutional layers except for the first one (Du et al., 2024). Furthermore, shallow models had 16 and 256 feature maps in their 2 convolutional layers, whereas deeper models had feature maps that progressively increased from 16 to 32, 64, and finally 256 across the network depth. For the shallow models, we had a pooling layer after each activation function and for the deeper models, we had the pooling layer after every other activation layer. For the shallow models, convolutional layers had either filter size of 9×9 or 17×17 pixels, leading to an effective receptive field of the models of 28×28 or 52×52 respectively. For the deeper models, we had filter sizes of 5×5 or 9×9 pixels leading to the same effective receptive fields as the sallow models (see supplementary Fig. S3.1 for an illustration of the detailed architecture of the shallow and deep 28×28 models). We tested a variety of activation functions including ReLU, ELU, Tanh, and Linear. Moreover, we considered average and maximum pooling operations, each with a pooling window of 2×2 pixels.

The three-dimensional activation maps of the convolutional block were transformed to the neural responses through a linear readout factorized using three one-dimensional weight vectors $\mathbf{w}_c$, $\mathbf{w}_x$, and $\mathbf{w}_y$ for the channels, and two spatial dimensions respectively. For the image classification tasks, global average pooling was applied to the three-dimensional activation maps to obtain the feature vector used for classification.

Complexity measurement

To calculate the complexity of the function represented by a neural network, we evaluate the network on a regularly sampled grid in its input space (Teney et al., 2024). Our networks were trained with input normalized to the range from -1 to 1. Therefore, we sampled 100 corners in the hypercube $[-1, 1]^d$, where d is the input dimension of the network. We sampled 50 points regularly on each of the 100 lines connecting each corner with its succeeding corner and evaluated the network at each sample input point.

Let x be the regularly sampled line in the range $[-1, 1]$, and y be the activation of a certain output node evaluated at the data points lying on that line in the input space hypercube. We compute the coefficients **c** of Chebyshev polynomials that fit the data (x, y) by minimizing the least square error:

$$E = \sum_{i=1}^{n} \left(y_i - \sum_{k=0}^{d} c_k T_k(x_i) \right)^2$$

where k is the polynomial order and $T_k(x_i)$ is the $k-$th Chebyshev polynomial evaluated at the i-th input x_i. The Chebyshev polynomials $T_k(x)$ are recursively defined as:

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$$

We define the complexity metric as the average of the polynomial orders weighted by their corresponding Chebyshev coefficients:

$$\text{Complexity}_{\textit{Chebychev}} = \frac{\sum_{k=0}^{d} c_k k}{\sum_{k=0}^{d} c_k}$$

For each output node, we average the 100 complexity metrics from the 100 lines to obtain one complexity measurement per one output node.

Comparison to experimental V1 data

To go beyond regression, we tested the alignment of the artificial neurons in our models with experimental data recorded from the V1 area by comparing their orientation selectivity distributions. We presented the models with Gabor patches of different orientations and phases (Kong et al., 2022). Orientations were sampled at $10°$ steps in the range from $0°$ to $180°$ and 20 phases were sampled evenly from $0°$ to $360°$. The spatial frequency of the Gabor patches was chosen to allow the

receptive field of the model neurons to contain two cycles i.e. for the receptive field of 28×28 pixels and input size of 128×128 pixels, we used a spatial frequency of 9.142 cycles/image. Specifically, we generated the Gabor patches according to the formula:

$$f(x, y; \sigma, \lambda, \psi, \theta, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(\frac{2\pi x'}{\lambda} + \psi\right)$$

and

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

where $\sigma = 20$ pixels is the standard deviation of the Gaussian envelope which controls the size of the Gabor patch, λ is the wavelength of the sinusoidal factor in pixels, ψ is the phase offset of the sinusoidal factor, θ is the orientation of the Gabor patch in radians and $\gamma = 1$ is the spatial aspect ratio, specifying the ellipticity of the support of the Gabor function.

We calculated the orientation tuning curves of the center location of every artificial neuron at the last convolutional layer of our models. The tuning curves were scaled to the range from 0 to 1 to avoid negative responses in activation functions such as Tanh, Elu, and Linear. To establish a baseline for artificial neuron responses, the minimum response value across the V1 test set was determined. This minimum value was then used in the scaling of the tuning curves. Then we followed the analysis steps mentioned in (Gur et al., 2005) to calculate the orientation selectivity distribution of the V1 neurons recorded from awake monkeys. Briefly, we linearly interpolated the tuning curves with 1° steps, then smoothed them with a Hanning filter with a 7° half-width at half-height. We then quantified the orientation selectivity of neurons from the tuning curves by calculating the circular variance (CV) (Mazurek et al., 2014). The circular variance was calculated from the smoothed tuning curves resampled at regular 15° intervals according to the equation:

$$CV = 1 - \frac{|\sum_k r_k e^{i2\theta_k}|}{\sum_k r_k}$$

where θ_k is the orientation in radians and r_k is the corresponding response.

For each model, we simulated 100 in-silico electrophysiology experiments by randomly sampling with replacement 339 neurons from its last convolutional layer. We calculated their CV as described. From each experiment, we compared the distribution of CV with the corresponding distribution obtained from 339 V1 neurons recorded from alert macaques (Figure 3 in (Gur et al., 2005)) by calculating a V1 deviation score and then averaging the scores over the 100 experiments to obtain

one score per model. The V1 deviation score was computed as the Wasserstein distance between the distribution of circular variance of the model's neurons and the corresponding distribution of experimental V1 neurons.

Acknowledgments

This project was financed by the BMF (Bundesministerium fuer Bildung und Forschung), Computational Life Sciences, project BINDA (031L0167); an ERC starting grant (850861) SPATEMP; DFG VI Grants (908/5-1 and 908/7-1); an NWO VIDI Grant; and the Dutch Brain Interface Initiative (DBI2).

Supplementary Figures

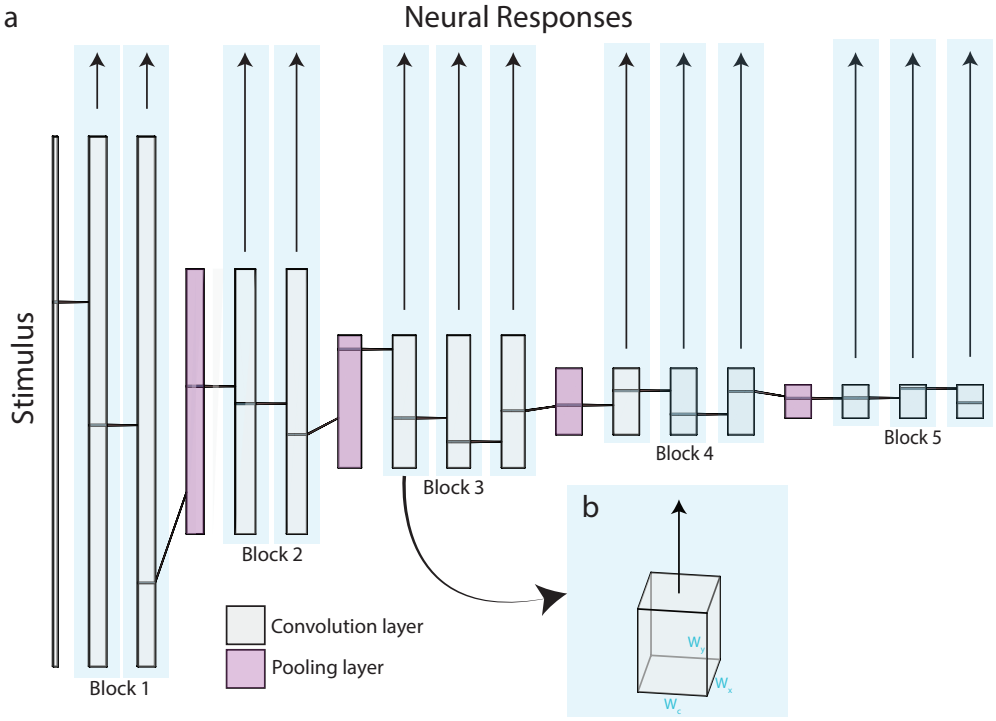


Figure S3.1. VGG16 encoding models. (a) On top of the three-dimensional feature maps of each convolutional layer (gray rectangles), we trained a linear readout (arrows) to predict the neural responses of a certain brain area. We used ImageNet-trained VGG16 model and randomly-initialized variants. **(b)** Linear readout was factorized into 3 one-dimensional weight vectors $\mathbf{w}_c$, $\mathbf{w}_x$, and $\mathbf{w}_y$ for the channels and the two spatial dimensions respectively.

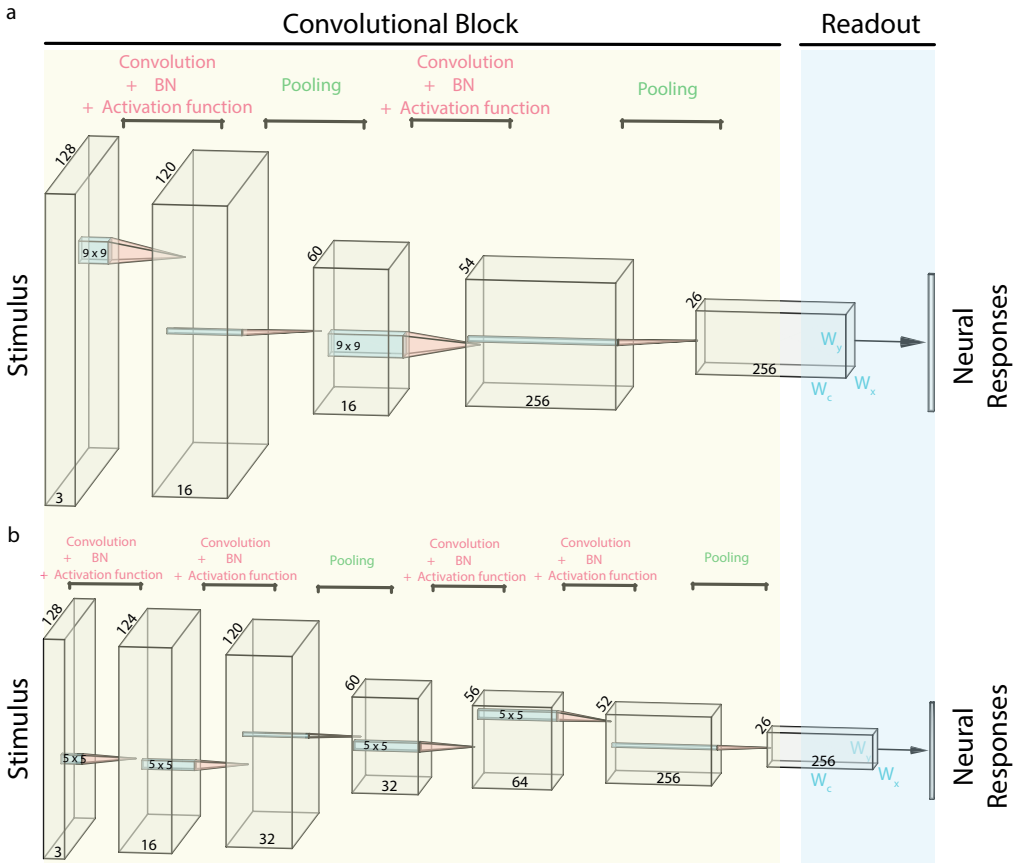


Figure S3.2. Architecture of CNNs used for encoding neural responses. All models are formed of a convolutional block and a linear readout. Readout is factorized into three one-dimensional vectors ($\mathbf{w}_c$, $\mathbf{w}_x$, and $\mathbf{w}_y$) that transform the channels and the two spatial dimensions of the feature maps of the last convolutional layer into neural activity. **(a)** The convolutional block of shallow models used for encoding early visual cortex activity is formed of two convolutional layers with 9×9 filter sizes. Each layer is followed by a batch normalization (BN) layer, an activation function (ReLU, ELU, Tanh, or linear), and a pooling layer of 2×2 window size and stride = 2 (maximum or average pooling). **(b)** The convolutional block of the deeper models used for encoding higher visual areas is formed of 4 convolutional layers with 5×5 filter sizes. Each layer is followed by a batch normalization (BN) layer and an activation function (ReLU, ELU, Tanh, or linear). Every other layer is followed by a pooling layer of 2×2 window size and stride = 2 (maximum or average pooling). The spatial resolution of feature maps is printed on top of each block of feature maps. Number of feature maps (channels) is printed at the bottom. All convolutional layers are depthwise separable except for the first one. The effective receptive field of neurons in the feature maps of the last convolutional layer of both models is 28×28 pixels.

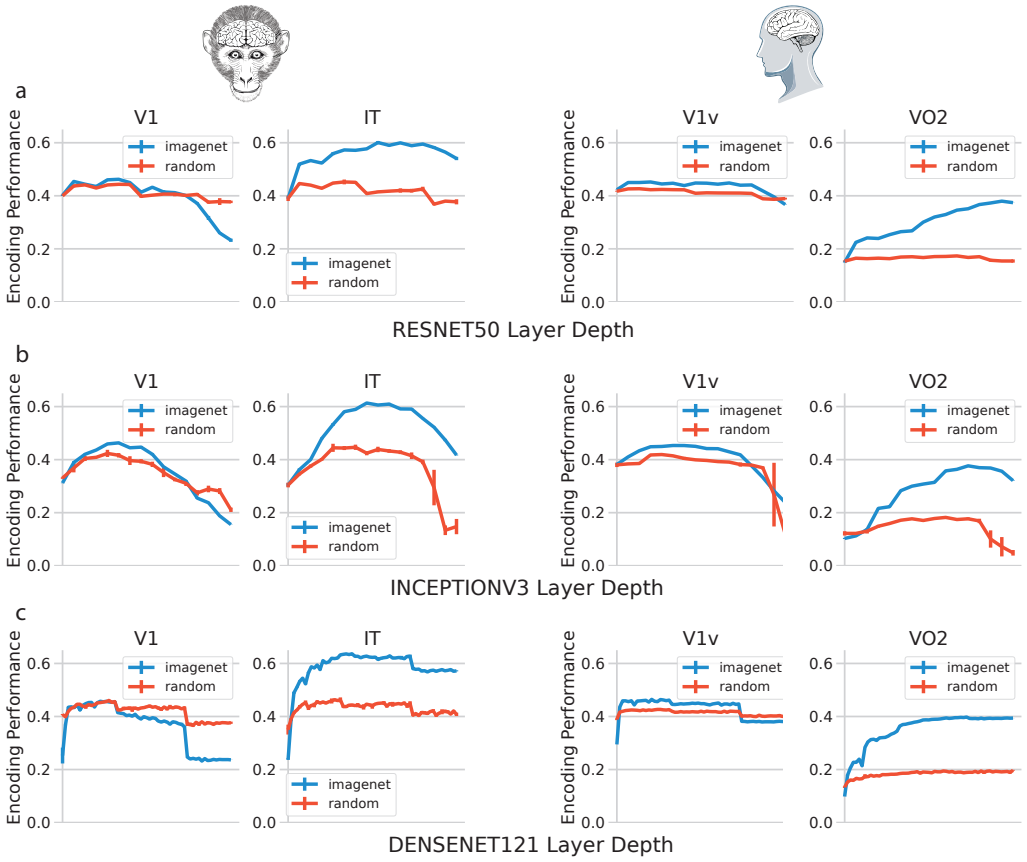


Figure S3.3. Encoding performance of popular convolutional architectures. **(a)** Encoding performance of a linear readout optimized on top of the representations of the convolutional layers of RESNET50 model either upon random initialization (in red) or pretraiend on ImageNet dataset for object recognition (in blue) for four neural datasets: Two electrophysiological datasets recorded from macaques: the early visual cortex V1 and the higher visual area IT and two fMRI datasets recorded from humans: the early visual cortex V1v and the higher visual area VO2. **(b)** same as **a** but for INCEPTIONV3 model. **(c)** same as **a** but for DENSENET121 model.

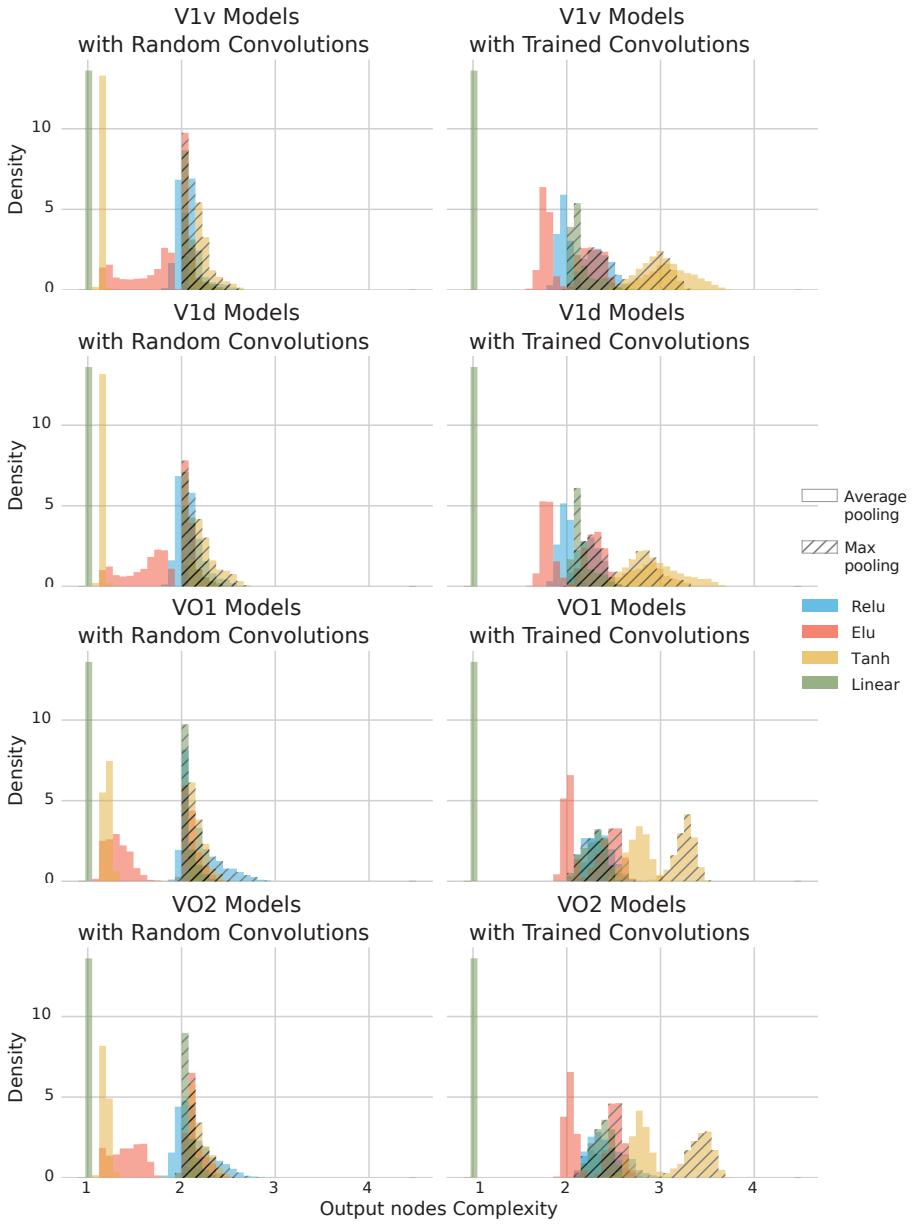


Figure S3.4. Distributions of encoding models' output nodes complexity. Each distribution represent the complexity of output nodes (see methods) of a certain model configuration (with respect to the pooling strategy and activation function) trained on human V1v, V1d (upper two rows), VO1 and VO2 (lower two rows) data. Only linear readout was trained on top of random convolutional features (left column) or the model was fully trained (right column). Each distribution is the average of 3 training iterations.

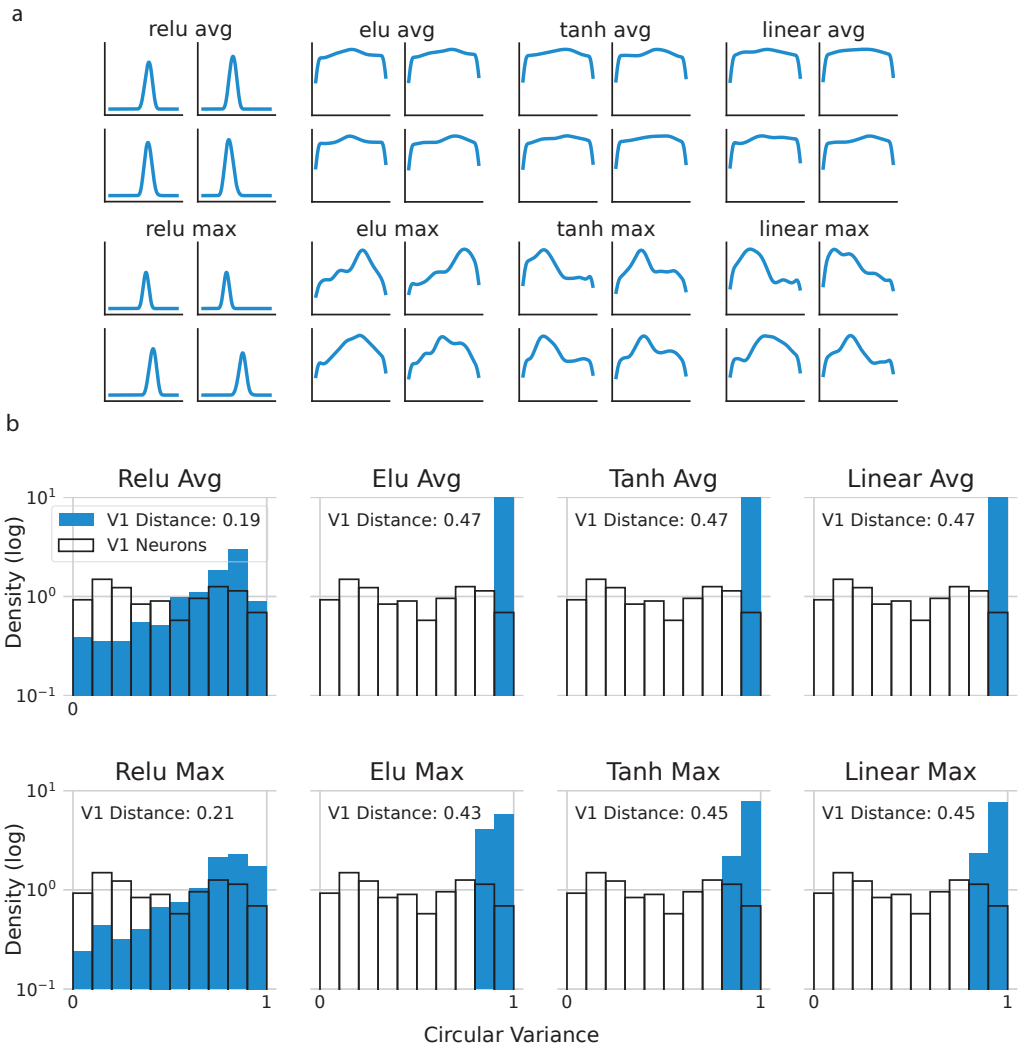


Figure S3.5. Orientation selectivity of random models. (a) Tuning curves of the most orientation-selective artificial neuron in the last convolutional layer of each model configuration upon random initialization. **(b)** Distributions of the circular variance of the artificial neurons of the last convolutional layer of each model configuration upon random initialization.

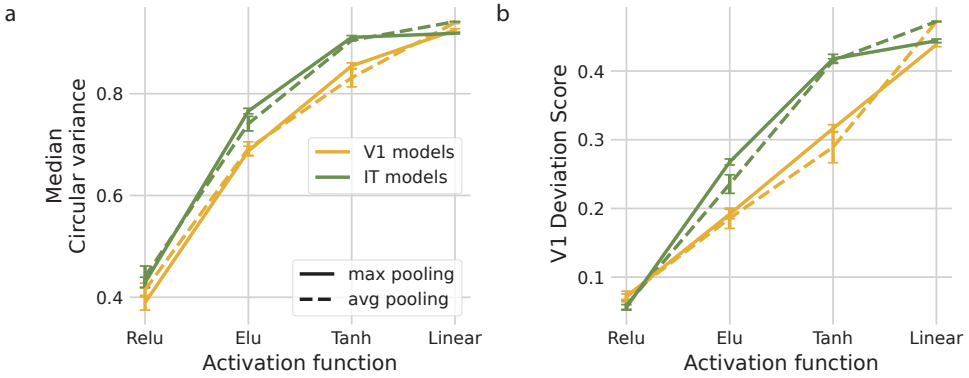


Figure S3.6. (a) Median circular variance of the artificial neurons of the last convolutional layer of V1 trained and IT trained models of different configurations averaged across 5 iterations. Error bars are the standard deviation. **(b)** V1 deviation scores of V1 trained and IT trained models averaged across 5 iterations. Error bars are the standard deviation.

Chapter 4

Diagnosing Epileptogenesis with Deep Anomaly Detection

Published as: **Farahat, Amr**, Diyuan Lu, Sebastian Bauer, Valentin Neubert, Lara Sophie Costard, Felix Rosenow, and Jochen Triesch. "Diagnosing epileptogenesis with deep anomaly detection." In Machine Learning for Healthcare Conference, pp. 325-342. PMLR, 2022.

Abstract

We propose a general framework for diagnosing brain disorders from Electroencephalography (EEG) recordings, in which a generative model is trained with EEG data from normal healthy brain states to subsequently detect any systematic deviations from these signals. We apply this framework to the early diagnosis of latent epileptogenesis prior to the first spontaneous seizure. We formulate the early diagnosis problem as an unsupervised anomaly detection task. We first train an adversarial autoencoder to learn a low-dimensional representation of normal EEG data with an imposed prior distribution. We then define an anomaly score based on the number of one-second data samples within one hour of recording whose reconstruction error and the distance of their latent representation to the origin of the imposed prior distribution exceed a certain threshold. Our results show that in a rodent epilepsy model, the average reconstruction error increases as a function of time after the induced brain injury until the occurrence of the first spontaneous seizure. This hints at a protracted epileptogenic process that gradually changes the features of the EEG signals over the course of several weeks. Overall, we demonstrate that unsupervised learning methods can be used to automatically detect systematic drifts in brain activity patterns occurring over long time periods. The approach may be adapted to the early diagnosis of other neurological or psychiatric disorders, opening the door for timely interventions.

Introduction

Epilepsy is a very common neurological disorder. Nearly 1% of the world's population will develop epilepsy at some point in their lives. Roughly 30% of these epilepsies will become drug-resistant (Kwan & Brodie, 2000), i.e., seizures cannot be controlled through medications. Epilepsy is often triggered by an initial brain injury, which is followed by a clinically silent so-called *latent* phase, during which the brain is undergoing a cascade of structural and functional changes. This process where the healthy brain transforms into an epileptic brain capable of generating spontaneous recurring seizures is called epileptogenesis (Löscher, 2019; Pitkänen & Engel, 2014). Importantly, the longer an epilepsy has been established, the more resistant to treatment it will be. Therefore, to issue early medical interventions and provide the potential epilepsy patients a better chance of living seizure-free lives, it may be helpful to identify epileptogenesis already before the first spontaneous seizure (FSS), which defines the beginning of an established epilepsy (Moshé et al., 2015).

EEG is a popular tool to measure brain activity at a high temporal resolution and it is often used in clinical settings and animal research (Löscher, 2019). The task of detecting epileptogenesis during the latent period, where there are no seizures yet, with EEG is very challenging and under-researched (Engel Jr & Pitkänen, 2020; Pitkänen et al., 2016), since it is often clinically silent. One contributing factor is that the data during this latent epileptogenesis phase is hard to acquire, especially in human patients. Usually, patients receive medical care only after experiencing at least one seizure. In animal epilepsy models, it is possible to acquire EEG data before the onset of the chronic seizures. However, due to a lack of well-established EEG biomarkers and well-annotated datasets, detecting epileptogenesis prior to the first spontaneous seizure remains a big challenge (Engel Jr & Pitkänen, 2020; Pitkänen et al., 2016).

Recent advances in machine learning (ML) offer promising directions for epilepsy research and have delivered encouraging results including seizure forecasting in canines with epilepsy (Nejedly et al., 2019), seizure forecasting and cyclic control in human patients (Stirling et al., 2021), epilepsy detection in clinical routine EEG data (Uyttenhove et al., 2020), as well as epileptogenesis detection and staging in animal epilepsy models (Lu et al., 2020a, 2020b). Specifically, there have been several studies on biomarker discovery for identifying epileptogenesis focusing on high-frequency-oscillations (HFOs) (Bragin et al., 2004; Burnos et al., 2014), dynamics of theta band activity (Milikovsky et al., 2017), asymmetry of background EEG (Bentes et al., 2018), and nonlinear dynamics of EEG signals (Rizzi

et al., 2019).

Generally, applying supervised ML to medical diagnosis problems is often hampered by the lack of large amounts of labeled training data. Therefore, we here consider a fully *unsupervised* learning framework that does not require any annotated data. Rather, the idea is to train a model to capture the statistics of normal healthy brain activity and use the model to subsequently detect systematic deviations from the healthy state. In our case, the types and the frequency of anomalous signals indicating the progression of epileptogenesis are not accessible and unpredictable during training. The signals are gradually evolving, which reflects the underlying changes taking place in the brain, evolving from a healthy brain to an epileptic one. This nature of the data renders a large amount of overlapping features between the healthy phase and the epileptogenic phase, which imposes grave difficulties for anomaly detection.

Inspired by the work from Schlegl et al. (2017) and Makhzani et al. (2015), we propose an adversarial autoencoder (AAE) network for anomaly detection in epilepsy progression. AAEs proposed by Makhzani et al. (2015) impose a prior distribution on to the latent codes learned by the encoder through the adversarial training. Here, we propose a flexible framework that makes use of different loss terms such as the reconstruction loss and the distance of the encoding distribution to the prior distribution to compute different anomaly scores.

Here, we would like to emphasize on one fundamental difference between our work and other works on seizure detection and prediction, i.e., there is **no seizure** yet in the data of interest in our work. We focus on detecting slow changes in brain activities before the very **first** unprovoked epileptic seizure aiming for early diagnosis of epilepsy (Fisher, 2015). This is a much more challenging problem that has only been recently addressed, but never with unsupervised methods (to the best of our knowledge).

Specifically, our contributions can be summarized as follows:

- We present an unsupervised adversarial autoencoder framework for detecting slowly evolving anomalies in brain activity.

- We validate our approach with data from a rodent epilepsy model and demonstrate good discriminative ability of signals from different phases of the epileptogenesis process.

Generalizable Insights about Machine Learning in the Context of Healthcare

In medical applications, massive amounts of data have been collected, however, obtaining expert annotations is extremely expensive and often infeasible. Especially, during the early disease progression phase, e.g., the case of early diagnosis of epilepsy, where the background normal activities are dominating the collected data and only gradual changes of certain features are involved. Our approach provides the opportunity of modeling the normal (healthy) data in an easy-to-acquire clinical setting and of detecting the slow evolution of disease progression in the collected query data. We emphasize that our framework is very general and could be applied to other neurological and psychiatric disorders, supporting early diagnosis and intervention. Moreover, our ablation studies show the significance of using adverserial training to further restrict the prior distribution of the latent space of the autoencoders trained on normal (healthy) EEG data. It led the autoencoders to learn an approximation to the normal (healthy) data distribution that maximized the separability between the normal (healthy) and anomalous (unhealthy) data.

Related Work

Early diagnosis of epilepsy holds great potential, since it might enable timely treatments that could potentially alter or even halt the disease progression. However, analysing large scale EEG data to discover bio-markers of epilepsy progression is very challenging. Recently, there has been an increasing interest in this area. For example, Rizzi et al. (2019) applied nonlinear dynamics analysis of EEG signals via recurrence quantification analysis. They found a significant decrease of the so-called embedding dimension in early epileptogenesis that correlates with the severity of the ongoing epileptogenesis. Buettner et al. (2019) identified two frequency sub-bands that are mostly effective in separating a healthy group from an epilepsy group with classic signal processing methods. Applying ML methods, Lu et al. (2020b) investigated the usage of raw EEG time series to distinguish mildly-injured and epileptogenic brain signals and demonstrated the potential of DNN-based methods in epileptogenesis detection. Furthermore, they extended the methods for staging the progression of epilepsy before the manifestation of the first spontaneous seizure (Lu et al., 2020a). In contrast to these supervised methods, we here propose an unsupervised anomaly detection approach, where the model is only trained with EEG signals that have been recorded prior to the disease-inducing injury in a rodent epilepsy model.

Anomaly detection (AD) describes a class of problems to detect samples that do not conform to the regularities of the training data. It can be addressed in a supervised learning, semi-supervised learning, or unsupervised learning fashion given the availability (or not) of sample labels (Gu et al., 2019). It can also be viewed as a one-class learning problem, where the training data are deemed to be the one class of interest. The models are trained to learn a classification boundary, either on a hyperplane (Schölkopf et al., 2001), or a hypersphere (Ruff et al., 2018; Tax & Duin, 2004) to separate anomalies from the nominal data (Ruff et al., 2019; Shen et al., 2020). Various AD methods are based on an encoder-decoder framework. In this framework, the model consists of two parts: an encoder and a decoder. The encoder maps the input into a lower-dimensional latent space representation, which the decoder uses to output a reconstructed version of the input. The reconstruction error between input and its reconstruction is usually used as the anomaly score, i.e., samples with high reconstruction error are deemed to be anomalous (P. Malhotra et al., 2016; B. Zhou et al., 2019). In addition, the error between the encoded latent vectors of the original input as well as that of the reconstructed input can be incorporated when defining the anomaly score (Kim et al., 2019). In the case where the knowledge of the anomalies is not accessible or is unpredictable during training, one can impose a regularizer on the learned latent distribution. Abati et al. (2019) propose to equip a deep autoencoder with a parametric density estimator, where the latent vector is generated in an autoregressive fashion. The overall model is trained to minimize the reconstruction error between the input and the output of the decoder network, as well as the log-likelihood of generating the latent vectors given the learned encoder network.

Adverserial autoencoders (AAEs) proposed by Makhzani et al. (2015) extend this notion of anomaly by imposing a prior distribution over the learned posterior by an encoder network through adversarial training. Specifically, an autoencoder is trained to reconstruct the input with low error, and an adversarial training process is applied to match the learned posterior distribution of the latent representation of the autoencoder to a prior distribution. One of the benefits of the AAE framework is the flexibility in choosing the prior distributions (Makhzani et al., 2015). The difference between AAEs and variational autoencoders (VAEs) is that VAEs use a KL-divergence term to impose a prior distribution on the latent code distribution, however AAEs achieve this by the adversarial training procedure. Schlegl et al. (2017) proposed a deep convolutional generative adversarial network trained to capture a manifold of normal anatomical variability in optical coherence tomography images of the retina based on the weighted sum of residual loss, a measure of reconstruction error, and discrimination loss. In Pidhorskyi et al. (2018), the proposed model consists of auto-encoders under the adversarial

training paradigm. Specifically, the probability distribution of the normal samples is learned through the encoder-decoder framework, and the anomaly score is computed through the evaluation of the probability of the test sample, i.e., normal samples will achieve high probabilities and anomalies will exhibit low probabilities.

It is common that the aforementioned methods assume that during the training there are no anomalous samples. However, in our case, we do not enforce this assumption, and in fact, we expect during the training phase, the model will encounter close-to-anomalous samples due to the nature of the experiment setup. Whilst many anomaly detection problems require label information during training (Gu et al., 2019; Tax & Duin, 2004), our method is completely unsupervised.

Dataset

The dataset used in this study stems from intracranial EEG recordings with a single depth electrode from a rodent mesial temporal lope epilepsy with hioppocampal sclerosis (mTLE-HS) model, where epilepsy is induced by electrical perforant pathway stimulation (PPS) (Costard et al., 2019; Norwood et al., 2011). Two groups of animals were considered by Costard et al., 2019: (1) PPS-stimulated rats, which developed epilepsy after an average epileptogenesis duration of 24 days (standard deviation 15 days), (2) control rats that had the depth electrode implantation as in the PPS group, but did not undergo the PPS and did not develop seizures by the end of recording (recording time was limited by the lifetime of the battery of the wireless transmitter). Continuous EEG recordings were obtained from the time of implantation of the depth electrodes. On average, a week of pre-stimulation (baseline) period was recorded for all rats. The EEG was recorded at the sampling rate of 512 Hz and band-pass filtered between 0.5 Hz and 176 Hz. Additionally, a notch filter at 50 Hz was applied to all the recordings.

The animal cohort used in this study consists of seven PPS-stimulated rats and three control rats. It is worth noting that during the data acquisition, there are several sources of noise in the signals: (1) electronic interference to the wireless transmission, which results in occasional extremely high amplitude peaks, (2) data loss during the transmission, which results in unchanging values for certain periods. To handle the these problems, we applied an outlier filtering method from MATLAB: `filloutliers` [1] with the parameters `method = 'pchip'; movmethod = 'movmedian'; window = 50`. Furthermore, we discarded the segments that have more than 20% data loss, which resulted in around 5% of the total recordings being discarded. Due to lack of annotations of artifacts such as movements, muscle

[1] **https://www.mathworks.com/help/matlab/ref/filloutliers.html**

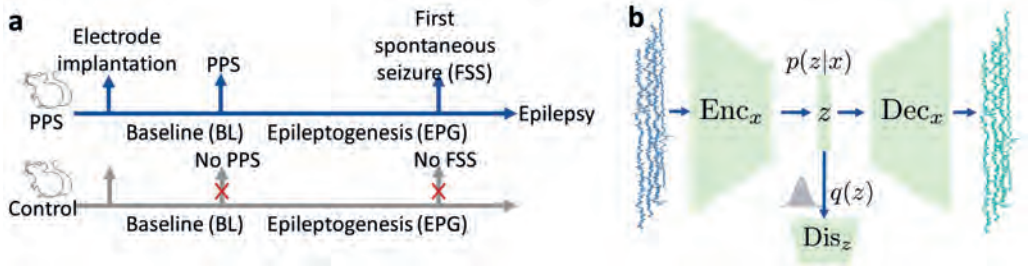


Figure S4.1. (a) Timeline of the experiment for the stimulated group (top) and the control group (bottom). Perforant pathway stimulation (PPS) is only performed on the PPS group but not the control group.(b) Proposed network structure. The backbone is a standard autoencoder, where the encoder Enc_x encodes the input into a latent representation z and the decoder Dec_x reconstructs the input from the vector z. Dis_z is a discriminator that distinguishes whether a sample z is from the encoded representation or drawn from the prior distribution $q(z)$.

twitching, chewing, etc., we do not discard them specifically. The time span of the experiment and the different phases are shown in Fig. S4.1a.

Methods

In this section, we describe the proposed adversarial autoencoder-based anomaly detection method in detail. The main idea is to train our model with only normal data from the training animals and measure the deviation of the test animal data from the learned distribution with an anomaly score based on two performance metrics: reconstruction error and distance of the latent code to the origin of the prior distribution. Code will be available online[2] for reproducability.

Proposed Model

We formulate our task as an unsupervised anomaly detection problem by learning only the distribution of the baseline EEG data through an adversarial autoencoder (AAE). The AAE is composed of three sub-networks: encoder, decoder, and discriminator (Fig. S4.1b). The encoder is trained to map the input data into a lower-dimensional latent space $p(z|X)$, which the decoder uses to reconstruct the input $p(X|z)$. By being trained to discriminate between true samples from the prior distribution and the fake samples generated by the encoder, the discriminator generates a teaching signal to the encoder to generate a latent code that matches the prior distribution. This adversarial loss serves two purposes: first it acts as a regularizer for the training and second it is used as an additional performance

[2] **https://github.com/amr-farahat/Epileptogenesis**

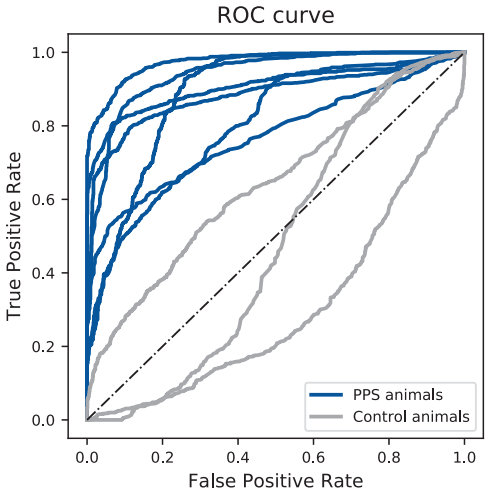


Figure S4.2. Receiver operating characteristic (ROC) curve for classifying baseline versus epileptogenesis periods for each animal in our dataset ($n = 10$). We show here the ROC curve for the $0.8\mathcal{R}$ anomaly score which is a weighted average of the $\mathcal{R}$ and $\mathcal{D}$ based anomaly scores. We use the count of supra-threshold one-second segments within one hour as an anomaly score. The threshold is selected to be the 99^{th} percentile of the training distribution of reconstruction errors and distances to the origin of the prior distribution of the latent space for the $\mathcal{R}$ and $\mathcal{D}$ based anomaly scores, respectively.

metric as we explain later. Specifically, the discriminator is trained with the loss function:

$$\mathcal{L}_{\text{Dis}} = -\log(\text{Dis}(\mathbf{z})) - \log(1 - \text{Dis}(\text{E}(\mathbf{X}))) \,, \tag{4.1}$$

where z are the true samples from the prior distribution and X are the data samples. On the other hand, the encoder and the decoder are trained with the loss function:

$$\mathcal{L}_{AE} = \|\mathbf{X} - \text{Dec}(\text{Enc}(\mathbf{X}))\|^2 \tag{4.2}$$

and the encoder/generator is trained with the loss function:

$$\mathcal{L}_{\text{Gen}} = -\log(\text{Dis}(\text{Enc}(\mathbf{X}))) \,. \tag{4.3}$$

Input data are one-second EEG segments collected as described in Section "Dataset". The encoder model is a residual convolutional neural network (He et al., 2016a) that consists of two blocks each composed of four residual units. Each residual unit is formed of two convolutional layers with kernel size $= 3 \times 3$ followed by batch normalization (Ioffe & Szegedy, 2015) and RELU activation functions. The number of kernels gradually doubles from 64 to 512 every two residual units and

the signal gets downsampled at the beginning of each block with $\texttt{stride} = 2$. At last, we have a convolutional layer with a kernel of size 1×1 to collapse the feature maps into the 128-dimensional latent code. The decoder model follows the same architecture, but with the use of transposed convolutions to upsample the latent code into the original 512-dimensional input size. The discriminator model is a fully connected network formed of two hidden layers each with 1000 units and followed by a leaky RELU activation function with $\alpha = 0.2$. The output layer is formed of one unit with a sigmoid activation function for binary classification.

The model is trained in two phases: a reconstruction phase and a regularization phase. In the reconstruction phase, both the encoder and the decoder are updated to minimize the reconstruction loss (Equation 4.2). In the regularization phase, the discriminator is first updated to distinguish between the true samples drawn from the prior distribution and the samples generated by the encoder (Equation 4.1). Then, the encoder/generator is updated to fool the discriminator (Equation 4.3). We balance the contributions of both $\mathcal{L}_{AE}$ and $\mathcal{L}_{Gen}$ to the trainable weights of the encoder/generator by a weighting parameter that we set to 0.99 and 0.01 respectively. All parts of the model are updated with the Adam optimizer (Kingma & Ba, 2014) with base learning rate = 0.0002, $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The prior distribution is a multivariate normal distribution with $\mu = 0$ and $\sigma = 0.1$ (see more in Section "Ablation"). We used MATLAB for preprocessing the data and used python for creating and training the models (specifically using the TensorFlow library (Martín Abadi et al., 2015)) and performing the post-hoc analysis of the results. It takes approximately 1.5 hours to train one epoch of 330-360 hours of EEG data.

After training, the AAE can be used to scan the query data to look for deviations from the training data distribution. In the data space, anomalous data are expected to have high reconstruction errors. On the other hand, depending on the nature of the changes in the brain activity due to the disease process (global or local changes in the signal), anomalous data can be expected to either lie in the the low or high probability density areas of the prior distribution used for training (Schreyer et al., 2019). For that reason, we additionally test the value of using the distance of the latent code to the origin of the prior distribution to define anomalous data.

Cross-validation Scheme

We adopt a leave-one-out (LOO) cross-validation scheme where we iterate over the list of all animals (seven PPS rats and three control rats) and in each iteration, we withhold the data from the test animal completely and train the

model on the normal data collected from all other (nine) animals. Since we aim for the model to capture the features of a normal EEG signal, we only use the data from the baseline period of the PPS groups. Additionally, we include the data from the control animals from the entire recording period. Note that it is shown that in longitudinal EEG recordings, various noise sources will be introduced due to the degradation of the implanted depth electrodes and changes in the electrode-tissue interface near the electrode (Kappenman & Luck, 2010; Straka et al., 2018). Hence, it is important to include the data from the control animals covering weeks of recording time in order to make sure that the model utilizes epileptogenesis-related features for discriminating between baseline and epileptogenesis periods and not the artifacts induced by the long-term recording. Specifically, we randomly selected 30 hours from the baseline period of each PPS animal and 75 hours from the whole recording period from each control animal to create the training dataset for each test animal in a LOO cross-validation scheme.

Detection Process

After training the full model on the training data from 9 out of 10 animals, we tested the ability of the trained model to distinguish between baseline and epileptogenesis periods of the data from the withheld test animal. Note that animals in the control group did not undergo the PPS. In order to keep the terms "baseline" and "epileptogenesis" consistent between the PPS group and the control group, we use the following notation for the control group. Baseline: one week period after the electrode implantation; epileptogenesis: starting 10 days after the electrode implantation. During testing, we apply the trained model to scan the data from the whole recording period of the test animal and compute the following metrics for each one-second segment $\mathbf{x}$: the reconstruction error ($\mathcal{R}$) and the distance of the latent code to the origin of the prior distribution ($\mathcal{D}$):

$$\mathcal{R}(\mathbf{x}) = \|\mathbf{x} - \mathrm{Dec}(\mathrm{Enc}(\mathbf{x}))\|^2 \tag{4.4}$$

$$\mathcal{D}(\mathbf{x}) = \|0 - \mathrm{Enc}(\mathbf{x})\|^2 \tag{4.5}$$

We set a threshold (λ) for these metrics based on the statistics of the training data, which is the 99^{th} percentile of the distribution of $\mathcal{R}$ and $\mathcal{D}$ computed from the training data. They are denoted by $\lambda_{\mathcal{R}}$ and $\lambda_{\mathcal{D}}$, respectively.

In order to aggregate the evidence from longer recording time periods and at the

same time simulate a clinical setting, we compute the number of suprathreshold segments within a certain time window ($\mathcal{T}$ = one hour), for both the baseline and the epileptogenesis data of the test animal and consider this number as the anomaly score ($\mathcal{S}$).

$$\mathcal{S}_{\mathcal{R}}(\mathcal{T}) = \sum_{i=1}^{n} I_{\mathcal{R}i} \tag{4.6}$$

where

$$I_{\mathcal{R}i} = \begin{cases} 1 & \text{if} \quad \mathcal{R}(\mathbf{x_i}) > \lambda_{\mathcal{R}} \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

and

$$\mathcal{S}_{\mathcal{D}}(\mathcal{T}) = \sum_{i=1}^{n} I_{\mathcal{D}i} \tag{4.8}$$

where

$$I_{\mathcal{D}i} = \begin{cases} 1 & \text{if} \quad \mathcal{D}(\mathbf{x_i}) > \lambda_{\mathcal{D}} \\ 0 & \text{otherwise} \end{cases} \tag{4.9}$$

where n is the number of one-second segments in time window $\mathcal{T}$, e.g., 3600 in one hour.

Consequently, we evaluate the ability of this aggregated anomaly score to distinguish between baseline and epileptogenesis data by computing the receiver operating characteristic (ROC) curve and calculating the area under the curve (AUC). We compute the ROC-AUC with the aggregated $\mathcal{R}$ and $\mathcal{D}$ metrics. Moreover, we investigate whether a weighted average of both anomaly scores would lead to better classification results.

Results

Epileptogenesis Detection

The main goal of this study is to investigate the potential of using electrical brain activity in an unsupervised way for predicting brain disorders and follow their development as the brain activity deviates from its baseline distribution. We trained an AAE on the baseline intracranial EEG data collected from PPS rats before stimulation and from control rats in a leave-one-out cross-validation scheme. For each test animal, we used the corresponding model to scan its whole data and record the average reconstruction error for each one-second segment in the data space. Additionally, we recorded the distance of the latent code to the origin of its prior distribution. We considered different metrics to compute the

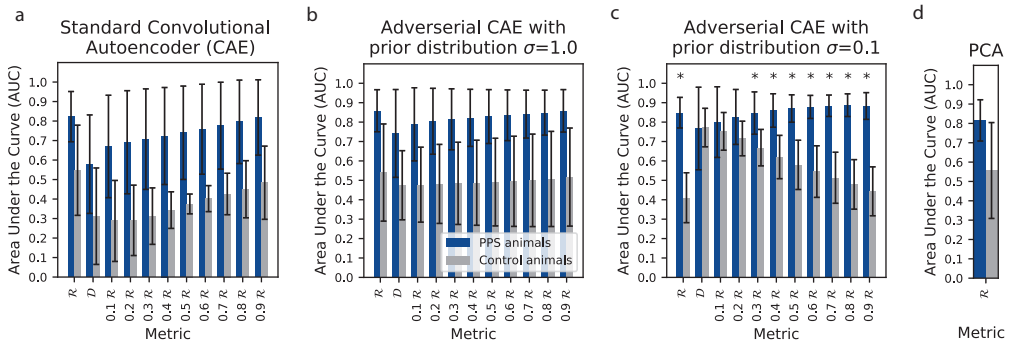


Figure S4.3. Average area under the curve (AUC) for the PPS animals and control animals for different definitions of the anomaly score and different models. (a) Standard convolutional autoencoder (CAE). (b,c) Adversarial CAE with different prior distributions. (d) Principal Component Analysis (PCA). In each panel, $\mathcal{R}$ denotes the reconstruction error metric, $\mathcal{D}$ denotes the metric based on distance of the latent code to the origin of its prior distribution. The remaining columns consider weighted averages of $\mathcal{R}$ and $\mathcal{D}$; the weight of $\mathcal{R}$ is indicated. The asterisks above the bars denote that the difference between PPS and control animals is statistically significant according to a Mann–Whitney U test (*:$p < 0.05$ FDR-adjusted for multiple comparisons.)

anomaly score: the reconstruction error in the data space ($\mathcal{R}$) and the distance to the origin of the Gaussian prior distribution in the latent space ($\mathcal{D}$). Using each of these metrics, we computed an anomaly score by counting the number of supra-threshold segments within one hour. The threshold was computed as the 99^{th} percentile of the training distribution of this metric. We randomly sampled 1000 hours from each of the baseline and the epileptogenesis periods of the test animal, computed the anomaly scores for them, and calculated the receiver operating characteristic (ROC) curve for discriminating between the two periods for each test animal in the dataset. We also computed additional anomaly scores as the weighted averages of the anomaly scores computed based on the $\mathcal{R}$ and $\mathcal{D}$ metrics which we denote $x\mathcal{R}$ where $x \in [0, 1]$ and represents the weight assigned to the $\mathcal{R}$-based anomaly score where the $\mathcal{D}$-based metric is assigned the weight $1 - x$ (see Fig. S4.2 for the ROC curve based on the $0.8\mathcal{R}$ metric as it was our best performing anomaly score and Fig. S4.3c for the average area under the curve (AUC) for all anomaly scores). We observe that control animals have their ROC curves around the diagonal which is expected since they were not exposed to PPS and therefore there should not be a significant difference between their baseline and hypothetical epileptogenesis periods. On the other hand, while there is variability among PPS animals, all their ROC curves lie above the diagonal, which denotes above chance discrimination performance. This is also reflected in the significant difference between the average AUC of PPS and control animals

(Fig. S4.3c first two bars). Contrarily, we note that the anomaly score based on the $\mathcal{D}$ metric alone does not show a difference between animal groups (Fig. S4.3c third and fourth bars), which means it is not a good metric for computing the anomaly score for discriminating between baseline and epileptogenesis periods. Next in the ablation study, we examine the value of the adversarial loss as a regularizer.

Ablation Study

In Fig. S4.3c, we noticed that the discriminative ability of the model using only the $\mathcal{R}$ metric is better than that with only the $\mathcal{D}$ metric. This is reflected in the AUC from control animals being around the chance level for the $\mathcal{R}$ metric and significantly above the chance level for the $\mathcal{D}$ metric. This suggests that the differences between the normal and anomalous data in the data space are too subtle for the encoder to push them into the low-density areas in the lower-dimensional latent space.

To further investigate the relevance of different loss components of the proposed method to the final epileptogenesis detection task, we performed ablation studies. To this end, we trained a standard convolutional autoencoder (CAE) (Fig. S4.3a) and an adverserial CAE with a standard Gaussian prior distribution ($\sigma = 1.0$) (Fig. S4.3b) rather than our proposed method with $\sigma = 0.1$ (Fig. S4.3c). We notice that even though the $\mathcal{D}$ metric did not prove useful alone for computing an anomaly score that maximizes the separability between baseline and epileptogenesis periods, adding the adversarial loss acted indirectly as a regularizer that boosted the discriminability of the $\mathcal{R}$-based anomaly score as evident by the high variability of the average AUC of the PPS and control animal groups in case of the standard CAE (Fig. S4.3a first two columns). Average AUC of PPS animals improved from 0.82 with $std = 0.13$ to 0.85 with $std = 0.08$. Additionally, using the weighted average of both $\mathcal{R}$ and $\mathcal{D}$ based anomaly scores improved the average AUC of PPS animals from 0.85 with $std = 0.08$ to 0.89 with $std = 0.06$ ($0.8\mathcal{R}$) but only when training with prior distribution with $\sigma = 0.1$ while there was no improvement for the standard CAE or when training with prior distribution with $\sigma = 1.0$. This can be explained by the fact that at the beginning of training with standardized inputs and random weights, the encoder already produces a latent code that approximates samples from a standard Gaussian distribution. Consequently, the discriminator does not get the chance to learn the prior distribution and send a teaching signal to the encoder/generator. Therefore, making the problem harder for the encoder/generator by restricting the standard deviation of the prior distribution has a better regularizing effect on the trained models.

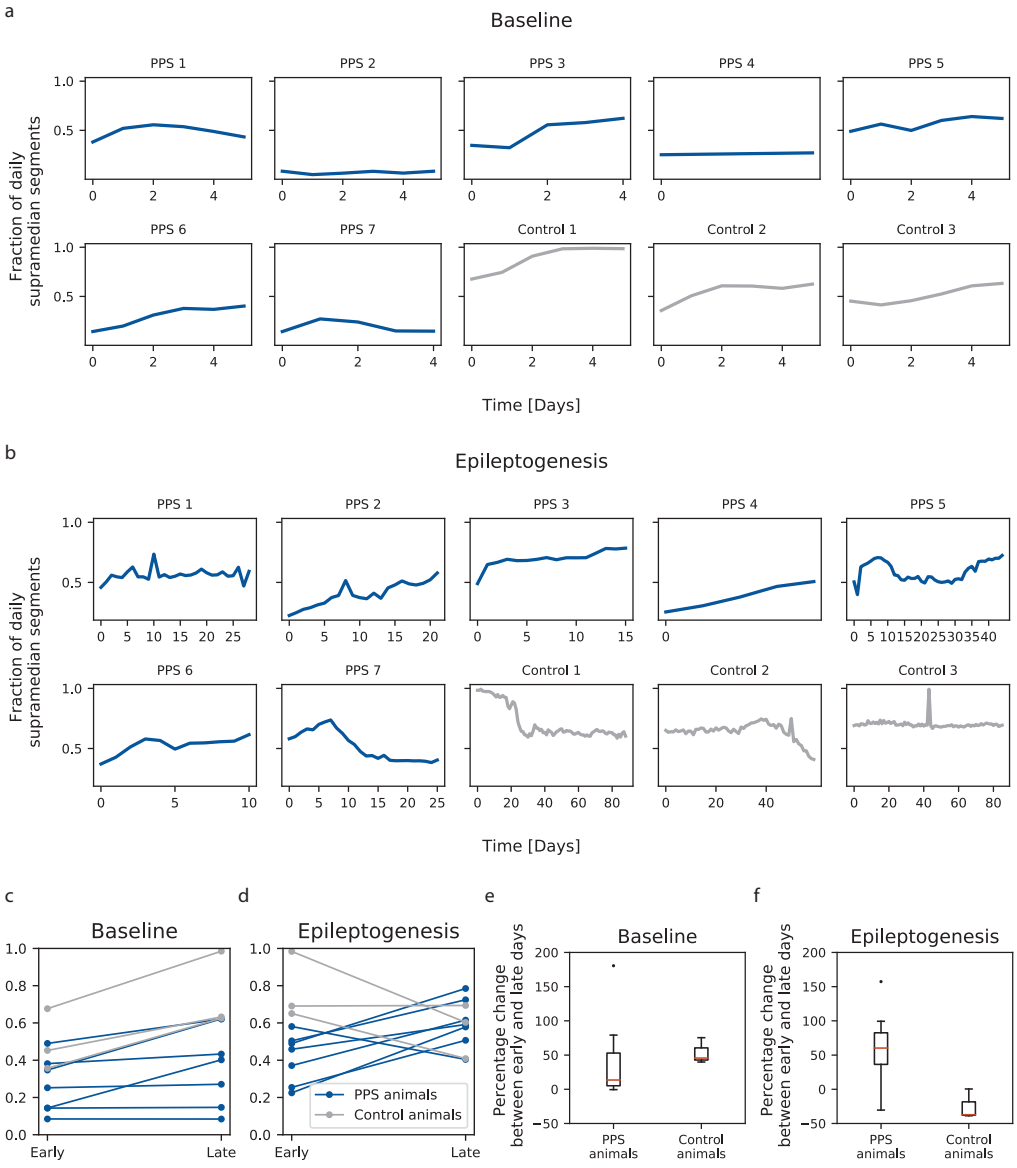


Figure S4.4. Fraction of one-second segments that exceeds the median of the training distribution of reconstruction errors (supra-median fractions) for each animal in our dataset ($n = 10$) for the baseline (a) and epileptogenesis (b) periods. Comparing supra-median fractions between the first (early) and last (late) day of the baseline (c) and epileptogenesis (d) periods for each animal in our dataset ($n = 10$). Average percentage of change between early and late supra-median fractions for each animals group (PPS or control) for baseline (e) and epileptogenesis (f) periods.

Moreover, we compared to a linear baseline for reconstruction-based anomaly detection by using Principal component analysis (PCA) to reduce the dimensionality of the data to 128 components and then project back to the data space and compute a reconstruction error. We merely obtain average AUC for PPS animals of 0.82 with $std = 0.11$ which does not show statistically significant difference to the control group AUCs (Fig. S4.3d). This is comparable to the standard CAE results, but falls short to our best achieved results with the adversarial CAE with a gaussian prior distribution with $\sigma = 0.1$ (Average AUC $= 0.89$ with $std = 0.06$).

Time Course of Epileptogenesis

We have shown so far that the $\mathcal{R}$-based anomaly score was successful at differentiating between EEG signals recorded during the baseline period and the epileptogenesis period after PPS. Next, we examined what the temporal evolution of reconstruction errors of the EEG signal can reveal about the epileptogenesis process. For each full 24-hour day in the baseline and epileptogenesis periods, we computed the fraction of one-second segments that have a reconstruction error that exceeds the median of the reconstruction error training distribution (fraction of daily supra-median segments in Fig. S4.4). We notice that the time course of supra-median fractions is complex and variable across animals in both periods. However, it is less variable in the baseline period specifically when we consider the difference between the control and PPS animal groups (Fig. S4.4a and quantified in Fig. S4.4c by contrasting the first and the last full-days of the whole period). On the one hand, all animals tend to have either stable or slightly increasing daily supra-median fractions across the whole baseline period. On the other hand, in the epileptogenesis period (Fig. S4.4b and quantified in Fig. S4.4c), control animals tend to have stable or decreasing daily supra-median fractions. This is in contrast to PPS animals, which mostly, with the exception of only one animal (PPS 7), have increasing daily supra-median fractions. Additionally, we computed the percentage change in the daily supra-median fractions between the first and last day for baseline and epileptogenesis periods for each animal in our dataset. Looking at the averages across animal groups for each period (Fig. S4.4e and f), we observe that both animal groups have comparable percentage change in daily supra-median fractions in the baseline period (p-value is 0.18 with Mann–Whitney U test). In contrast, PPS animals show significantly higher percentage change than control animals in daily supra-median fractions during the epileptogenesis period (p-value is 0.02 with Mann–Whitney U test). These results show that the epileptogenesis process causes alternations to the brain that are reflected in its electrical activity, which is in turn reflected in the ability of the model to reconstruct this electrical signal. These changes in brain activity get

progressively stronger and consequently, the reconstruction errors increase.

Discussion

Machine learning techniques have been transforming many domains of investigation, in particular those that require detecting patterns in vast amounts of data. Healthcare applications have been at the top of the list of these domains, specifically when it comes to diagnosing diseases or rehabilitating patients by training machine learning models on labeled biomedical data like X-Rays (Rajpurkar et al., 2017), magnetic resonance imaging (MRI) (Lundervold & Lundervold, 2019), EEG (Farahat et al., 2019; Lu et al., 2020a), and electrocardiograms (ECG) (Hannun et al., 2019). Machine learning algorithms trained on large amounts of data can discover new patterns in the data, e.g., diagnostic biomarkers, that may be too subtle to be detected by humans. However, one problem is that the data collected in the medical domain are usually imbalanced. There is a scarcity of abnormal data that corresponds to certain diseases and disorders relative to normal data from healthy subjects. Also, collecting data from patients is subject to regulations that protect the privacy of the patients which makes it harder to obtain.

One potential approach to overcome this problem of scarcity of abnormal data is to leverage the abundance of normal data by training machine learning models to learn the distribution of normal data and then survey the query data for deviations from this learned distribution. Clinically, this approach can work as a screening procedure for individuals with risk factors who can then be further evaluated by professionals. Technically, this approach has the advantage of only requiring the relatively cheap data of healthy subjects. However, this approach is challenging when the deviations from normal data caused by the disease process are subtle (especially early in the disease) and develop gradually over a long period of time, which is the case in epileptogenesis.

In this study, we have given a proof of concept that such an approach can be implemented through an adversarial convolutional autoencoder model. We trained the model on normal EEG data collected from a rodent epilepsy model and used it in an anomaly detection paradigm to screen the data of test animals to discriminate between the data collected before and after PPS, i.e., to detect a developing epilepsy. The anomaly scores were computed based on how the reconstructed signal deviates from the original signal and could be viewed as a proxy of how the epileptogenesis process develops over time after PPS. This is important as anticipating epilepsy before the FSS could urge medical intervention

that significantly improves the patients' long-term quality of life (Moshé et al., 2015). Note that we chose the time window of our anomaly score computation to be one hour — which is clinically feasible — to act as a simulation for a clinical routine.

Limitations The main goal of this study was to test the potential of an unsupervised deep anomaly detection paradigm in detecting subtle changes in brain electrical activity as a consequence of a brain-altering disease process. Despite the success of the approach, it still falls short of a fully supervised approach. In particular, using the same dataset, a previous supervised approach achieved an average AUC = 0.93 for distinguishing between baseline and epileptogenesis in PPS rats (Lu et al., 2020a), in contrast to 0.89 for our approach. This is expected as in our approach, the model does not have access to any epileptogenesis data. Another difference is that the authors of that study used five-second segments instead of one-second segments used here. However, we also experimented with five-second segments and obtained similar results. Nevertheless, given the advantages of our approach mentioned earlier, it is worth pursuing and with further advances in unsupervised and self-supervised learning techniques, we expect further improvements.

Another limitation of our approach is that we computed anomaly scores on relatively short one-second (or five-second) EEG segments. While our approach aggregates these scores over longer periods of one hour, it does not look for patterns at these longer time scales. This choice was motivated by the fact that identified frequency bands that effectively differentiate healthy subjects from epileptics in the epileptogenesis period lay above 1 Hz (Buettner et al., 2019). However, we can not exclude the possibility that there is additional valuable information in lower frequency bands that are not usually considered in EEG analysis.

A final limitation is the relatively small number of individuals considered in this study.

Outlook In the future, we plan to pursue two broad directions with this approach. First, we aim to translate the results to human patients at risk of developing epilepsy. Second, we would like to test the generality of the approach by applying it to other neurological or psychiatric disorders. In particular, several psychiatric disorders are characterized by the alternation of episodes of different "states". Examples are bipolar disorder or schizophrenia. Detecting transitions between these states early and automatically could improve the management of such disorders. Critically for both research directions is to investigate the applicability of the approach to

non-invasive surface EEG recordings.

Acknowledgments

This work was supported by the China Scholarship Council (No. (2016)3100), the LOEWE Center for Personalized Translational Epilepsy Research (CePTER), and the Johanna Quandt Foundation.

Chapter 5
Summary and General Discussion

Summary

In **Chapter 1**, I introduced DNNs as pivotal tools for analyzing complex, high-dimensional data across scientific fields, including neuroscience. DNNs serve a dual role: as powerful analytical tools and as computational models of brain systems (e.g., CNNs for the primate visual system). However, a trade-off exists between a model's predictive power and its ability to provide insight into the underlying phenomenon; replacing a complex biological system with a "black box" model may not enhance understanding. Therefore, modelers must strategically select model types based on whether prediction or explanation is prioritized for a given research question.

In **Chapter 2**, I investigated whether CNNs utilize spatial relationships between features for object recognition, employing a novel feature-scrambling approach. CNNs with restricted Effective Receptive Fields (ERFs) were trained on various datasets (Sketchy, Animals, ImageNet). Pretrained features from these CNNs were spatially scrambled and fed into a follow-up network. This follow-up network was trained for object recognition, allowing assessment of the impact of spatial scrambling on performance. Minimal Recognizable Configurations (MIRC) analysis quantified the minimal image patch sizes necessary for correct classification. The results demonstrated that CNNs are capable of using spatial relationships between features for object classification, particularly for textureless datasets like sketches. The extent of this reliance on spatial relationships depends on the dataset and even varies between classes within heterogeneous datasets (e.g., ImageNet). However, CNNs learn the spatial arrangement of features only up to an intermediate level of granularity, not capturing the global shape holistically. This limitation may stem from optimization pressures, as intermediate features offer an optimal balance between sensitivity and specificity, crucial for object recognition.

In **Chapter 3**, I explored the relative importance of network architecture and training in enabling CNNs to predict neural responses in the primate visual cortex, focusing on early (V1) and higher visual areas (IT, VO). The study compared task-optimized CNNs (trained for object recognition on ImageNet) and brain-optimized CNNs (trained directly to predict neural activity) with random-weight CNNs (without convolutional filter training). For V1, random-weight CNNs with ReLU activations and max pooling achieved performance comparable to trained networks in predicting neural responses, unlike higher visual areas (IT, VO). This

suggests that, for V1, the architecture itself, specifically its non-linear complexity, is a primary factor in encoding visual information, and convolutional filter training is less critical. In contrast, higher visual areas require the precise weight configurations learned through training to effectively encode complex visual information. Furthermore, model complexity, quantified using Chebyshev polynomials, correlated strongly with V1 encoding performance for both random and trained networks. Random-weight ReLU networks exhibited orientation selectivity similar to V1 neurons, and training on V1 data further enhanced this selectivity, indicating that the inherent structure of ReLU networks can capture basic visual features like orientation tuning, even without explicit training. Random-weight networks performed surprisingly well on a texture discrimination task, nearing the performance of trained networks. However, for object recognition (digit recognition), trained networks were significantly superior, suggesting that random networks suffice for simpler visual tasks like texture processing, while object recognition necessitates task-specific training.

In **Chapter 4**, I proposed a novel, unsupervised deep learning framework for the early diagnosis of epileptogenesis (the process of developing epilepsy) using EEG recordings. An Adversarial Autoencoder (AAE) was trained on EEG data representing healthy brain states. The model learned to reconstruct normal brain activity and represent it in a lower-dimensional latent space. Subsequently, when presented with new EEG data, the model's reconstruction error and the distance of the latent representation from the learned normal distribution were used as anomaly scores. The model successfully detected subtle, gradual changes in EEG signals associated with epileptogenesis before the first spontaneous seizure. The value of this unsupervised approach, compared to supervised methods, lies in its independence from labeled epileptic activity data, making it practical when disease data are scarce. This framework could be adapted for the early diagnosis of other neurological and psychiatric disorders characterized by gradual changes in brain activity.

"Unfortunately, nature seems unaware of our intellectual need for convenience and unity, and very often takes delight in complication and diversity."

— Santiago Ramón y Cajal

Prioritizing accurate predictions over model explanation in DNN models for neurological screening

In **Chapter 1**, I introduced the trade-off between a model's prediction accuracy and its interpretability: the ability to explain its predictions/decisions is often inversely correlated with its predictive performance (Chirimuuta, 2021). Highly predictive models, like DNNs, tend to be non-linear, complex, and consequently opaque (i.e., "black boxes"). Ideally, we should strive for models that are both highly predictive and highly interpretable, especially in high-stakes scenarios like medical diagnostics. High accuracy can improve patient outcomes through correct diagnoses, while interpretability builds trust and supports clinical decision-making (Rudin, 2019; Yoon et al., 2022). However, when a trade-off is unavoidable, the prioritization of prediction accuracy or interpretability becomes a debatable issue.

Some researchers argue that interpretability is essential for deploying computational models in healthcare (Kundu, 2021; Rudin, 2019; Yoon et al., 2022). Interpretability is crucial for identifying and mitigating biases in data and algorithms, which can perpetuate health inequities. It also builds trust, facilitates clinical acceptance, aids in error auditing, improves model development efficiency, and is critical for establishing accountability in cases of model failures. Some even argue against using black-box models in high-stakes scenarios, even with post-hoc explanations, claiming these explanations are often inaccurate and incomplete (Rudin, 2019). For instance, saliency maps might highlight image regions relevant for a classifier but not how those regions are used and integrated within the model.

Conversely, other researchers argue that explaining model decisions is not strictly necessary for adopting computational models in healthcare (Durán & Jongsma, 2021; Kawamleh, 2023; London, 2019). They argue that model reliability trumps explainability in establishing trust (Durán & Formanek, 2018). Reliability can be assessed through indicators like verification/validation, robustness analysis, a history of successful implementations, and expert knowledge (Durán & Formanek, 2018; Durán & Jongsma, 2021). Therefore, model developers should prioritize em-

pirical validation on diverse datasets, across different populations and hospitals, to mitigate biases (McKinney et al., 2020; Ting et al., 2017). Manipulating and corrupting datasets can also test model robustness and identify failure scenarios. Furthermore, clinical decision-making often relies on correlational evidence even when the underlying causal mechanisms of a disease or intervention are unknown. In this sense, some routine medical practices are not fundamentally different from opaque machine learning models (London, 2019). Randomized controlled trials (RCTs) can provide evidence for the efficacy of complex, opaque models, just as they do for many medical interventions, without requiring a full explanation of their mechanisms (Hernström et al., 2025). Moreover, clinicians often trust and operate machines they cannot fully explain (e.g., MRI machines) (Durán & Jongsma, 2021). Just as post-hoc explanations of complex models are criticized, similar critiques apply to doctors providing post-hoc justifications for their decisions, which may rely on intuitive judgment and unconscious biases (Carruthers, 2011; Kawamleh, 2023). Expert radiologists, for example, often struggle to articulate rule-based explanations for their diagnoses (Hegdé & Bart, 2018; Kawamleh, 2023; Sevilla & Hegde, 2017). Therefore, demanding higher explainability standards for computational models than for human experts may be unreasonable.

However, clinical practices are diverse. Prioritizing either prediction accuracy or interpretability likely depends on the specific clinical context, the stakes, and the model's intended use. Whether models are designed to replace physicians or merely aid their decision-making also influences the required level of explainability. For example, in ethically sensitive areas like resource allocation (e.g., organ transplants), interpretability is paramount for accountability, fairness, and transparency. In contrast, for screening and triage, the primary goal is efficient and accurate identification of individuals needing further attention, reducing the burden on the healthcare system (Hernström et al., 2025; McKinney et al., 2020; Ting et al., 2017). Here, high sensitivity is crucial to avoid missing potential cases. A highly accurate black-box model, even with limited interpretability, can be valuable, especially if experts review positive cases. Indeed, when primary care providers were surveyed, they valued sensitivity most when considering black-box AI models for breast cancer screening (Hendrix et al., 2021). They did not prefer a radiologist confirming the diagnosis of all images over confirming only the likely positive images suggested by the model. This illustrates a framework where DNN models collaborate with, rather than replace, humans in screening workflows. Recent RCT results show that this approach improved early breast cancer detection, reduced workload, and did not increase false positives (Hernström et al., 2025).

Similarly, in **Chapter 4**, I introduced a DNN-based proof-of-concept framework for

screening EEG signals for early signs of post-traumatic epilepsy during its development. Due to the relative scarcity of disease data compared to normal data, the model was not trained in a supervised manner to discriminate between normal and disease EEG signals. Instead, a generative adversarial convolutional autoencoder learned the normative distribution of intracranial EEG data across multiple rodent subjects. The model could then flag EEG segments as anomalous based on their reconstruction errors. To improve sensitivity, crucial in clinical screening (Hendrix et al., 2021), evidence of anomalous segments was aggregated over a one-hour period, which is clinically feasible. Crucially, the model was built with the intended clinical application in mind: 1) One-dimensional convolutions were used to fit the time-series EEG data. 2) Unsupervised anomaly detection training leveraged the relatively abundant and easily obtainable normal data. 3) Evidence aggregation over an extended time (one hour) improved sensitivity. Multiple anomaly scores were evaluated to identify the most accurate. 4) Temporal progression of the fraction of anomalous data points was assessed, mirroring common medical practice of repeated measurements to confirm diagnoses (e.g., multiple high blood pressure readings for hypertension diagnosis). 5) The model was validated using a leave-one-out cross-validation scheme, testing generalizability across individuals in the population. 6) A simpler, more interpretable PCA model was compared. However, its lower accuracy led to prioritizing the higher accuracy of the less interpretable model. Finally, any clinical finding, whether from a complex opaque model or a simple interpretable lab measurement, must be contextualized and integrated with other clinical findings (medical history, symptoms, other diagnostics) before making a clinical decision. Therefore, as long as the DNN model outputs are integrated into clinical workflows under human physician supervision, their explainability should not be a barrier to clinical adoption.

Explanation of biological and artificial visual intelligence using DNNs

As argued above, researchers must be aware of the prediction-explanation trade-off and develop models accordingly. If the goal is understanding brain function, model interpretability is naturally essential. However, accurately predicting responses of a complex organ like the brain and the complex behaviors it supports (e.g., object recognition) may be impossible with simple, interpretable models (Wichmann & Geirhos, 2023). Therefore, employing complex, opaque models is often unavoidable. Similar to the argument for integrating multiple clinical findings, including outputs from complex models, scientific understanding of a

phenomenon depends on integrating information from multiple sources. Therefore, we can distinguish between model explanation and model-induced explanation (Kästner & Crook, 2023; Lawler & Sullivan, 2021).

In model explanation, the model's content is the explanation; the explanation of the phenomenon is found directly within the model's structure and inner workings. For example, if a DNN predicts neural activity or performs object recognition, a model explanation approach would focus on deciphering the internal mechanisms, parameters, and computations that transform input stimuli into neural activity or behavioral output (e.g., object class). Achieving complete model explanation for highly non-linear, complex DNNs is challenging, if not impossible (Lipton, 2016; Rudin, 2019). Even if a trained DNN were converted into mathematical equations, those equations would likely be unintelligible to humans (Chirimuuta, 2021).

In model-induced explanation, the relevant explanatory information is independent of the model itself. The model acts as a tool or mediator to uncover understanding, but that understanding isn't necessarily contained within the model (Kästner & Crook, 2023; Lawler & Sullivan, 2021). Using the same example of a DNN trained to predict neural activity or perform object recognition, a model-induced explanation approach would focus on generating questions or suggesting hypotheses about the modeled phenomenon that can be further validated using other methods. Essentially, researchers move beyond understanding the black-box model to illuminate the real black box – the brain itself.

Texture and shape bias of CNNs as examples of model explanations

CNNs excel at object recognition. However, uncovering the perceptual dimensions (e.g., shape, texture, color) CNNs utilize for categorization has been an open research question. Initially, it was assumed that CNNs primarily used shape information, as suggested by feature visualization techniques (LeCun et al., 2015; Olah et al., 2017; Zeiler & Fergus, 2014), which often showed object parts as the most activating features in deep layers. More recent, hypothesis-driven techniques yielded conflicting results (Baker et al., 2018; Brendel & Bethge, 2019; Geirhos et al., 2019; Kubilius et al., 2016; Ritter et al., 2017; Tartaglini et al., 2022).

For example, Ritter et al., 2017 used a dataset inspired by cognitive psychology experiments, with image triplets (probe, shape-match, color-match), to test shape bias. Shape bias was estimated by the proportion of shape labels assigned to the probe, calculated using distances in the CNN representation space. This study concluded that ImageNet-trained CNNs exhibited a strong shape bias. Another

study used images with texture-shape cue conflict images, generated using style transfer (Gatys et al., 2016), to quantify shape bias in both humans and CNNs (Geirhos et al., 2019). Shape bias was measured as the percentage of trials where participants/models responded with the shape category. This study found that ImageNet-trained CNNs - unlike humans - were strongly biased towards texture. A more recent study replicated both approaches with modifications (Tartaglini et al., 2022). By decreasing the opacity of background texture in cue-conflict stimuli, they found that removing the influence of background texture resulted in a preference for shape over texture. Moreover, using image triplets with texture-match probes instead of color-match probes and manipulating stimulus size, they found shape bias for small stimuli, decreasing with increasing size. These conflicting results highlight the fragility and incompleteness of post-hoc model explanation techniques, as conclusions can depend on subtle design choices. Moreover, these studies often make an implicit, and likely incorrect, assumption equating texture with local features and shape with global features.

In **Chapter 2**, I developed a novel feature-scrambling approach to address CNN object representations in a hypothesis-free manner. The goal was to determine the granularity of learned features sufficient for reasonable object recognition and whether CNNs could combine finer-grained features to create coarser-grained features in a way sensitive to spatial relations. Importantly, this approach did not involve manipulated stimuli or special datasets that could predetermine the observed bias. It also did not assume what constitutes shape or texture, but rather tested how CNNs integrate features along their depth and whether spatial relations are crucial for this integration. That is because encoding spatial relations between features is essential for the emergence of shape object representations (Biederman, 1987; J. Hummel & Biederman, 2002; J. Hummel, 2013). Analysis revealed that CNNs can use spatial relations to integrate fine-grained features, constructing coarser-grained ones. However, the extent of this integration depends on the dataset and object class. It was also limited to an intermediate level of granularity, not capturing the object's global form, even for textureless datasets (e.g., sketches). These results were validated using another explainability technique, MIRCs. MIRC analysis showed that CNNs could correctly recognize natural objects and sketches from partial image crops that did not include the objects' global forms. These intermediate features are optimal for object recognition, balancing sensitivity and specificity (Ullman et al., 2002). These findings help reconcile previous conflicting conclusions, suggesting that human shape representations may not be an emergent property of optimizing for object recognition (G. Malhotra et al., 2022) and might not originate in the ventral visual stream, which is primarily associated with object recognition (Ayzenberg & Behrmann, 2022a,

2022b; Jagadeesh & Gardner, 2022; Long et al., 2018). The intermediate features CNNs learn could be combinations of local shape and texture features. When texture covers the entire image in texture-shape cue conflict stimuli, texture-based evidence overwhelms local shape features, leading the CNN to classify based on texture. Removing the background allows local shape features (edges, contours) to provide relatively stronger evidence for their associated class.

In summary, modeling complex behaviors like object recognition necessitates complex, opaque models like DNNs. We can still gain insights into how DNNs perform these tasks using novel post-hoc explanation techniques. However, researchers must recognize that post-hoc explanations can be incomplete and deficient. Their results must be integrated within a broader literature to yield valuable insights.

Model-induced explanations of the importance of both architecture and training in modeling different stages of the visual hierarchy

Modeling the mapping from high-dimensional input stimuli to high-dimensional neural activity requires complex, non-linear models like DNNs. But how can we learn about brain computations by replacing one black box (the brain) with another (the model)? Can DNN models induce understanding of the modeled system despite being complex and opaque? (Kästner & Crook, 2023; Lawler & Sullivan, 2021).

Cadena et al., 2019 used an ImageNet-trained VGG19 model to predict V1 neural responses in macaques, achieving better performance than traditional models like linear-non-linear Poisson models and Gabor filter banks. The best predictive layer was *conv3_1*, five non-linear transformations from the input. Two conclusions could be drawn: First, V1 neurons might perform more complex computations than previously thought. However, a more recent study showed that with an AlexNet model, the first layer was the best predictor of V1 responses, suggesting that V1 prediction might not require as many non-linear transformation steps (Miao & Tong, 2024). The key difference is that AlexNet uses larger convolutional filters, increasing receptive fields faster along the network's depth. Du et al., 2024 similarly found that model performance saturated after only two convolutional layers when training simple CNNs directly on V1 data. These findings challenge the conclusion that modeling V1 computations requires many non-linear transformations.

Second, object recognition training might be important for predicting V1 activity. However, the study lacked a random-weight CNN control to validate this. In **Chapter 3**, I showed that random-weight VGG16 models performed comparably

to their ImageNet-trained counterparts in predicting V1 neural data, suggesting that Cadena et al., 2019's findings can be largely explained by CNN architecture, not object recognition training. Similar to Cadena et al., 2019, Du et al., 2024 did not report random-weight CNN control model performance besides their models directly trained on V1 data. Surprisingly, I found that simple 2-layer CNNs trained directly on V1 neural data did not considerably outperform their random-weight counterparts on held-out test data, particularly when using ReLU activations instead of ELU as Du et al., 2024. This highlights ReLU's importance in providing the appropriate architectural bias for efficient V1 response modeling, given the number of trainable parameters. Importantly, random-weight ReLU CNNs performed comparably to trained counterparts in predicting V1 responses in macaques and humans, but not in higher visual areas (IT, VO). This indicates that the relative contribution of architecture and training varies across the visual hierarchy, aligning with the idea that higher areas perform more specialized, task-dependent computations.

Thus, careful control experiments, accounting for potential confounding factors, can provide insights into brain computations, even with opaque models. We learned that V1 computations can be approximated by relatively shallow (2-layer) random-weight CNNs with pooling and ReLU activations. It was possible to identify the critical components (e.g. ReLU) and show that training the convolutional filters was not essential. However, fully training models with different architectures (e.g., ELU or Tanh) yielded comparable neural encoding performance, highlighting a limitation of relying on a single metric like prediction accuracy. Therefore, I adopted a multidimensional assessment approach, going beyond neural response prediction. I evaluated model neuron orientation selectivity and compared it to that of an independent set of experimental V1 neurons. I found high variability in V1 deviation scores, contrasting with the low variability of V1 prediction accuracy. ReLU models were the most V1-like in both orientation selectivity and prediction accuracy. Random-weight ReLU models V1 deviation scores were comparable to those of fully trained models with other activation functions, providing further evidence that basic V1 features (like orientation tuning) can emerge without task-specific training. These findings support the idea that ReLU activations are crucial for V1 computation modeling, especially considering their biologically-inspired history (Glorot et al., 2011). I also showed that ReLU random-weight CNNs are functionally relevant, significantly outperforming other random-weight models in visual tasks like texture discrimination, often associated with early visual cortex (Bolaños et al., 2024; Schwartz et al., 2002; Ziemba et al., 2016, 2019).

In summary, in **Chapter 3**, I showed that even opaque DNNs can be powerful

tools for generating hypotheses and providing insights into brain computations. The key is to iteratively eliminate confounding factors in control experiments to pinpoint critical architectural and training components. Moreover, researchers should move beyond a single brain alignment score like prediction accuracy and adopt multiple metrics computed on independent neural datasets. Furthermore, neuroscience research could benefit from shifting focus from explaining the best predictive model to using multiple models and alignment metrics to generate and eliminate hypotheses about the structure and function of the brain area under investigation.

Conclusions and future directions

I presented studies investigating the trade-off between prediction and explanation when using complex and opaque DNNs in neuroscience. I argued that researchers should be aware of this trade-off, but that this does not mean sacrificing DNNs' predictive power for simpler, less accurate, but more interpretable models, simply for the sake of interpretability. This applies to both basic neuroscience research and neurological clinical applications. Rigorously validated, complex, opaque DNNs can be useful on their own, as in screening for neurological disorders. They can also be used to induce understanding, either through post-hoc model explanation techniques or by shifting focus to understanding the phenomenon itself, using models as tools for generating questions and hypotheses.

Appendicies

A1 Data management

Ethical Approval

No ethical approval was required as no data collection was performed as part of the presented thesis.

Findability and Accessibility

All the data is stored permanently in servers at The Ernst Struengmann Institute (ESI) for Neuroscience in Cooperation with Max Planck Society and Frankfurt Institute for Advanced Studies (FIAS). All data used in Chapters 2 and Chapter 3 are public and can be accessed online through links in the publications cited. Data used in chapter 4 can be requested from the collaborators cited in the chapter.

Interoperability and Reusability

Data is publicly available and code used for training the models and analyzing the data in chapter 4 is available at **https://github.com/amr-farahat**. The code for chapters 2 and 3 will be available soon in the same github account of the author. All scripts used Python programming language with standard libraries such as Numpy, Scipy, Matplotlib and Tensorflow for constructing and training DNNs.

A2 Abbreviations

AAE = Adversarial Autoencoder

AD = Anomaly Detection

AUC = Area Under the Curve

BCI = Brain Computer Interface

BL = Baseline

CAE = Convolutional Autoencoder

CNN = Convolutional Neural Network

CT = Computer Tomography

DNN = Deep Neural Network

DTI = Diffusion Tensor Imaging

EEG = Electroencephalography

ELU = Exponential Linear Unit

EPG = Epileptogenesis

ERF = Effective Receptive Field

ET = Essential Tremor

fMRI = functional Magnetic Resonance Imaging

GAN = Generative Adversarial Network

GAP = Global Average Pooling

IT = Inferotemporal

LOO = Leave One Out

MDS = Multidimensional Scaling

MIRC = Minimal Recognizable Configuration

MRI = Magnetic Resonance Imaging

MSA = Multiple System Atrophy

NSD = Natural Scenes Dataset

PCA = Principal Component Analysis

PD = Parkinson Disease

PET = Positron Emission Tomography

PPS = Perforant Pathway Stimulation

PSP = Progressive Supranuclear Palsy

RDM = Representational Dissimilarity Matrix

ReLU = Rectified Linear Unit

ROC = Receiver Operating Characteristic

RSA = Representational Similarity Analysis

SPECT = Single-Photon Emission Computed Tomography

Tanh = Hyperbolic Tangent

V1 = Primary Visual Cortex

VO = Ventral Occipital

A3 Bibliography

Abati, D., Porrello, A., Calderara, S., & Cucchiara, R. (2019). Latent space autoregression for novelty detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 481–490.

Abbas, A., & Deny, S. (2023). Progress and limitations of deep networks to recognize objects in unusual poses. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(1), 160–168.

Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical ai. *Nature Medicine*, *28*(9), 1773–1784.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, *31*.

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., & Nguyen, A. (2019). Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4845–4854.

AlKhamissi, B., Tuckute, G., Bosselut, A., & Schrimpf, M. (2024). Brain-like language processing via a shallow untrained multihead attention network. *arXiv preprint arXiv:2406.15109*.

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al. (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

An, J., & Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, *2*(1), 1–18.

Attwell, D., & Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism*, *21*(10), 1133–1145.

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, *18*(10), 1196–1203.

Ayzenberg, V., & Behrmann, M. (2022a). Does the brain's ventral visual pathway compute object shape? *Trends in Cognitive Sciences*.

Ayzenberg, V., & Behrmann, M. (2022b). The dorsal visual pathway represents object-centered spatial relations for object recognition. *Journal of Neuroscience*, *42*(23), 4693–4710.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Baker, N., & Elder, J. H. (2022). Deep learning models fail to capture the configural nature of human shape perception. *iScience*, *25*(9), 104913.

Baker, N., & Kellman, P. J. (2018). Abstract shape representation in human visual perception. *Journal of Experimental Psychology: General*, *147*(9), 1295.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). *Deep convolutional networks do not classify based on global object shape* (Vol. 14). Public Library of Science San Francisco, CA USA.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2020). Local features and global shape information in object classification by deep convolutional neural networks. *Vision Research*, *172*(October 2019), 46–61.

Ballard, D. H. (1987). Modular learning in neural networks. *Proceedings of the sixth National conference on Artificial intelligence-Volume 1*, 279–284.

Barenholtz, E., & Tarr, M. J. (2006). Reconsidering the role of structure in vision. *Psychology of learning and motivation*, *47*, 157–180.

Bengs, M., Behrendt, F., Krüger, J., Opfer, R., & Schlaefer, A. (2021). Three-dimensional deep learning with spatial erasing for unsupervised anomaly segmentation in brain mri. *International journal of computer assisted radiology and surgery*, *16*, 1413–1423.

Bentes, C., Martins, H., Peralta, A. R., Morgado, C., Casimiro, C., Franco, A. C., Fonseca, A. C., Geraldes, R., Canhão, P., Pinho e Melo, T., et al. (2018). Early eeg predicts poststroke epilepsy. *Epilepsia open*, *3*(2), 203–212.

Bianchini, S., Müller, M., & Pelletier, P. (2020). Deep learning in science. *arXiv preprint arXiv:2009.01575*.

Biederman, I. (1987). Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, *94*(2), 115–147.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, *20*(1), 38–64.

Biscione, V., Yin, D., Malhotra, G., Dujmovic, M., Montero, M. L., Puebla, G., Adolfi, F., Heaton, R. F., Hummel, J. E., Evans, B. D., et al. (2024). Mindset: Vision. a toolbox for testing dnns on key psychological experiments. *arXiv preprint arXiv:2404.05290*.

Bolaños, F., Orlandi, J. G., Aoki, R., Jagadeesh, A. V., Gardner, J. L., & Benucci, A. (2024). Efficient coding of natural images in the mouse visual cortex. *Nature Communications*, *15*(1), 2466.

Boon, M., & Knuuttila, T. (2009). Models as epistemic tools in engineering sciences. In *Philosophy of technology and engineering sciences* (pp. 693–726). Elsevier.

Borkakoti, N., & Thornton, J. M. (2023). Alphafold2 protein structure prediction: Implications for drug discovery. *Current opinion in structural biology*, *78*, 102526.

Borowski, J., Zimmermann, R. S., Schepers, J., Geirhos, R., Wallis, T. S. A., Bethge, M., & Brendel, W. (2021). Exemplary natural images explain {cnn} activations better than state-of-the-art feature visualization.

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. *Neuron*, *107*(4), 603–616.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., et al. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385.

Bragin, A., Wilson, C. L., Almajano, J., Mody, I., & Engel Jr, J. (2004). High-frequency oscillations after status epilepticus: epileptogenesis and seizure genesis. *Epilepsia*, *45*(9), 1017–1023.

Brendel, W., & Bethge, M. (2019). Approximating NNs with Bag-of-Local-Features models works surprisingly well on ImageNet. *7th International Conference on Learning Representations, ICLR 2019*, 1–15.

Buettner, R., Frick, J., & Rieg, T. (2019). High-performance detection of epilepsy in seizure-free eeg recordings: A novel machine learning approach using very specific epileptic eeg sub-bands. *ICIS*.

Burnos, S., Hilfiker, P., Sürücü, O., Scholkmann, F., Krayenbühl, N., Grunwald, T., & Sarnthein, J. (2014). Human intracranial high frequency oscillations (HFOs) detected by automatic time-frequency analysis. *PloS one*, *9*(4).

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, *15*(4), e1006897.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Computational Biology*, *10*(12), 1–35.

Calhoun, V. D., Pearlson, G. D., & Sui, J. (2021). Data-driven approaches to neuroimaging biomarkers for neurological and psychiatric disorders: Emerging approaches and examples. *Current opinion in neurology*, *34*(4), 469–479.

Canatar, A., Feather, J., Wakhloo, A., & Chung, S. (2024). A spectral theory of neural prediction and alignment. *Advances in Neural Information Processing Systems*, *36*.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. OUP Oxford.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications biology*, *5*(1), 134.

Chiang, P.-y., Ni, R., Miller, D. Y., Bansal, A., Geiping, J., Goldblum, M., & Goldstein, T. (2022). Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. *The Eleventh International Conference on Learning Representations*.

Chirimuuta, M. (2021). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*, *199*(1), 767–790.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in cognitive sciences*, *23*(4), 305–317.

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*(1), 27755.

Cimpian, A., & Markman, E. M. (2005). The absence of a shape bias in children's word learning. *Developmental Psychology*, *41*(6), 1003.

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2024). A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nature Communications*, *15*(1), 9383.

Costard, L. S., Neubert, V., Venø, M. T., Su, J., Kjems, J., Connolly, N. M., Prehn, J. H., Schratt, G., Henshall, D. C., Rosenow, F., et al. (2019). Electrical stimulation of the ventral hippocampal commissure delays experimental epilepsy and is associated with altered microrna expression. *Brain Stimulation*, *12*(6), 1390–1401.

Couronné, R., Vernhet, P., & Durrleman, S. (2021). Longitudinal self-supervision to disentangle inter-patient variability from disease progression. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, 231–241.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671–675.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D., & DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Bal-

can, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 13073–13087, Vol. 33). Curran Associates, Inc.

Davari, M., Horoi, S., Natik, A., Lajoie, G., Wolf, G., & Belilovsky, E. (2022). Reliability of cka as a similarity measure in deep learning. *The Eleventh International Conference on Learning Representations*.

Del Giudice, M. (2021). Effective dimensionality: A tutorial. *Multivariate behavioral research*, *56*(3), 527–542.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Diesendruck, G., & Bloom, P. (2003). How specific is the shape bias? *Child development*, *74*(1), 168–178.

DiLuca, M., & Olesen, J. (2014). The cost of brain diseases: A burden or a challenge? *Neuron*, *82*(6), 1205–1208.

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., Van Gerven, M. A., Kriegeskorte, N., et al. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*(7), 431–450.

Dong, Y., Ruan, S., Su, H., Kang, C., Wei, X., & Zhu, J. (2022). Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. *Advances in Neural Information Processing Systems*, *35*, 36789–36803.

Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A., & Suarez, H. H. (1995). Recurrent excitation in neocortical circuits. *Science*, *269*(5226), 981–985.

Du, F., Núñez-Ochoa, M. A., Pachitariu, M., & Stringer, C. (2024). Towards a simplified model of primary visual cortex. *bioRxiv*, 2024–06.

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines*, *28*, 645–666.

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai. *Journal of medical ethics*, *47*(5), 329–335.

Dwivedi, K., Bonner, M. F., Cichy, R. M., & Roig, G. (2021). Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS computational biology*, *17*(8), e1009267.

Edelman, S. (1993). Representing three-dimensional objects by sets of activities of receptive fields. *Biological Cybernetics*, *70*(1), 37–45.

Egger, J., Pepe, A., Gsaxner, C., Jin, Y., Li, J., & Kern, R. (2021). Deep learning—a first meta-survey of selected reviews across scientific disciplines, their com-

monalities, challenges and research impact. *PeerJ Computer Science*, *7*, e773.

Elmoznino, E., & Bonner, M. F. (2024). High-performing neural network models of visual cortex benefit from high latent dimensionality. *PLOS Computational Biology*, *20*(1), e1011792.

Engel Jr, J., & Pitkänen, A. (2020). Biomarkers for epileptogenesis and its treatment. *Neuropharmacology*, *167*, 107735.

Evans, B. D., Malhotra, G., & Bowers, J. S. (2022). Biological convolutions improve dnn robustness to noise and generalisation. *Neural Networks*, *148*, 96–110.

Farahat, A., Effenberger, F., & Vinck, M. (2023). A novel feature-scrambling approach reveals the capacity of convolutional neural networks to learn spatial relations. *Neural networks*, *167*, 400–414.

Farahat, A., Reichert, C., Sweeney-Reed, C. M., & Hinrichs, H. (2019). Convolutional neural networks for decoding of covert attention focus and saliency maps for eeg feature visualization. *Journal of neural engineering*, *16*(6), 066010.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, *1*(1), 1–47.

Ferruz, N., Schmidt, S., & Höcker, B. (2022). Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, *13*(1), 4348.

Fisher, R. S. (2015). Redefining epilepsy. *Current opinion in neurology*, *28*(2), 130–135.

Frankle, J., Schwab, D. J., & Morcos, A. S. (2021). Training batchnorm and only batchnorm: On the expressive power of random features in cnns. *International Conference on Learning Representations*.

Fukushima, K., Miyake, S., & Ito, T. (1983). Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (5), 826–834.

Gallicchio, C., & Scardapane, S. (2020). Deep randomized neural networks. *Recent Trends in Learning From Data: Tutorials from the INNS Big Data and Deep Learning Conference (INNSBDDL2019)*, 43–68.

Gao, J., Li, P., Chen, Z., & Zhang, J. (2020). A survey on deep learning for multimodal data fusion. *Neural Computation*, *32*(5), 829–864.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, *33*, 13890–13902.

Geirhos, R., Michaelis, C., Wichmann, F. A., Rubisch, P., Bethge, M., & Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing

shape bias improves accuracy and robustness. *7th International Conference on Learning Representations, ICLR 2019*, (100), 1–22.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, *34*, 23885–23899.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, *31*.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323.

González-Rodríguez, L., & Plasencia-Salgueiro, A. (2021). Uncertainty-aware autonomous mobile robot navigation with deep reinforcement learning. *Deep learning for unmanned systems*, 225–257.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* (**http://www.deeplearningbook.org**). MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Gorenstein, L., Onn, A., Green, M., Mayer, A., Segev, S., & Marom, E. M. (2023). A novel artificial intelligence based denoising method for ultra-low dose ct used for lung cancer screening. *Academic Radiology*, *30*(11), 2588–2597.

Grand, R. L., Mondloch, C. J., Maurer, D., & Brent, H. P. (2004). Impairment in holistic face processing following early visual deprivation. *Psychological science*, *15*(11), 762–768.

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE international conference on acoustics, speech and signal processing*, 6645–6649.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., et al. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471–476.

Grill-spector, K., Kushnir, T., Hendler, T., Edelman, S., Itzchak, Y., & Malach, R. (1998). A Sequence of Object-Processing Stages Revealed by fMRI in the Human Occipital Lobe. *328*, 316–328.

Gu, X., Akoglu, L., & Rinaldo, A. (2019). Statistical analysis of nearest neighbor methods for anomaly detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (pp. 10923–10933, Vol. 32). Curran Associates, Inc.

Güçlü, U., & Van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014.

Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, *7*, 87–93.

Gur, M., Kagan, I., & Snodderly, D. M. (2005). Orientation and direction selectivity of neurons in v1 of alert monkeys: Functional relationships and laminar distributions. *Cerebral Cortex*, *15*(8), 1207–1221.

Gur, M., & Snodderly, D. M. (2007). Direction selectivity in v1 of alert monkeys: Evidence for parallel pathways for motion processing. *The Journal of physiology*, *585*(2), 383–400.

Han, C., Rundo, L., Araki, R., Nagano, Y., Furukawa, Y., Mauri, G., Nakayama, H., & Hayashi, H. (2019). Combining noise-to-image and image-to-image gans: Brain mr image augmentation for tumor detection. *Ieee Access*, *7*, 156966–156977.

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, *25*(1), 65–69.

Hansel, D., & van Vreeswijk, C. (2012). The mechanism of orientation selectivity in primary visual cortex without a functional map. *Journal of Neuroscience*, *32*(12), 4049–4064.

Hansson, O. (2021). Biomarkers for neurodegenerative diseases. *Nature medicine*, *27*(6), 954–963.

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, *2016-Decem*, 770–778.

He, K., Zhang, X., Ren, S., & Sun, J. (2016b). Identity mappings in deep residual networks. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 630–645.

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

Hegdé, J., & Bart, E. (2018). Making expert decisions easier to fathom: On the explainability of visual object recognition expertise. *Frontiers in Neuroscience*, *12*, 670.

Henderson, M. M., Tarr, M. J., & Wehbe, L. (2022). A texture statistics encoding model reveals hierarchical feature selectivity across human visual cortex. *bioRxiv*.

Hendrix, N., Hauber, B., Lee, C. I., Bansal, A., & Veenstra, D. L. (2021). Artificial intelligence in breast cancer screening: Primary care provider preferences. *Journal of the American Medical Informatics Association*, *28*(6), 1117–1124.

Hernström, V., Josefsson, V., Sartor, H., Schmidt, D., Larsson, A.-M., Hofvind, S., Andersson, I., Rosso, A., Hagberg, O., & Lång, K. (2025). Screening performance and characteristics of breast cancer detected in the mammography screening with artificial intelligence trial (masai): A randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study. *The Lancet Digital Health*.

Hewage, P., Behera, A., Trovati, M., Pereira, E., Ghahremani, M., Palmieri, F., & Liu, Y. (2020). Temporal convolutional neural (tcn) network for an effective weather forecasting using time-series data from the local weather station. *Soft Computing*, *24*, 16453–16482.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., & Lerchner, A. (2017). Beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, *3*.

Holzinger, Y., Ullman, S., Harari, D., Behrmann, M., & Avidan, G. (2019). Minimal Recognizable Configurations Elicit Category-selective Responses in Higher Order Visual Cortex. *Journal of Cognitive Neuroscience*, *31*(9), 1354–1367.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359–366.

Hroub, N. A., Alsannaa, A. N., Alowaifeer, M., Alfarraj, M., & Okafor, E. (2024). Explainable deep learning diagnostic system for prediction of lung disease from medical images. *Computers in Biology and Medicine*, *170*, 108012.

Huang, G., Liu, Z., & Weinberger, K. Q. (2016). Densely connected convolutional networks. corr. *arXiv preprint arXiv:1608.06993*.

Hubel, D. H., Wiesel, T. N., et al. (1959). Receptive fields of single neurones in the cat's striate cortex. *J physiol*, *148*(3), 574–591.

Hughes, A. J., Daniel, S. E., Kilford, L., & Lees, A. J. (1992). Accuracy of clinical diagnosis of idiopathic parkinson's disease: A clinico-pathological study of 100 cases. *Journal of neurology, neurosurgery & psychiatry*, *55*(3), 181–184.

Hummel, J., & Biederman, I. (2002). Dynamic Binding in a Neural Network for Shape Recognition. *Psychological review*, *99*(3), 480–517.

Hummel, J. (2013, March). Object recognition. In *The oxford handbook of cognitive psychology*. Oxford University Press.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning*, 448–456.

Jacob, G., Pramod, R., Katti, H., & Arun, S. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature communications*, *12*(1), 1872.

Jagadeesh, A. V., & Gardner, J. L. (2022). Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences*, *119*(17), e2115302119.

Jang, H., & Tong, F. (2021). Convolutional neural networks trained with a developmental sequence of blurry to clear images reveal core differences between face and object processing. *Journal of vision*, *21*(12), 6–6.

Jo, J., & Bengio, Y. (2017). Measuring the tendency of cnns to learn surface statistical regularities. *ArXiv*, *abs/1711.11561*.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of big data*, *6*(1), 1–54.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *nature*, *596*(7873), 583–589.

Kappenman, E. S., & Luck, S. J. (2010). The effects of electrode impedance on data quality and statistical significance in erp recordings. *Psychophysiology*, *47*(5), 888–904.

Kästner, L., & Crook, B. (2023). Don't fear the bogeyman: On why there is no prediction-understanding trade-off for deep learning in neuroscience.

Kawamleh, S. (2023). Against explainability requirements for ethical artificial intelligence in health care. *AI and Ethics*, *3*(3), 901–916.

Kay, K. N. (2018). Principles for models of neural information processing. *NeuroImage*, *180*, 101–109.

Kazemian, A., Elmoznino, E., & Bonner, M. F. (2024). Convolutional architectures are cortex-aligned de novo. *bioRxiv*, 2024–05.

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644.

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

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854–21863.

Kim, K. H., Shim, S., Lim, Y., Jeon, J., Choi, J., Kim, B., & Yoon, A. S. (2019). Rapp: Novelty detection with reconstruction along projection pathway. *International Conference on Learning Representations*.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., & Kim, B. (2019). The (un) reliability of saliency methods. *Explainable AI: Interpreting, explaining and visualizing deep learning*, 267–280.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*.

Kong, N. C., Margalit, E., Gardner, J. L., & Norcia, A. M. (2022). Increasing neural network robustness improves match to macaque v1 eigenspectrum, spatial frequency preference and predictivity. *PLOS Computational Biology*, *18*(1), e1009739.

Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of neural network representations revisited. *International conference on machine learning*, 3519–3529.

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.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249.

Krizhevsky, A., Sutskever, I., & Geoffrey E., H. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25 (NIPS2012)*, 1–9.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep Neural Networks as a Computational Model for Human Shape Sensitivity. *PLOS Computational Biology*, *12*(4), e1004896.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. (2019). Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, *32*.

Kundu, S. (2021). Ai in medicine must be explainable. *Nature medicine*, *27*(8), 1328–1328.

Kwan, P., & Brodie, M. J. (2000). Early identification of refractory epilepsy. *New England Journal of Medicine*, *342*(5), 314–319.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321.

Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., Rahman, M. M., Di Santo, V., Soberanes, D., Feng, G., et al. (2022). Multi-animal pose estimation, identification and tracking with deeplabcut. *Nature Methods*, *19*(4), 496–504.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). Eegnet: A compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, *15*(5), 056013.

Lawler, I., & Sullivan, E. (2021). Model explanation versus model-induced explanation. *Foundations of Science*, *26*, 1049–1074.

Le, H., & Borji, A. (2017). What are the receptive, effective receptive, and projective fields of neurons in convolutional neural networks? *arXiv preprint arXiv:1705.07049*.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

LeCun, Y., Yoshua, B., & Geoffrey, H. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lei, J., & Gillespie, K. (2024). Projected global burden of brain disorders through 2050 (p7-15.001). *Neurology*, *102*(17_supplement_1), 3234.

Levins, R. (1966). The strategy of model building in population biology. *American scientist*, *54*(4), 421–431.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in cognitive sciences*, *6*(1), 9–16.

Lin, W., Lin, W., Chen, G., Zhang, H., Gao, Q., Huang, Y., Tong, T., Du, M., & Initiative, A. D. N. (2021). Bidirectional mapping of brain mri and pet with 3d reversible gan for the diagnosis of alzheimer's disease. *Frontiers in Neuroscience*, *15*, 646013.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, *33*(10), 2017–2031.

Lindsay, G. W., Merel, J., Mrsic-Flogel, T., & Sahani, M. (2021). Divergent representations of ethological visual inputs emerge from supervised, unsupervised, and reinforcement learning. *arXiv preprint arXiv:2112.02027*.

Lipton, Z. C. (2016). The mythos of model interpretability. corr abs/1606.03490 (2016). *arXiv preprint arXiv:1606.03490*, *2*.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, *49*(1), 15–21.

Long, B., Yu, C. P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(38), E9015–E9024.

Löscher, W. (2019). The holy grail of epilepsy prevention: Preclinical approaches to antiepileptogenic treatments. *Neuropharmacology*, *167*, 107605.

Lu, D., Bauer, S., Neubert, V., Costard, L. S., Rosenow, F., & Triesch, J. (2020a). Staging epileptogenesis with deep neural networks. *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 1–10.

Lu, D., Bauer, S., Neubert, V., Costard, L. S., Rosenow, F., & Triesch, J. (2020b). Towards early diagnosis of epilepsy from eeg data. *Machine Learning for Healthcare Conference*, 80–96.

Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, *29*(2), 102–127.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., & Frey, B. (2015). Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Malhotra, G., Dujmović, M., Hummel, J., & Bowers, J. S. (2022, December). *Human shape representations are not an emergent property of learning to classify objects* (preprint).

Malhotra, G., Evans, B. D., & Bowers, J. S. (2020). Hiding a plane with a pixel: Examining shape-bias in cnns and the benefit of building in biological constraints. *Vision Research*, *174*, 57–68.

Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., & Shroff, G. (2016). Lstm-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*.

Manole, I., Butacu, A.-I., Bejan, R. N., & Tiplica, G.-S. (2024). Enhancing dermatological diagnostics with efficientnet: A deep learning approach. *Bioengineering*, *11*(8), 810.

Margalit, E., Biederman, I., Tjan, B. S., & Shah, M. P. (2017). What is actually affected by the scrambling of objects when localizing the lateral occipital complex? *Journal of cognitive neuroscience*, *29*(9), 1595–1604.

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Jia, Y., Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, ... Xiaoqiang Zheng. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems (Software available from tensorflow.org). **https://www.tensorflow.org/**

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). Deeplabcut: Markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, *21*(9), 1281–1289.

Matthewson, J. (2011). Trade-offs in model-building: A more target-oriented approach. *Studies in History and Philosophy of Science Part A*, *42*(2), 324–333.

Mazurek, M., Kager, M., & Van Hooser, S. D. (2014). Robust quantification of orientation selectivity and direction selectivity. *Frontiers in neural circuits*, *8*, 92.

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

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., et al. (2020). International evaluation of an ai system for breast cancer screening. *Nature*, *577*(7788), 89–94.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8), e2011417118.

Miao, H.-Y., & Tong, F. (2024). Convolutional neural network models applied to neuronal responses in macaque v1 reveal limited nonlinear processing. *Journal of Vision*, *24*(6), 1–1.

Milikovsky, D. Z., Weissberg, I., Kamintsky, L., Lippmann, K., Schefenbauer, O., Frigerio, F., Rizzi, M., Sheintuch, L., Zelig, D., Ofer, J., et al. (2017). Electrocorticographic dynamics as a novel biomarker in five models of epileptogenesis. *Journal of Neuroscience*, *37*(17), 4450–4461.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, *518*(7540), 529–533.

Moshé, S. L., Perucca, E., Ryvlin, P., & Tomson, T. (2015). Epilepsy: new advances. *The Lancet*, *385*(9971), 884–898.

Nazir, M., Shakil, S., & Khurshid, K. (2021). Role of deep learning in brain tumor detection and classification (2015 to 2020): A review. *Computerized medical imaging and graphics*, *91*, 101940.

Nejedly, P., Kremen, V., Sladky, V., Nasseri, M., Guragain, H., Klimes, P., Cimbalnik, J., Varatharajah, Y., Brinkmann, B. H., & Worrell, G. A. (2019). Deep-learning for seizure forecasting in canines with epilepsy. *Journal of neural engineering*, *16*(3), 036031.

Nguyen, B., Feldman, A., Bethapudi, S., Jennings, A., & Willcocks, C. G. (2021). Unsupervised region-based anomaly detection in brain mri with adversarial image inpainting. *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, 1127–1131.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4).

Norwood, B. A., Bauer, S., Wegner, S., Hamer, H. M., Oertel, W. H., Sloviter, R. S., & Rosenow, F. (2011). Electrical stimulation-induced seizures in rats: A "dose-response" study on resultant neurodegeneration. *Epilepsia*, *52*(9), e109–e112.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), e7.

Pahwa, R., & Lyons, K. E. (2010). Early diagnosis of parkinson's disease: Recommendations from diagnostic clinical guidelines. *Am J Manag Care*, *16*(4), 94–99.

Pattadkal, J. J., Mato, G., van Vreeswijk, C., Priebe, N. J., & Hansel, D. (2018). Emergent orientation selectivity from random networks in mouse visual cortex. *Cell reports*, *24*(8), 2042–2050.

Peissig, J. J., & Tarr, M. J. (2007). Visual object recognition: Do we know more now than we did 20 Years ago? *Annual Review of Psychology*, *58*, 75–96.

Pereira, A. F., James, K. H., Jones, S. S., & Smith, L. B. (2010). Early biases and developmental changes in self-generated object views. *Journal of vision*, *10*(11), 22–22.

Pereira, A. F., & Smith, L. B. (2009). Developmental changes in visual object recognition between 18 and 24 months of age. *Developmental Science*, *12*(1), 67–80.

Pidhorskyi, S., Almohsen, R., & Doretto, G. (2018). Generative probabilistic novelty detection with adversarial autoencoders. *Advances in Neural Information Processing Systems*, *31*, 6822–6833.

Pinaya, W. H., Tudosiu, P.-D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., Ourselin, S., & Cardoso, M. J. (2022). Brain imaging generation with latent diffusion models. *MICCAI Workshop on Deep Generative Models*, 117–126.

Pitkänen, A., & Engel, J. (2014). Past and present definitions of epileptogenesis and its biomarkers. *Neurotherapeutics*, *11*(2), 231–241.

Pitkänen, A., Löscher, W., Vezzani, A., Becker, A. J., Simonato, M., Lukasiuk, K., Gröhn, O., Bankstahl, J. P., Friedman, A., Aronica, E., et al. (2016). Advances in the development of biomarkers for epilepsy. *The Lancet Neurology*, *15*(8), 843–856.

Portelance, E., Frank, M. C., Jurafsky, D., Sordoni, A., & Laroche, R. (2021). The emergence of the shape bias results from communicative efficiency. *arXiv preprint arXiv:2109.06232*.

Rahimi, A., & Recht, B. (2007). Random features for large-scale kernel machines. *Advances in neural information processing systems*, *20*.

Rahimi, A., & Recht, B. (2008a). Uniform approximation of functions with random bases. *2008 46th annual allerton conference on communication, control, and computing*, 555–561.

Rahimi, A., & Recht, B. (2008b). Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, *21*.

Rainer, G., Augath, M., Trinath, T., & Logothetis, N. K. (2002). The effect of image scrambling on visual cortical BOLD activity in the anesthetized monkey. *NeuroImage*, *16*(3 I), 607–616.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*(33), 7255–7269.

Rajesh, N., Jacob, G., & Arun, S. (2024). Brain-like emergent properties in deep networks: Impact of network architecture, datasets and training. *arXiv preprint arXiv:2411.16326*.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Ramasinghe, S., MacDonald, L., Farazi, M., Saratchandran, H., & Lucey, S. (2022). How you start matters for generalization. *arXiv preprint arXiv:2206.08558*.

Ren, F., Ding, X., Zheng, M., Korzinkin, M., Cai, X., Zhu, W., Mantsyzov, A., Aliper, A., Aladinskiy, V., Cao, Z., et al. (2023). Alphafold accelerates artificial intelligence powered drug discovery: Efficient discovery of a novel cdk20 small molecule inhibitor. *Chemical science*, *14*(6), 1443–1452.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

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

Ritter, S., Barrett, D. G., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 2940–2949.

Rizzi, M., Brandt, C., Weissberg, I., Milikovsky, D. Z., Pauletti, A., Terrone, G., Salamone, A., Frigerio, F., Löscher, W., Friedman, A., et al. (2019). Changes of dimension of EEG/ECoG nonlinear dynamics predict epileptogenesis and therapy outcomes. *Neurobiology of disease*, *124*, 373–378.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2021). High-resolution image synthesis with latent diffusion models. 2022 ieee. *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, *1*.

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

Rosenfeld, A., & Tsotsos, J. K. (2019). Intriguing properties of randomly weighted networks: Generalizing while learning next to nothing. *2019 16th conference on computer and robot vision (CRV)*, 9–16.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206–215.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., & Kloft, M. (2018). Deep one-class classification. *International Conference on Machine Learning*, 4393–4402.

Ruff, L., Zemlyanskiy, Y., Vandermeulen, R., Schnake, T., & Kloft, M. (2019). Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4061–4071.

Rumelhart, D. E., McClelland, J. L., Group, P. R., et al. (1986). *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press.

Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it. *Journal of Neuroscience*, *30*(39), 12978–12995.

Sangkloy, P., Burnell, N., Ham, C., & Hays, J. (2016). The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Trans. Graph.*, *35*(4).

Saxe, A. M., Koh, P. W., Chen, Z., Bhand, M., Suresh, B., & Ng, A. Y. (2011). On random weights and unsupervised feature learning. *Icml*, *2*(3), 6.

Scardapane, S., & Wang, D. (2017). Randomness in neural networks: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(2), e1200.

Schaeffer, R., Khona, M., Chandra, S., Ostrow, M., Miranda, B., & Koyejo, S. (2024). Position: Maximizing neural regression scores may not identify good models of the brain. *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*.

Schaeffer, R., Khona, M., & Fiete, I. (2022). No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. *Advances in neural information processing systems*, *35*, 16052–16067.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., & Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. *International conference on information processing in medical imaging*, 146–157.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, *13*(7), 1443–1471.

Schreyer, M., Sattarov, T., Schulze, C., Reimer, B., & Borth, D. (2019). Detection of accounting anomalies in the latent space using adversarial autoencoder neural networks. *arXiv preprint arXiv:1908.00734*.

Schwartz, S., Maquet, P., & Frith, C. (2002). Neural correlates of perceptual learning: A functional mri study of visual texture discrimination. *Proceedings of the National Academy of Sciences*, *99*(26), 17137–17142.

Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S. E., & Van Gerven, M. (2018). Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage*, *180*, 253–266.

Seiler, M., & Ritter, K. (2024). Pioneering new paths: The role of generative modelling in neurological disease research. *Pflügers Archiv-European Journal of Physiology*, 1–19.

Self, M. W., van Kerkoerle, T., Supèr, H., & Roelfsema, P. R. (2013). Distinct roles of the cortical layers of area v1 in figure-ground segregation. *Current biology*, *23*(21), 2121–2129.

Sevilla, J., & Hegde, J. (2017). Deep visual patterns are informative to practicing radiologists in mammograms in diagnostic tasks. *Journal of Vision*, *17*(10), 90–90.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., & Netrapalli, P. (2020). The pitfalls of simplicity bias in neural networks. *arXiv preprint arXiv:2006.07710*.

Shen, L., Li, Z., & Kwok, J. (2020). Timeseries anomaly detection using temporal hierarchical one-class network. *NIPS 2020*.

Shoeibi, A., Khodatars, M., Jafari, M., Ghassemi, N., Moridian, P., Alizadehsani, R., Ling, S. H., Khosravi, A., Alinejad-Rokny, H., Lam, H.-K., et al. (2023). Diagnosis of brain diseases in fusion of neuroimaging modalities using deep learning: A review. *Information Fusion*, *93*, 85–117.

Shoeibi, A., Moridian, P., Khodatars, M., Ghassemi, N., Jafari, M., Alizadehsani, R., Kong, Y., Gorriz, J. M., Ramírez, J., Khosravi, A., et al. (2022). An overview of deep learning techniques for epileptic seizures detection and prediction based on neuroimaging modalities: Methods, challenges, and future works. *Computers in biology and medicine*, *149*, 106053.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484–489.

Simoncelli, E. P., Paninski, L., Pillow, J., Schwartz, O., et al. (2004). Characterization of neural responses with stochastic stimuli. *The cognitive neurosciences*, *3*(327-338), 1.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Singer, J. J., Seeliger, K., Kietzmann, T. C., & Hebart, M. N. (2022). From photos to sketches-how humans and deep neural networks process objects across different levels of visual abstraction. *Journal of vision*, *22*(2), 4–4.

Smith, L. B. (2009). From fragments to geometric shape: Changes in visual object recognition between 18 and 24 months. *Current Directions in Psychological Science*, *18*(5), 290–294.

Soni, A., Srivastava, S., Khosla, M., & Kording, K. P. (2024). Conclusions about neural network to brain alignment are profoundly impacted by the similarity measure. *bioRxiv*, 2024–08.

Soska, K. C., & Johnson, S. P. (2008). Development of three-dimensional object completion in infancy. *Child development*, *79*(5), 1230–1236.

Stahlschmidt, S. R., Ulfenborg, B., & Synnergren, J. (2022). Multimodal deep learning for biomedical data fusion: A review. *Briefings in Bioinformatics*, *23*(2), bbab569.

Stirling, R. E., Cook, M. J., Grayden, D. B., & Karoly, P. J. (2021). Seizure forecasting and cyclic control of seizures. *Epilepsia*, *62*, S2–S14.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex

well, after training and fitting. *Journal of cognitive neuroscience*, *33*(10), 2044–2064.

Straka, M. M., Shafer, B., Vasudevan, S., Welle, C., & Rieth, L. (2018). Characterizing longitudinal changes in the impedance spectra of in-vivo peripheral nerve electrodes. *Micromachines*, *9*(11), 587.

Stringer, C., & Pachitariu, M. (2024). Analysis methods for large-scale neuronal recordings. *Science*, *386*(6722), eadp7429.

Suchowersky, O., Reich, S., Perlmutter, J., Zesiewicz, T., Gronseth, G., & Weiner, W. (2006). Practice parameter: Diagnosis and prognosis of new onset parkinson disease (an evidence-based review)(retired) report of the quality standards subcommittee of the american academy of neurology. *Neurology*, *66*(7), 968–975.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International conference on machine learning*, 3319–3328.

Szegedy, C. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the inception architecture for computer vision. 2015. *arXiv preprint arXiv:1512.00567*.

Tanaka, J. W., & Simonyi, D. (2016). The "parts and wholes" of face recognition: A review of the literature. *The Quarterly Journal of Experimental Psychology*, *69*(10), 1876–1889.

Tartaglini, A. R., Vong, W. K., & Lake, B. M. (2022). A Developmentally-Inspired Examination of Shape versus Texture Bias in Machines. *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, 8.

Tax, D. M., & Duin, R. P. (2004). Support vector data description. *Machine learning*, *54*(1), 45–66.

Teney, D., Nicolicioiu, A. M., Hartmann, V., & Abbasnejad, E. (2024). Neural redshift: Random networks are not random functions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4786–4796.

Ting, D. S. W., Cheung, C. Y.-L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I. Y., Lee, S. Y., et al. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama*, *318*(22), 2211–2223.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016). Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(10), 2744–2749.

Ullman, S., Sali, E., & Vidal-Naquet, M. (2001). A fragment-based approach to object representation and classification. *Visual Form 2001*, 85–100.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, *5*(7), 682–687.

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2018). Deep image prior. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9446–9454.

Uyttenhove, T., Maes, A., Van Steenkiste, T., Deschrijver, D., & Dhaene, T. (2020). Interpretable epilepsy detection in routine, interictal eeg data using deep learning. *Machine Learning for Health*, 355–366.

Valliani, A. A.-A., Ranti, D., & Oermann, E. K. (2019). Deep learning and neurology: A systematic review. *Neurology and therapy*, *8*(2), 351–365.

van der Ouderaa, T. F., & Worrall, D. E. (2019). Reversible gans for memory-efficient image-to-image translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4720–4728.

Vezoli, J., Magrou, L., Goebel, R., Wang, X.-J., Knoblauch, K., Vinck, M., & Kennedy, H. (2021). Cortical hierarchy, dual counterstream architecture and the importance of top-down generative networks. *Neuroimage*, *225*, 117479.

Vinck, M., Uran, C., Spyropoulos, G., Onorato, I., Broggini, A. C., Schneider, M., & Canales-Johnson, A. (2023). Principles of large-scale neural interactions. *Neuron*, *111*(7), 987–1002.

Vogels, R. (1999). Effect of image scrambling on inferior temporal cortical responses. *NeuroReport*, *10*(9), 1811–1816.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in neurobiology*, *51*(2), 167–194.

Wang, R., Bashyam, V., Yang, Z., Yu, F., Tassopoulou, V., Chintapalli, S. S., Skampardoni, I., Sreepada, L. P., Sahoo, D., Nikita, K., et al. (2023). Applications of generative adversarial networks in neuroimaging and clinical neuroscience. *Neuroimage*, *269*, 119898.

Wang, W., Huang, Y., Wang, Y., & Wang, L. (2014). Generalized autoencoder: A neural network framework for dimensionality reduction. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 490–497.

Wichmann, F. A., & Geirhos, R. (2023). Are deep neural networks adequate behavioral models of human visual perception? *Annual Review of Vision Science*, *9*(1), 501–524.

Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., Kamdar, F., Glasser, M. F., Hochberg, L. R., Druckmann, S., et al. (2023). A high-performance speech neuroprosthesis. *Nature*, *620*(7976), 1031–1036.

Willmore, B., Prenger, R. J., Wu, M. C.-K., & Gallant, J. L. (2008). The berkeley wavelet transform: A biologically inspired orthogonal wavelet transform. *Neural computation*, *20*(6), 1537–1564.

Willsey, M. S., Shah, N. P., Avansino, D. T., Hahn, N. V., Jamiolkowski, R. M., Kamdar, F. B., Hochberg, L. R., Willett, F. R., & Henderson, J. M. (2025). A high-performance brain–computer interface for finger decoding and quadcopter game control in an individual with paralysis. *Nature Medicine*, 1–9.

Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2019). Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(9), 2251–2265.

Xua, B., & Yang, G. (2024). Interpretability research of deep learning: A literature survey. *Information Fusion*, 102721.

Yamins, D. L. K., 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*, *111*(23), 8619–8624.

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

Yang, Z., Nasrallah, I. M., Shou, H., Wen, J., Doshi, J., Habes, M., Erus, G., Abdulkadir, A., Resnick, S. M., Albert, M. S., et al. (2021). A deep learning framework identifies dimensional representations of alzheimer's disease from brain structure. *Nature communications*, *12*(1), 7065.

Yee, M., Jones, S. S., & Smith, L. B. (2012). Changes in visual object recognition precede the shape bias in early noun learning. *Frontiers in Psychology*, *3*(DEC), 1–13.

Yoon, C. H., Torrance, R., & Scheinerman, N. (2022). Machine learning in medicine: Should the pursuit of enhanced interpretability be abandoned? *Journal of Medical Ethics*, *48*(9), 581–585.

Yoshida, H., & Smith, L. B. (2003). Shifting ontological boundaries: How japanese- and english-speaking children generalize names for animals and artifacts. *Developmental Science*, *6*(1), 1–17.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, 818–833.

Zhou, B., Liu, S., Hooi, B., Cheng, X., & Ye, J. (2019). BeatGAN: Anomalous Rhythm Detection using Adversarially Generated Time Series. *IJCAI*, 4433–4439.

Zhou, C., & Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 665–674.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118.

Ziemba, C. M., Freeman, J., Movshon, J. A., & Simoncelli, E. P. (2016). Selectivity and tolerance for visual texture in macaque v2. *Proceedings of the National Academy of Sciences*, *113*(22), E3140–E3149.

Ziemba, C. M., Perez, R. K., Pai, J., Kelly, J. G., Hallum, L. E., Shooner, C., & Movshon, J. A. (2019). Laminar differences in responses to naturalistic texture in macaque v1 and v2. *Journal of Neuroscience*, *39*(49), 9748–9756.

Zimmermann, R. S., Borowski, J., Geirhos, R., Bethge, M., Wallis, T., & Brendel, W. (2021). How well do feature visualizations support causal understanding of cnn activations? *Advances in Neural Information Processing Systems*, *34*, 11730–11744.

A4 Dutch Summary

Deze dissertatie onderzoekt de toepassing van diepe neurale netwerken (DNN's) in de neurowetenschappen, waarbij de nadruk ligt op de afweging tussen voorspellend vermogen en verklarend inzicht. Het onderzoek benadrukt een aantal belangrijke bevindingen. Ten eerste heeft een nieuwe, niet gesuperviseerde leerbenadering met behulp van een Adversarial Autoencoder (AAE) met succes vroege tekenen van epileptogenese gedetecteerd in EEG-opnamen, waarmee het potentieel van DNN's voor proactieve diagnose wordt aangetoond, zelfs wanneer gelabelde gegevens schaars zijn. Ten tweede toonden experimenten met Convolutionele Neurale Netwerken (CNN's) aan dat ze ruimtelijke relaties tussen kenmerken gebruiken voor objectherkenning, vooral bij textuurloze beelden, maar dat ze de algemene vorm niet op een holistische manier vastleggen. Deze bevinding helpt ons uit te zoeken hoe deze modellen "zien" en toont een praktische toepassing van post-hoc modelinzicht. Ten derde onderzocht het onderzoek wat belangrijker is voor het nabootsen van verschillende delen van de visuele cortex: de structuur van het netwerk of de training die het krijgt. Voor het vroege visuele gebied (V1) was de inherente complexiteit van zelfs willekeurig geïnitialiseerde netwerken verrassend effectief in het voorspellen van reacties, terwijl gebieden op een hoger niveau (IT, VO) specifieke training nodig hadden. Dit onderscheid laat zien hoe verschillende hersengebieden kunnen vertrouwen op verschillende computationele strategieën, een inzicht dat werd verkregen door het model vergelijkend te gebruiken en zonder volledige uitleg van het model zelf. Concluderend stelt deze dissertatie dat DNN's waardevolle hulpmiddelen zijn in de neurowetenschappen, niet alleen voor voorspellingen, maar ook voor het verkrijgen van meer inzicht. Hoewel het onderzoek erkent dat complexe modellen "zwarte dozen" kunnen zijn, benadrukt het dat zorgvuldige validatie ons in staat stelt om ze effectief te gebruiken. Deze modellen kunnen krachtige voorspellers zijn (zoals in vroege ziektedetectie) en, wat cruciaal is, hulpmiddelen die ons helpen nieuwe vragen te genereren en ons begrip van de hersenen te vergroten, waarbij onze focus verschuift van het volledig transparant maken van de modellen zelf naar het beter begrijpen van de hersenen.

A5 Acknowledgement I Dankwoord

I am thankful to my supervisor, Martin Vinck, for providing me with the opportunity to conduct this research in his lab and for his guidance in developing my research skills. I would also like to express my appreciation to Jochen Triesch and Felix Effenberger for their insightful contributions. My sincere thanks go to all the members of the Vinck lab and the wider ESI community for making the past few years in Frankfurt enjoyable, particularly during the challenging times of the COVID-19 pandemic. I also cherish the time spent with international colleagues at conferences and summer schools around the world. I am especially grateful to the organizers, teaching assistants, and fellow students of the IBRO-Simons Imbizo summer school for creating unforgettable memories and fostering what I hope will be enduring friendships. Finally, I would like to express my deepest gratitude to my family for their unwavering support and belief in me.

A6 Curriculum Vitae

Work and Research Experience

Neuroradiology Resident Doctor — Neuroradiology Department – Otto von Guericke University Hospital, Magdeburg, Germany. 2019 — 2019

Research Assistant — Neurology Department – Otto von Guericke University Hospital, Magdeburg, Germany. 2017 — 2019

Intern Doctor and General Practitioner — Mansoura University Hospitals, Egypt. 2014 — 2015

Education

Ph.D. Candidate in Ernst Strüngmann Institute for Neuroscience in Cooperation with Max Planck Society, Frankfurt, Germany and Donders Centre for Neuroscience, Department of Neurophysics, Radboud University Nijmegen, the Netherlands. 2020 — 2025

M.Sc. Integrative Neurosciences in Otto von Guericke University, Magdeburg, Germany. 2015 — 2018
Thesis title: *Deep Learning for EEG Decoding and Automatic Feature Discovery*.

MBBCh (Bachelor of Medicine, Bachelor of Surgery) in Mansoura University, Egypt. 2007 — 2014

Peer-Reviewed Publications

- Voegtle, A., Terzic, L., **Farahat, A.**, Hartong, N., Galazky, I., Hinrichs, H., ... & Sweeney-Reed, C. M. (2024). Ventrointermediate thalamic stimulation improves motor learning in humans. Communications Biology, 7(1), 798.

- **Farahat, A.**, Effenberger, F., & Vinck, M. (2023). A novel feature-scrambling approach reveals the capacity of convolutional neural networks to learn spatial relations. Neural networks, 167, 400-414.
- Terzic, L., Voegtle, A., **Farahat, A.**, Hartong, N., Galazky, I., Nasuto, S. J., ... & Sweeney-Reed, C. M. (2022). Deep brain stimulation of the ventrointermediate nucleus of the thalamus to treat essential tremor improves motor sequence learning. Human Brain Mapping, 43(15), 4791-4799.
- Voegtle, A., Terlutter, C., Nikolai, K., **Farahat, A.**, Hinrichs, H., & Sweeney-Reed, C. M. (2023). Suppression of motor sequence learning and execution through anodal cerebellar transcranial electrical stimulation. The Cerebellum, 22(6), 1152-1165.
- **Farahat, A.**, Lu, D., Bauer, S., Neubert, V., Costard, L. S., Rosenow, F., & Triesch, J. (2022). Diagnosing epileptogenesis with deep anomaly detection. In Machine Learning for Healthcare Conference (pp. 325-342). PMLR.
- **Farahat, A.**, Reichert, C., Sweeney-Reed, C. M., & Hinrichs, H. (2019). Convolutional neural networks for decoding of covert attention focus and saliency maps for EEG feature visualization. Journal of neural engineering, 16(6), 066010.

Preprints

- **Farahat, A.**, & Vinck, M. (2025). Neural responses in early, but not late, visual cortex are well predicted by random-weight CNNs with sufficient model complexity. bioRxiv, 2025-02.

Awards

- **Poster prize** during the Neurorad Conference in Frankfurt 2019 awarded from the German and Austrian Societies of Neuroradiology. 2019
- **Scholarship** granted from Leibniz institute for Neurobiology in the context of the project "SFB 779 – MGK" for scoring the best grades in the integrative neuroscience master program. 2016— 2017

A7 Donders Graduate School

For a successful research Institute, it is vital to train the next generation of scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School in 2009. The mission of the Donders Graduate School is to guide our graduates to become skilled academics who are equipped for a wide range of professions. To achieve this, we do our utmost to ensure that our PhD candidates receive support and supervision of the highest quality.

Since 2009, the Donders Graduate School has grown into a vibrant community of highly talented national and international PhD candidates, with over 500 PhD candidates enrolled. Their backgrounds cover a wide range of disciplines, from physics to psychology, medicine to psycholinguistics, and biology to artificial intelligence. Similarly, their interdisciplinary research covers genetic, molecular, and cellular processes at one end and computational, system-level neuroscience with cognitive and behavioural analysis at the other end. We ask all PhD candidates within the Donders Graduate School to publish their PhD thesis in de Donders Thesis Series. This series currently includes over 750 PhD theses from our PhD graduates and thereby provides a comprehensive overview of the diverse types of research performed at the Donders Institute. A complete overview of the Donders Thesis Series can be found on our website: **https://www.ru.nl/donders/donders-series**

The Donders Graduate School tracks the careers of our PhD graduates carefully. In general, the PhD graduates end up at high-quality positions in different sectors, for a complete overview see **https://www.ru.nl/donders/destination-our-former-phd**. A large proportion of our PhD alumni continue in academia (>50%). Most of them first work as a postdoc before growing into more senior research positions. They work at top institutes worldwide, such as University of Oxford, University of Cambridge, Stanford University, Princeton University, UCL London, MPI Leipzig, Karolinska Institute, UC Berkeley, EPFL Lausanne, and many others. In addition, a large group of PhD graduates continue in clinical positions, sometimes combining it with academic research. Clinical positions can be divided into medical doctors, for instance, in genetics, geriatrics, psychiatry, or neurology, and in psychologists, for instance as healthcare psychologist, clinical neuropsychologist, or clinical psychologist. Furthermore, there are PhD graduates who continue to work as researchers outside academia, for instance at non-profit or government organiza-

tions, or in pharmaceutical companies. There are also PhD graduates who work in education, such as teachers in high school, or as lecturers in higher education. Others continue in a wide range of positions, such as policy advisors, project managers, consultants, data scientists, web- or software developers, business owners, regulatory affairs specialists, engineers, managers, or IT architects. As such, the career paths of Donders PhD graduates span a broad range of sectors and professions, but the common factor is that they almost all have become successful professionals.

For more information on the Donders Graduate School, as well as past and upcoming defences please visit: **http://www.ru.nl/donders/graduate-school/phd/**

DONDERS INSTITUTE

Radboud University

Radboudumc