École Polytechnique Fédérale de Lausanne
School of Computer Science



Master Thesis
in
Computer Science

Bridging Artificial and Primate Vision:
The Impact of Visual Angle, Scene Context and IT-Alignment

Sara Djambazovska

*C*arried out in the Kreiman lab
at Harvard Medical School
Under the supervision of Prof. Gabriel Kreiman

Evaluated by Prof. Pavan Ramdya
Neuroengineering Laboratory
EPFL

*B*oston, MA, USA, August 2023

Bridging Artificial and Primate Vision: The Impact of Visual Angle, Scene Context and IT-Alignment

Abstract

The ventral visual stream is a biological neural network that plays a critical role in object recognition and scene understanding. Convolutional neural networks (CNNs) developed to mirror its capabilities, achieve impressive performance in these tasks. Yet, easily fooled by simple image distortions, these artificial models are not perfect replicas of the primate ventral stream. Assessing the similarity between these two systems of vision is an important step toward building more brain-like artificial models. Current methods use the neural predictivity metric, or the ability to predict neural responses given the internal activations in an artificial neural network, when both are presented with the same stimulus. However, rigorously controlled predictivity analyses are rare. The models are often evaluated with neurobehavioral datasets where the exact visual angle of stimuli presentation and the image-context varies widely across studies. In this work we systematically compare the late neural representations of the macaque ventral visual stream, recorded by chronically implanted multi-electrodes in the inferior temporal (IT) cortex, and deep CNNs while varying the image-context and field of view (FOV), to quantify their respective effects. We show that artificial models have a preferred visual angle range of 7-11 degrees that best aligns them with the neural data. We explore the similarities and differences in how artificial and biological neural networks process visual information, particularly in handling natural scenes in varying context during behavior and physiology. We show that the neural alignment of the late representations of the two models is optimal when removing the context from an image, and impaired when placing the object in incongruent context. Moreover, we reveal a substantial explanatory gap for these models for image-level primate behavioral consistency across contextual manipulations. We show that the neural population activity in IT from a context-naive macaque, is able to accurately predict these behavioral patterns, suggesting that IT-aligned models would better approximate primate behavior. Finally, we introduce a data-driven approach, leveraging large-scale multi-electrode recordings to align the artificial model's late feature space to that of the macaque IT population. Our resulting model, validated on new sessions and held-out animals across large image sets, showed an increased IT-likeness, generalization to new subjects, and remarkably, increased robustness to different image perturbations and adversarial attacks. Overall, our results suggest that with appropriate adjustments and careful comparisons, modern computer vision models can come closer to replicating the intricacies of the primate visual system, and in turn deepen our understanding of visual processing in the brain.

Contents

List of figures

Acknowledgments

I would like to express my sincere gratitude to professor Gabriel Kreiman for giving me the opportunity to complete my Master thesis in his lab and the welcoming environment he created, allowing me to explore different projects. His exceptional guidance and advice during our weekly meetings were vital for accomplishing my work presented here. I am also deeply thankful to professor Kohitij Kar, who co-supervised two of my projects, for providing me with the neural data, the knowledge and skills to approach it and invaluable supervision and support. I was fortunate enough to also have some of my work overseen by Will Xiao, a PhD at the Kreiman lab, who generously provided me with the large neural datasets from the Livingstone lab and brilliant advice and feedback along the development of my project, for which I am truly grateful.

My gratitude goes to the Bertarelli Foundation for supporting my research throughout this year. I would also like to thank my EPFL supervisor, professor Pavan Ramdya who kindly agreed to oversee this project.

I would like to extend my warmest gratitude to all my colleagues at the Kreiman lab, who made my working environment that much more enjoyable, especially Dianna Hidalgo, Elisa Pavarino and Leonardo Pollina, for the many insightful and joyful discussions, their incredible kindness and friendship.

Finally, my deepest thanks is to my sister who has always been there, I owe much of my journey to her unconditional support.

1 Introduction

The fields of artificial intelligence and neuroscience have a rich and deeply connected history,[43,44,124,90], with each field informing and advancing the other. The deepening of our knowledge of biological neural networks (BNNs) has influenced the development of artificial models, while advances in deep learning have contributed to a better understanding of how the brain processes information. Historically, intelligence has been defined by looking at the capabilities of complex biological beings, primarily humans. As a consequence, research on artificial intelligence has focused on systems that can emulate human intelligence, building Artificial General Intelligence[40] (AGI) or "strong AI". It comes naturally that the development of AI has taken inspiration from the brain's complex neural mechanisms. Artificial neural networks (ANNs) however, are an extreme abstraction of BNNs. The first generation of ANNs developed in the 1950s used perceptrons[108], abstract mathematical models of biological neurons. Then, in the 1980s, the backpropagation algorithm[70] was developed, which allowed neural networks to learn from data and improve their performance over time. This

led to the emergence of a new wave of AI capable of intelligent skills such as speech [9] and image recognition [82]. More recently, the development of attention networks [125] was motivated by the observation that human brains "attend to" certain parts of inputs when processing large amounts of information. Similarly, the emergence of spiking neural networks [100] can be attributed to the concept of approximating stochastic potential-based communications that occur between neurons.

Conversely, the development of artificial intelligence has facilitated advancements in the realm of neuroscience, with AI rapidly emerging as an essential tool in neuroscience research. The core strength of artificial intelligence lies in its capability to analyze vast amounts of complex data and extract meaningful patterns from within it. Artificial models have been widely used in neuroscience to analyze large-scale neuroimaging data, thereby facilitating the timely prediction and detection of psychiatric disorders [117]. The advent of Brain Computer Interfaces (BCIs) has enabled the possibility to link artificial systems with the brain, showing the capability of these systems to decode neurological signals and issue directives to devices such as robotic arms that enable movement for disabled patients [133]. AI has also revolutionized the field of connectomics - the study of neural connections in the brain. By automating and refining the process of mapping intricate neural networks, deep learning algorithms have expedited the identification and analysis of neural pathways [63]. Additionally, artificial neural networks help examine neuroscience hypotheses by simulating complex neural circuits [39]. In particular, deep learning has found application in the modeling of the primate cortex's convolutional layers and recurrent connections, responsible for important functions such as visual processing [129], memory [2], and motor control [93]. Moreover, these deep models have been used to assess the structural features of the visual system of the brain and precisely predict neural activity patterns [130,122,42,114,30]. Nevertheless, thorough and rigorously controlled comparisons of artificial and primate vision remain scarce.

Here, we will build on this interdisciplinary history by further investigating the concept of neural and behavioral predictivity of deep ANNs, measuring their similarity to the primate ventral visual stream. In particular, we probe the effects of scene context, visual angle, and neural data tuning on the alignment between the primate and artificial models of vision.

1.1 The structure of this thesis

The research presented in this thesis was done in the field of computational neuroscience, at the intersection of neuroscience and computer science.
Following the introduction presented above, chapter 2 is a review of the literature regarding this thesis' main pillars:

1. **the ventral visual stream** - the critical circuitry for primate core object recognition,
2. **convolutional neural networks** - artificial models for object recognition, and
3. **neural predictivity** - similarity metric for these visual hierarchies.

The subsequent three chapters present the core research work in this thesis.
In chapter 3 we systematically investigate the impact of visual angle of the alignment on the late representations of the primate ventral visual stream and deep convolutional neural networks. Chapter 4 contains our work on probing the role of image-context on the neural, 4.2.1 and behavioral 4.2.2 predictivity of artificial models. In chapter 5, we introduce a data-driven neural alignment approach, using an extensive set of neural data to fine-tune a model that more closely mirrors neural representations and explore its benefits.
For each of the three chapters, in the first section we present a detailed description of the data and methods implemented, followed by a section that outlines the key results and a discussion on their limitations and broader implications.
The thesis ends with chapter 6 where we put it all together :) and summarize the crucial results and explore future directions.

2

Foundational Concepts

2.1 The ventral visual stream

The complex details of primate vision provide an insight into a refined mechanism that has been perfected over millions of years of evolutionary development. The primate visual system begins with the retina, which detects light and converts it into electrical signals. These signals travel through the optic nerve to the Lateral Geniculate Nucleus (LGN) into the thalamus for initial processing. From there, they move to the primary visual area (V1) in the occipital cortex, where specialized neurons discern light's orientation and direction [56,58,59].

Beyond V1, the system splits into the dorsal and ventral streams. The dorsal stream, known as the *Where* pathway, processes motion and spatial awareness, while the ventral stream, the *What* pathway, with its hierarchical and feed-forward organization drives core object recognition [26]. The latter pathway is often divided in the following areas: V1, V2, V4 and IT (inferior temporal cortex), further split into posterior, central and anterior IT cortices (pIT, cIT,

aIT)[27], represented in Figure 2.1.A. These regions progressively process visual signals, evolving from basic feature detection in earlier stages (like V1) to more intricate object representations in areas like IT. This progression is also evident in the increasing size of neuron receptive fields (RFs) as one moves from V1 to IT[34].

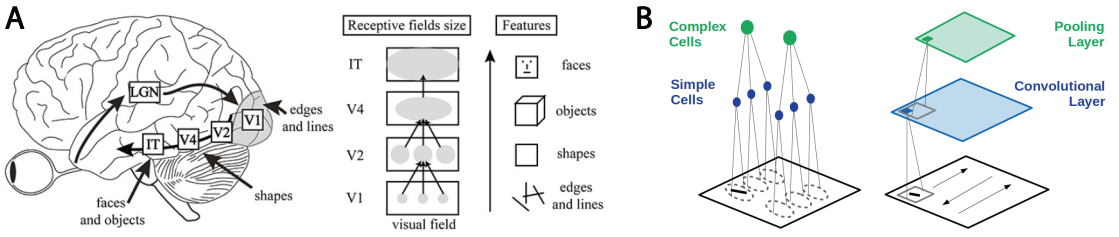

Figure 2.1: The primate ventral visual stream and CNNs A Left: The primate ventral visual stream processes information from the retina through the LGN, V1, V2, V4, to IT. Center: Receptive fields expand from lower to higher visual areas, aggregating more of the visual field. Right: V1 codes basic features like edges, while higher-level neurons integrate these, representing more intricate features. (Taken from Herzog et al.[50]) **B** The relationship between components of the visual system (left) and the base operations of a convolutional neural network (right). Simple cells (left, blue) respond to specific image orientations within preferred locations (dashed ovals). Complex cells (green) integrate inputs from multiple simple cells, achieving spatial invariance. In a CNN (right), the first convolutional layer (blue) is formed by convolving the image with filters (gray box), generating feature maps. Max-pooling, which takes the highest activation within a feature map section (gray box), downsamples the image and mimics complex cell responses (green), forming a pooling layer. (Taken from Lindsay et al.[86])

While V1 plays a role in basic feature extraction like detecting orientations and edges[55], V2 dives deeper, processing complex attributes like contours and textures[99]. V4, though still involved in orientation processing[98], is predominantly linked with color processing[111]and plays an essential part in maintaining color constancy, allowing consistent color perception regardless of illumination variations[94]. The final recognition of core object attributes predominantly happens in the IT cortex. While earlier stages in the visual pathway recognize simple features like edges and basic shapes, the IT cortex can detect and represent more intricate patterns, such as complex object parts and even entire objects[123,26]. Here, neuron groups are often tailored to be selective towards specific categories, including fruits, faces, bodies, and locations[72,26]. One of the remarkable properties of IT neurons is their invariance. A neuron that responds to a specific object will typically continue to do so regardless of changes in the object's size, position, or rotation in the visual field[127]. This invariance allows primates to recognize objects under various conditions[60,83].

2.2 Convolutional Neural Networks in computer vision

Convolutional Neural Networks (CNNs) have become foundational in computer vision. Originating from ideas inspired by the visual cortex's hierarchical organization in mammals, CNNs have achieved groundbreaking results in various visual recognition tasks. Structurally, a CNN clearly reflects the core work of Hubel and Wiesel [55] (Figure 2.1.B). It consists of multiple layers specifically designed to automatically and adaptively learn spatial hierarchies of features from input images. These layers include convolutional layers that apply a series of filters (feature detectors) to input data, producing "feature maps". This is followed by pooling layers that reduce spatial dimensions. After repeating these steps multiple times, non-convolutional, fully connected layers are integrated eventually categorizing the image into respective classes (Figure 2.2.C).

In 1989, the capabilities of CNNs were first highlighted when a relatively simple CNN model, using supervised learning via backpropagation, managed to classify handwritten numbers effectively [82]. But the real turning point for CNNs was in 2012, when an 8-layered model named AlexNet [77] set new performance standards in the ImageNet competition. This dataset [24], with over a million of diverse images, challenges models to categorize each image into one of a thousand distinct classes. AlexNet's achievement showed that the essential components of the visual system had the potential for broad vision applications when combined with the right training strategies and ample data. Following that, numerous CNN designs emerged, experimenting with depth, pooling layer positions, feature map counts, training methods, and the use of residual connections [106]. The main aim of this body of work was to improve image classification benchmark performance, with a growing emphasis on efficiency and training data reduction. Aligning with biological systems was no longer a driving factor.

2.2.1 CNNs as models of the primate ventral visual stream

CNNs as described above, by design, reflect the architecture of the mammalian ventral visual stream. Similar to the processing in the retina, CNNs normalize and segment images into RGB channels. Their layers, from convolution to pooling, mirror the progression from visual areas V1 to IT, as represented by Yamins et al. [130] in Figure 2.2. Each convolutional layer in a CNN can be thought of as a feature detector, with earlier layers often capturing simple

patterns like edges and textures, while deeper layers capture complex structures and objects. Their hierarchical feature extraction mimics the way the ventral visual system processes visual information, with simple cells in the visual cortex detecting local features and complex cells capturing more intricate patterns (Figure 2.1.B). Though these parallels were intentionally engineered, CNNs also exhibit unexpected non-engineered similarities with the visual system, especially in their alignment with neural data[130], underscoring their potential as representative models at a neural and behavioral level.

2.3 Neural predictivity

A significant reason for the renewed attention to artificial neural networks among neuroscientists stems from discoveries that they can capture the visual information representation in the ventral visual stream. Specifically, when both CNNs and animals view an identical image, the behavior of artificial units aligns with and can predict the activity of actual neurons—achieving an accuracy surpassing prior techniques. In 2014, Yamins et al.[130] first demonstrated this connection by recording neural activity in macaques as they looked at images of objects. By comparing the activity of actual V4 or IT neurons to the behavior of artificial units in hierarchical CNNs and verifying the predictive capability on a separate test set, they showed that networks excelling in object recognition were also better at predicting neural activity. This observation was consistent even with video classification[122]. Notably, the neural activity in IT was most accurately predicted by the network's final layer, while V4 activity aligned with the network's penultimate layer (Figure 2.2). This correlation, where later network layers more accurately represent the upper parts of the ventral stream, has been confirmed in other research including human fMRI[42], MEG[114], and dynamic videos over static images[30].

2.3.1 Metrics

Regression

In a classic neural predictivity analysis, a model layer's responses are linearly mapped to macaque IT neural responses involving techniques such as Partial Least Squares[46] or Ridge regression[51].

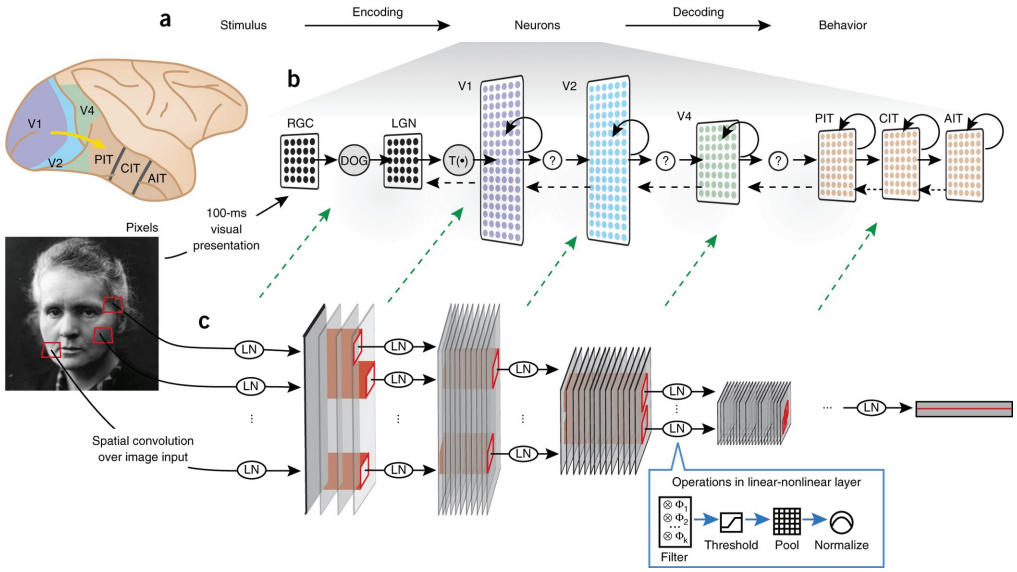

Figure 2.2: Hierarchical CNNs as models of the sensory system. **a** Sensory cortex studies focus on encoding (how stimuli turn into neural activity) and decoding (how neural activity drives behavior). Hierarchical CNNs (HCNNs) model the encoding phase, depicting how stimuli relate to observed brain responses. **b** The ventral visual pathway, represented most extensively, as a chain of interconnected cortical regions in the macaque brain. This includes areas like PIT (posterior inferior temporal cortex), CIT (central), AIT (anterior), RGC (retinal ganglion cell), and LGN (lateral geniculate nucleus). Symbols like DoG denote the difference of Gaussians model, and T(•) represents a transformation. **c** A representation of an HCNN. Each layer combines operations like filtering and pooling in a linear-nonlinear (LN) manner. Operations in a layer focus on specific input sections, simulating small receptive fields (red boxes). Layer stacking results in a complex transformation of input. As layers progress, retinopy decreases while receptive field size grows. (Taken from Yamins et al.[129])

Some recent non-linear approaches have also been developed[3], however the linear method remains the benchmark in the field, especially given its prominence in widely recognized metrics like Brainscore[112,113]. This goal-driven approach uses models pre-trained on a specific task, such as object classification, and regresses their resulting intermediate feature representations to model the neural responses. The mapping's performance is defined to be the noise-corrected Pearson's correlation between the model predictions and the observed neural responses. This is done in a cross-validated way by first obtaining the goodness of fit R^2 for each neural site, which is then corrected, by dividing it with the square root of the Spearman-Brown corrected self-consistency of that neural site over the image presentation repetitions. For noisy models we also correct by the internal consistency of the model.

For a neural site i, the normalized explained variance is given by $EV_i = \frac{R^2}{\sqrt{\rho_m \times \rho_n}}$, where ρ_n and ρ_m are the neural and model corrected split-half correlation, and R is the correlation between the predicted and actual neural response. This average explained variance computed across all neural sites is the metric for neural predictivity that will henceforth be used in this thesis.

3

Visual angle

For any analysis predicting animal physiology or behavior, it is *essential* that both the animal and the model - typically an Artificial Neural Network (ANN), view the *same stimuli*. A significant aspect here is the *visual angle* θ — the angle under which the animal sees the image. This angle is influenced by the screen size and its distance from the viewer, their relation is described in equation 3.1.

$$\theta = 2 \times \arctan\left(\frac{\text{size of the object}}{2 \times \text{distance from the object}}\right) \tag{3.1}$$

An extensive review of studies in vision over the past decade predicting animal empirical data reveals considerable variation in the visual angles used. This variability is not only inter-species but also intra-species, as illustrated by Figure3.1. Across various animals including mice, rats, marmoset and macaques, this angle ranges between 1 and 120 degrees, while in human studies it varies from 2.9 to 20 degrees.

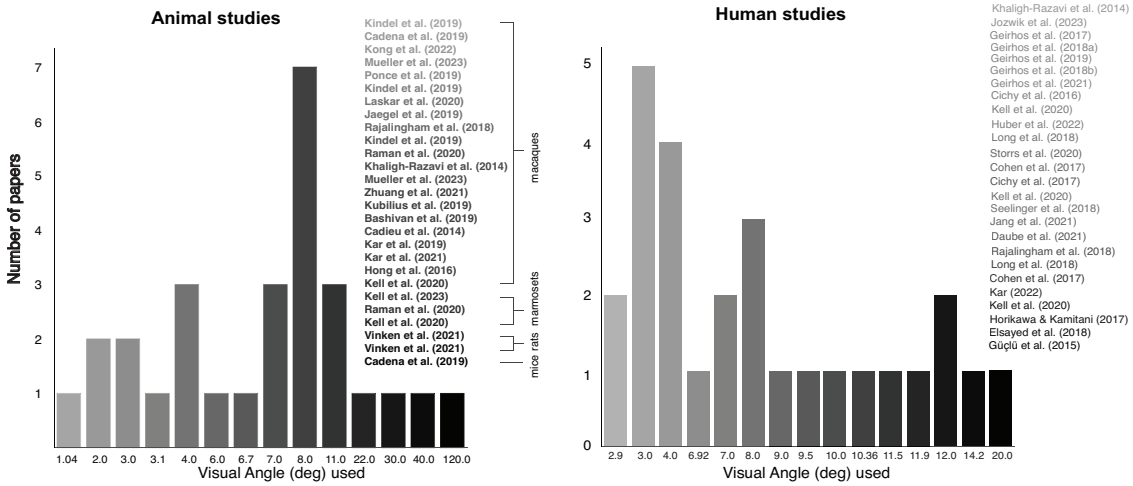

Figure 3.1: Survey of studies evaluating the alignment between ANNs and empirical data. A review of studies done over the past decade predicting animal neural or behavioural responses to a visual stimuli. For each study we show the visual angle reported during stimulus presentation. The left panel shows non-human animal studies and the right panel shows studies that used human subjects. The field of view (FOV) variability is present across species as well as within. We show results for mice [13], rats [126], marmoset [69,68] macaques [67,66,79,105,7,12,61,71,73,74,81,95,101,134] and human studies [18,16,17,19,23,31,128,38,35,37,36,42,52,57,103,62,65,68,71,88,114,116].

In contrast, the ANNs being compared with the animal data invariably receive the full image, with the strong assumption that the neural population's FOV spans the entire image. This raises the question of whether there is a match between what the models and the neurons are "seeing". When the visual angle is too large, the neurons recorded may primarily respond to the center of the image rather than processing the larger periphery. Conversely, if the stimulus is too far or relatively small, neurons may respond to something happening in the background that is not part of the image. Such disparities challenge the fairness of comparisons, as ANNs are not exposed to the same stimuli as the neural population they aim to explain. If these models are built as a hypothesis of a visual system, they should commit to a visual input size which should remain unchanged during all animal experiments.

To investigate this effect, we conducted an experiment in which we presented a large stimulus of 20 degrees of visual angle to a macaque chronically implanted with two Utah electrode arrays (Figure 3.2.A). We recorded its neural responses while the animal was performing a passive viewing task (details in 3.1) and fed the same images to deep CNNs for a typical neural predictivity analysis. In order to model the visual angle change, we performed center crops of

smaller sizes on the same images and repeated the analysis to assess the effects on the models' ability to explain the neural data.

3.1 Data and methods

3.1.1 Visual Stimuli

We used 75 gray-scale natural images of 5 object categories - car, airplane, bird, bear, and elephant with equal distribution. The images were taken from Microsoft Common Objects in Context (COCO)[85] and modified, such that their context was removed and swapped with noise, to avoid any context modulated effects. To simulate the visual angle change from the model side, the images were center cropped in a range from 2 to 20 degrees (the full image size). After cropping they were resized to the original 512x512 pixels dimension. As a control, we also kept the original, full context 75 images and did the same transformations on them.

3.1.2 Macaque neural data collection and processing

The neural activity was recorded using two micro-electrode arrays (Utah arrays) implanted in the IT cortex. A total of 96 electrodes were connected per array (grid arrangement, 400 um spacing, 4mm x 4mm span of each array). The array placements allowed us to sample neural sites from different parts of IT, along the posterior to anterior axis. However, for all the analyses, we did not consider the specific spatial location of the sites, and treated each site as a random sample from a pooled IT population. During the passive viewing task, the animal fixated on a white dot (0.2° of visual angle) for 300 ms to initiate a trial. We then presented a sequence of 5 to 10 images, each ON for 100 ms followed by a 100 ms gray (background) blank screen. This was followed by fluid (water) reward and an inter trial interval of 500 ms, followed by the next sequence. During each daily recording session, band-pass filtered (0.1 Hz to 10 kHz) neural activity was recorded continuously at a sampling rate of 20 kHz using Intan Recording Controllers (Intan Technologies, LLC). The majority of the data presented here were based on multiunit activity. All surgical and animal procedures were performed in accordance with National Institutes of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care.

The neural data analyzed for this study was collected when the macaque was performing the passive fixation task, on the dataset of 75 images, presented for 59 repetitions, at a large visual angle of 20°. Each neural site's response to an image was taken as the mean rate during a time window of 100-170ms following image onset, a window that has been chosen based on the neuronal population's self-consistency (Figure 3.2.B) and aligns with the visually-driven latency of IT neurons[92]. We selected the most reliable (self-consistency over 0.6) neural sites from both pooled arrays, resulting with an IT population of 88 neural sites.

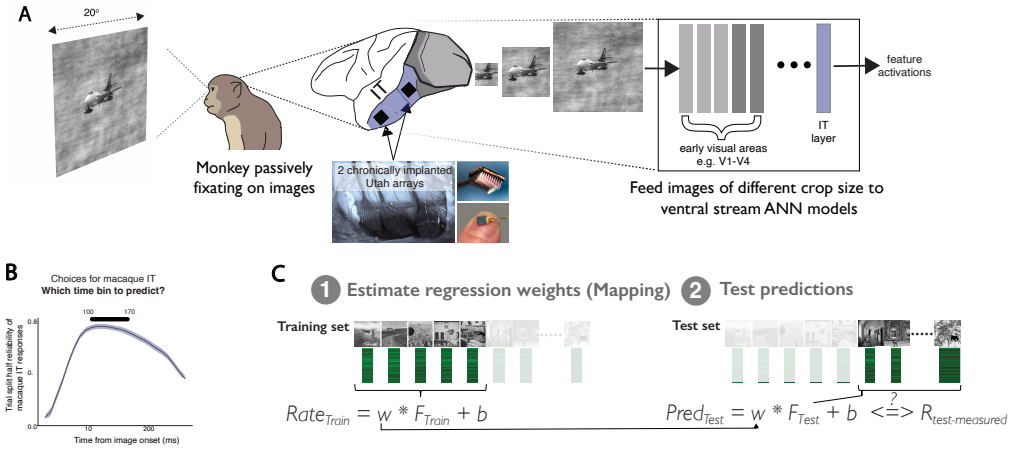

Figure 3.2: Conceptual schematic of the experiment. A We recorded activity from one adult macaque in cortical area IT while monkeys passively viewed large 20 degree images. We fed the same images to a series of DCNNs and extracted their features from the most IT-like layer. We did central crops for each image ranging from 2 degrees to 19 degrees of visual angle and repeated the feature extraction for each crop size. **B** Choosing the most reliable time-bin from the neural data. The self-consistency, measured by the trial split-half correlation for all images, is shown on the y axis. The x axis depicts the time elapsed since the image onset. The optimal time-frame chosen for the neural prediction was 100 to 170 ms. **C** Neural predictivity with a Ridge regression in a cross validated way. The regression weights predicting the neural responses were estimated from the split train-set of images and evaluated on the test set. This was repeated for 10 random iterations.

3.1.3 Models

We evaluated nine different state-of-the-art DCNN models, on the exact images shown to the macaque. Each of these models had been trained using the extensive ImageNet 1000-way object categorization task. We focused on publicly available PyTorch DCNN model architectures that have demonstrated significant success in this computer vision benchmark:

AlexNet[77],GoogleNet[118], VGG-19[115], MobileNet[53], Resnet-18, ResNet-50, ResNet-101[45], DenseNet-201[54] and Inception-v3[120].

We extracted the most IT-like features from each model for every stimulus, from the layers chosen based on BrainScore[112,113], the field's predominant benchmark. To ensure consistency in results across the models, given the varying layer sizes for each, we standardized the dimension for every model down to 1,000 features. This was done by using Gaussian random projection with 1000 components to project the full extracted features space on a randomly generated linear subspace in such a way that distances between the points are nearly preserved. These features were then regressed onto the neural responses in a cross validated way, using a 20 fold cross validation (hyperparameter tuned in the range 2-25, with similar results for lower number of folds) for ten different variations of the splits. To avoid overfitting, we used a Ridge regression[51] with a strong regularization parameter (α=100, tuned with a cross validated logarithmic scale search) to predict the response of each neural site and reported the mean explained variance for all predictions.

3.2 Results

We assessed the predictivity of a series of DCNN model across various image crops and identified the visual angle corresponding to the maximum explained variance. An illustration for one model, Inception-V3 is depicted in the inset of Figure 3.3.A. This model is most effective at explaining neural responses when limited to a 9 to 10° of visual angle (mean preferred angle = 9.6$\pm$1.3°). It is important to note that the model features for every crop are explaining the same neural data, we are only changing their input image. This finding generalizes to all tested DCNNs, depicted in the supplementary Figure A.2. All models seem to be most effective at explaining neural responses when limited to a small window of 7-11 degrees of visual angle (Figure 3.3.A). This suggests that presenting the full stimulus to the models might provide an inequitable representation of their ability to interpret neural data responding to a narrower visual angle.

Our results showed that the central reduced visual angle is the best predicted location of the image for the neural population. As a control, we analyzed this for individual neural sites for different degree crops covering the full image. For each image, we extracted smaller patches

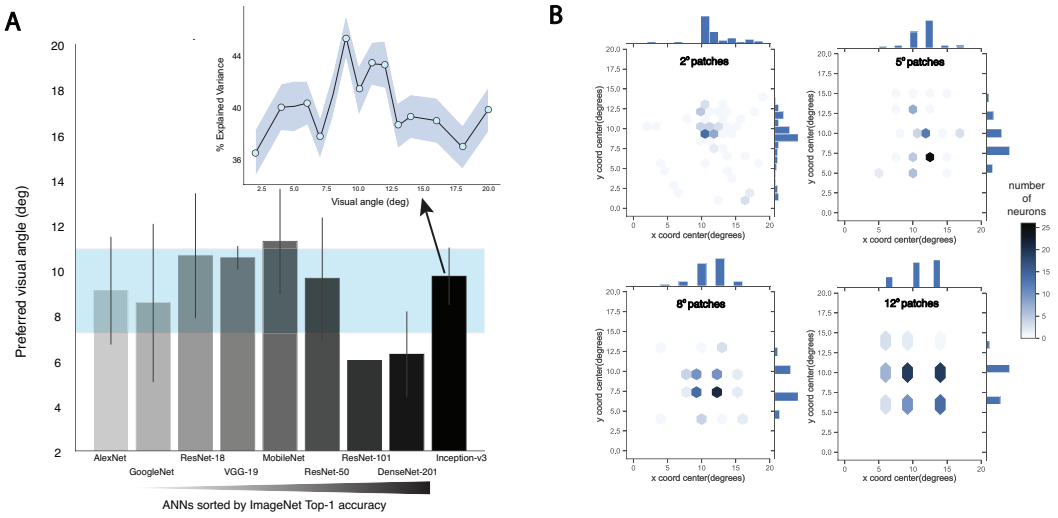

Figure 3.3: The most macaque IT aligned ANN image viewing angle is 7-11 degrees.A Shows the preferred visual angle for each model, chosen based on the model's highest explained variance for the same neural response. The standard deviation is shown across different sub-sampling of images used in the linear mapping to estimate the parameters. The inset illustrates the explained variance changes for Inception-v3 across different degrees of visual angle, with the standard error for one sub-sampling seed. The average optimal visual angle for this model is 9.6±1.3 degrees, EV=45.37±1.7%. The light blue band shows the mean with the standard deviation for the preferred visual angle across models. **B** We show the crop center location relative to the full image size, which led to the highest predictivity for a neural site. The crops were done by starting at the edge of the image and sliding for 1 degree in both directions. Every panel shows the count of neural sites which were best predicted when the model was fed the patch in the depicted location. The model (Inception-v3) is best at predicting most neurons when viewing only the central image patches.

of 2,5,8, and 12 degrees at every location and ran our analysis for each patch. We found that most neural sites were best predicted for central crops. Figure 3.3.B shows the preferred center location (x and y coordinates within the image) of the crop that best explains each neural site. Although some periphery-preferring neurons were identified, the majority were unreliable and the difference in EV from the central and periphery crops were not significant (see supplementary Figure A.1). The images presented were modified by eliminating their original context and substituting it with random noise, to avoid any contextual effects on this analysis. Nevertheless, we also verified the influence of the FOV alteration on the unaltered, full context images (see supplementary Figure A.2). Consistently, our findings indicated that the models most accurately explained the neural population responding to these images when limited to a narrower central FOV, with an average angle of 12.01±2.13°.

3.3 Discussion

Overall, our research revealed the vital role of visual angle in neural predictivity. The performance of all of our evaluated models was affected by the change in visual angle when predicting the same neural responses. Notably, most models were able to predict these responses with a higher accuracy when only shown a smaller center crop of the image, while the macaque was fixating on the full 20° image. We speculate that this is due to the IT neural population responding largely to the central ≈10 degrees of the image, so only showing these crops would allow the model to better fit the data. Taken together our results urge the need for careful consideration of the visual angle when presenting stimuli, to ensure a true comparison between neural responses and artificial units.

There are several limitations of this work and future directions to consider. Our methodology mainly adjusted the model's FOV. However, a similar approach could be applied to empirical data: recording neural activity while the subject views images from varying visual angles (2-20°). Here, the models would be presented the entire image, allowing us to validate if their extracted features still optimally predict the neural responses to images viewed at smaller visual angles. It is also worth investigating these effects in different brain regions, such as the primary visual cortex and V4, allowing for an analysis of overlapping receptive fields. If the preferred FOV remains within a similar range for all regions, we can speculate a need for a fixed universal visual angle size for primate neural predictivity studies. Moreover, it would be insightful to evaluate these effects on a behavioural level, with the aim of finding the optimal visual angle size for which the model's binary discrimination accuracies align most closely with primate behaviour. Finally, the relatively small and simple set of stimuli used here might fail to adequately capture the models' performance. Repeating these experiments with a larger image set and data from multiple subjects would be necessary to validate our results. Nevertheless, our goal was not to find the "optimal" visual angle, rather to point out the significant effects that visual angle has on the alignment of artificial units with empirical data, and underscore the importance of careful experimental design that takes these effects into consideration. This is crucial for consistent cross-study comparisons and further advancing our understanding of visual perception and neural processing.

4

Context

Context matters. In real-world situations objects do not appear alone, rather surrounded with other objects and scene properties. In the visual domain, our perception of the environment is shaped not just by what lies directly in our focal point, but also by the surrounding scene elements and our past experiences [132,5]. Beyond psychophysical demonstrations of how context can impact vision [6,41], we know very little about the mechanisms that help integrate objects and surrounding information during scene understanding. "Low-level" contextual effects have been studied extensively, including extra-classical receptive fields [49], temporal adaptation [49], and surround suppression [1]. However, major lacunae remain in our understanding of how context impacts "higher-level" visual recognition. Understanding context and relationships between objects is essential for computer vision as well. Deep neural networks used for object recognition, especially those trained on natural image datasets like ImageNet [78], heavily and implicitly depend on context [37]. In fact, these algorithms tend to struggle when objects are situated in incongruent (wrong) contexts [8,29].

In this chapter, we evaluate the effects of scene context on neural predictivity and in visual object recognition. Our findings reveal a consistent contextual modulation in macaque physiology and the behavioral accuracies of both humans, monkeys and deep CNNs.

4.1 Data and methods

4.1.1 Visual Stimuli

We used a dataset of 600 gray-scale images from 10 object categories, including bear, elephant, person, car, dog, apple, chair, plane, bird and zebra. For each object category, we selected 6 natural images from the Microsoft Common Objects in Context (COCO)[85] dataset, varying in object size and location, which were center square cropped (if needed), converted to gray-scale, and finally re-scaled to 512x512 pixels. We then generated 10 different contextual variations for each image, including congruent, incongruent, no context, masked object, blurred context, blurred object, blurred incongruent boundary, minimal, textured, and jigsaw context. We extracted the object from each image using the COCO object annotation masks and did the defined contextual manipulations. We used a Gaussian blurring kernel of size 2 to blur the object, background, and object-context boundary. The jigsaw context was generated by cropping the context into 25x25 pixel chunks and randomly shuffling them around the object. The textured context was created using the Portilla and Simoncelli method[102] applied to the original grayscale image with five iterations. An example of the contextual manipulations can be seen in Figure 4.2.A.

Low level features

For every image, we extracted a range of basic features, such as object size, location and category, spectral mean and std, and contrast mean and std. The standard contrast metric for gray-scale images was obtained using the highest and lowest pixel values $\frac{P_{max}-P_{min}}{P_{max}+P_{min}}$. The contrast standard deviation was derived from the pixel-wise standard deviation of the grayscale image. From the COCO[85] object annotations, we determined the object size, represented in degrees of visual angle. The x and y coordinates, relative to the image, captured the object's central position. Using the Fast Fourier Transform[10] (FFT) we transformed the image in the

spectral domain, and noted its spectral mean and standard deviation. The peak power was set as the maximum value of the spatial frequency (magnitude of the FFT) for each image.

4.1.2 Macaque neural data

The neural activity was recorded from one adult macaque monkey, using two micro-electrode arrays (Utah arrays) implanted in IT cortex. All neural data analyzed for this study was collected when the macaque was performing the passive fixation task, on the contextually manipulated dataset of 600 images presented 31 times. For more details of the neural data collection and processing see 3.1.2. Using the mean time window of 100-170ms post image onset (selected as the most reliable time-bin, see Figure 4.1), we selected the most consistent responses (self-consistency>0.6) from both pooled arrays resulting with an IT population of 46 most reliable sites.

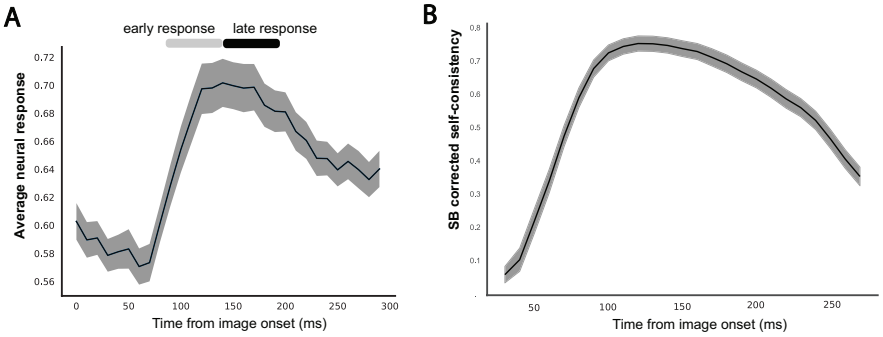

Figure 4.1: Averaged neural response and neural data consistency. **A** Overall averaged neural population response. Mean across all 600 images, 30 trials per image, and 192 neural sites. Shown for increasing averaged 10ms response time bins (with a 10ms time-step) from image onset. The lighter band indicates the standard error. **B** Spearman-Brown corrected split-half reliability for the neural sites used, as a function of time from image onset. Each time-bin is an average of 30 ms, and we are showing a 10ms time-step. The lighter band indicates the standard error.

4.1.3 Active binary object discrimination task

Macaque active binary object discrimination task

We measured monkey behavior from 2 male rhesus macaques. Images were presented on a 24-inch LCD monitor (1920 × 1080 at 60 Hz) positioned 42.5 cm in front of the animal.

Monkeys were head fixed. Monkeys fixated a white dot (0.2°) for 300 ms to initiate a trial. The trial started with the presentation of a sample image (from a set of 600 images) for 100 ms. This was followed by a blank gray screen for 100 ms, after which the choice screen was shown containing a standard image of the target object (the correct choice) and a standard image of the distractor object. The monkey was allowed to view freely the choice objects for up to 1500 ms and indicated its final choice by holding fixation over the selected object for 400 ms. Trials were aborted if gaze was not held within ±2° of the central fixation dot during any point until the choice screen was shown. Prior to testing in the laboratory, monkeys were trained in their home-cages to perform the delayed match to sample tasks on the same object categories (but with a different set of images). We obtained a minimum of 31 trials per image from the pooled monkey responses.

Eye Tracking

Macaque behavioral testing was performed using standard operant conditioning (fluid reward), head stabilization, and real-time video eye tracking. We monitored eye movements using video eye tracking (SR Research EyeLink 1000). Our 2 macaque subjects were trained to fixate a central white square (0.2°) within a square fixation window that ranged from ±2°. At the start of each behavioral session, monkeys performed an eye-tracking calibration task by making a saccade to a range of spatial targets and maintaining fixation for 500 ms. Calibration was repeated if drift was noticed over the course of the session. Real-time eye-tracking was employed to ensure that eye jitter did not exceed ±2°, otherwise the trial was aborted, and data discarded. Stimulus display and reward control were managed using the MWorks Software.

Human active binary object discrimination task

We measured human behavior using the online Amazon MTurk platform which enables efficient collection of large-scale psychophysical data from crowd-sourced "human intelligence tasks" (HITs). The reliability of the online MTurk platform has been validated by comparing results obtained from online and in-lab psychophysical experiments[92,104]. Each trial started with a 100 ms presentation of the sample image. This was followed by a blank gray screen for 100 ms; followed by a choice screen with the target and distractor objects, similar to Rajal-

ingham et al., 2018[103]. The subjects (n=90) indicated their choice by touching the screen or clicking the mouse over the target object. Each subject saw an image only once. We collected the data such that, there were a minimum of 25 valid pooled subject responses per image, with varied distractor objects.

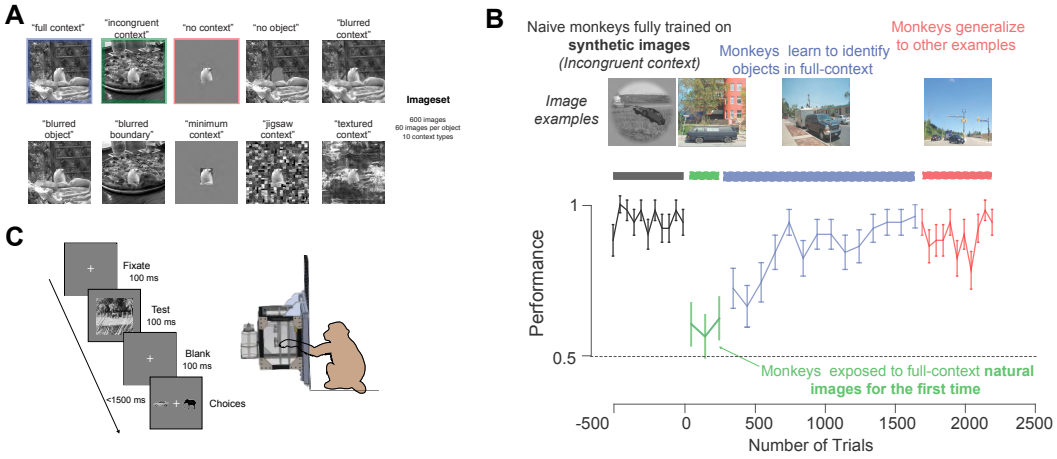

Figure 4.2: Conceptual schematic of the binary discrimination task. A An example of the ten contextual manipulations done for one image of the image-set used for both the neural and behavioral experiment. **B** Training process for the each macaque. The monkey is initially trained with incongruent-context images (black curve). This training does not generalize and results in low starting performance in full-context (green curve) . However, monkeys quickly learn to recognize images in full context (blue curve). Furthermore, this ability generalizes to new images (red curve). **C** Binary object discrimination task, showing the timeline of events for each trial. Subjects fixate on a cross, then the test image at 8 degrees containing one of ten possible objects and contextual manipulations is shown for 100 ms. After a 100-ms delay, a canonical view of the target object (the same as that presented in the test image) and a distractor object (one of the other nine objects) appears, and the human or monkey indicates which object was present in the test image by clicking on or making a saccade, respectively, to one of the two choices.

CNNs active binary object discrimination task

We evaluated a series of DCNN models, trained on ImageNet (described in section 3.1.3), using the same images and tasks that were shown to humans and monkeys. To make these ImageNet-trained models compatible with our specific 10-way object recognition task, we extracted their most IT similar feature representations (based on BrainScore[112]), and trained a multiclass logistic regression classifier using these features to calculate the cross validated prob-

abilities for each object class, mimicking the binary object discrimination task. We trained the regression on a different dataset of 800 unaltered (full-context), natural COCO images and then assessed its performance on our set of 600 contextually-modified images.

4.1.4 Behavioral metrics and signatures

To characterize the behavior of the visual systems, we used two behavioral metrics, the hit rate resolution at context-level - C1, and more fine grained image-level - $I1_n$ (refer to [103] for more details). Each behavioral metric computes a pattern of unbiased behavioral performance, using the model's accuracies per image averaged across all trials. We obtained a biological or artificial "signature" for each system by applying each metric to its behavioral accuracies. The one-versus-all context-level performance metric (termed C1) estimates the discriminability of each context category c, essentially pooling the accuracies across all images of context type c and all object/distractor pairs within. Because we here tested 10 context categories, the resulting C1 signature has 10 independent values. Figure 4.3.A shows the C1 behavioral signatures for primates and all tested models.

The one-versus-all image-level performance metric (termed I1) estimates the discriminability of each image containing object o from all other objects, pooling across all possible distractor choices. Because we focused on the primary image test set of 600 images (10 per object, see above), the resulting I1 signature has 600 independent values, see Figure 4.3.B. Given an image i of object o, and all nine distractor objects (d≠o) we computed the average performance per image as:

$$I1_i^o = \frac{\sum_{d=1}^{10} Pc_i^{o,d\neq o}}{9}$$

where Pc - percent correct, is the fraction of correct responses for the binary task between objects o and d. Considering every image i_c of context type c, the C1 performance for each context type is the mean across all images' (60 per context type) performance:

$$C1_c = \frac{\sum_{i_c=1,}^{60} I1_{i_c}}{60}$$

Both of these behavioral signatures are however tightly linked. For instance, images with context that is challenging to discriminate generally would display lower performance metrics as

opposed to images with "easier" context manipulations. To pinpoint the behavioral variability strictly influenced by variations in images, and not determined by the context embedded (as already represented in C1), we introduced normalized behavioral metrics at the image level - $I1_n$. For an image i of context type c and object o, this metric is then given by subtracting this context mean from the image-level performance:

$$I1n^o_{i_c} = I1^o_{i_c} - C1_c$$

4.1.5 Behavioral consistency

To quantify the similarity between a model visual system and the primate visual system with respect to a given behavioral metric, we used the "primate consistency" measure, similar to the "human consistency" previously defined[64,103]. Primate consistency is computed, for the two behavioral metrics, as a noise-adjusted correlation of behavioral signatures[25]. To obtain the primate behavioral ceiling, we randomly split all behavioral trials into two equal halves and applied each behavioral metric to each half, resulting in two independent estimates of the system's behavioral signature with respect to that metric. The self-consistency of the system is then obtained by the Pearson correlation between these two estimates of the behavioral signature, Spearman-Brown corrected. This is a measure of the reliability of that behavioral signature given the amount of data collected, and is noted as the primate self-consistency. The primate consistency for every model is then found by correlating the primate and model behavioral signatures. For the neural model this is noise-adjusted by the neural self-reliability. The purpose of using the primate ceiling is to consider the unpredictable variances in behavioral patterns due to differing factors - "noise" not reproducible by the experimental condition, which a model cannot anticipate.

4.2 Results and discussion

We tested humans and monkeys on contextual information for real-world objects, such as cars, animals, and fruits. We introduce multiple variations of the contextual information to further our understanding of what aspects of the object's surround impact recognition including re-

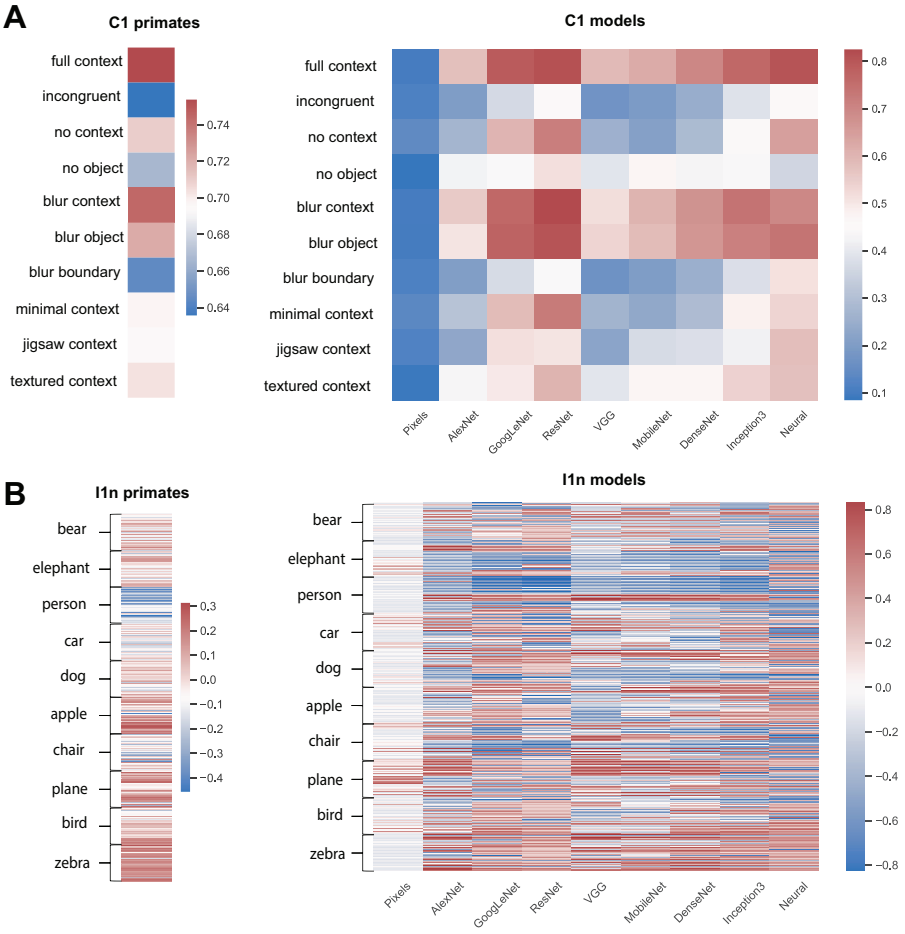

Figure 4.3: Behavioral signatures. A Showing the 10 dimensional C1 behavioral signature for primates (left) and models (right). Each value represents the accuracy grouped across all images of that context type, for all distractor objects. **B** Similar as A but for the $I1_n$ behavioral signature. Each line in the signature is the accuracy for an image, averaged across distractors. Images are shown grouped by object category.

moving the context, swapping it with incongruent context, and blurring different parts of the image.

4.2.1 Physiology

To rigorously quantify the effects of contextual changes on the neural predictivity, we recorded the IT neural responses of one context-naive monkey while passively fixating on our set of 600 contextually modified images.

Effects of scene context on neural data

We first examined the impact of context on the neural responses. Using the IT population, we decoded the context category through cross-validated one-vs-all classification. The confusion matrix derived from the decoding accuracies is presented in Figure 4.4.A. The neural data demonstrated the capability to decode the image context category at an accuracy surpassing the chance level (≈26%), with notable variability across different context types. Neural responses for contexts such as no context, minimal context, absence of object, textured, and jigsaw contexts were decoded with greater certainty. Several patterns of contextual alterations appeared to be frequently misclassified, implying their neural responses were closely related. This was evident for no context versus minimal context, the full context when compared to its blurred variants (either on the object or the context itself), and the incongruent context compared to its "blurred boundary" version (where the sharp transition between the context and object was softened).

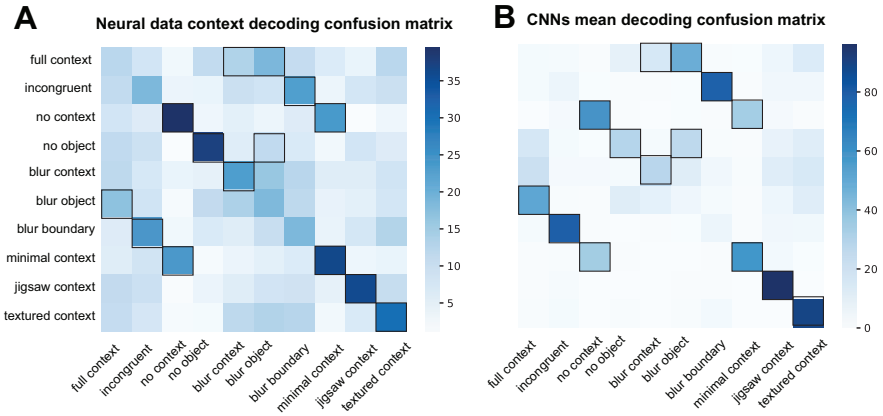

Figure 4.4: Confusion matrix when decoding context for the neural population and CNN features. A Confusion matrix for the decoding accuracy of the neural data model, when decoding context. The diagonal shows the percent of correctly predicted images of the appropriate context type. The i-th row and j-th column entry indicates the percent of images with true context label being i-th context category and predicted label being j-th category. **B** Similar as A but for the model features,showing the mean across all models' confusion matrices. The outlined results indicate the same pattern across the two plots.

Effects of scene context on CNN activations

Similarly, we decoded the context type from the CNN activations, with a higher overall accuracy ($\approx 45\%$), confusion matrix shown in Figure 4.4.B. A congruent pattern with the neural data is evident, where again the most distinct responses typically occur when presented with images with no context, minimal, jigsaw, and textured context; these contexts are predicted with notable accuracy. The parallel confusion trends with the neural data are highlighted in Figure 4.4.B. We subsequently analyzed the model activations' similarities across different context types for identical images, using the cosine similarity metric[131]. This was done for features representations across context types of the same image and then averaged across all images and models, see supplementary Figure A.6. Again, we see an analogous pattern of similarity for the model features as those most confused from the neural data.
These results indicate that context has an impact on the IT population responses as well as on the deep CNN activations, and it suggests that they might encode for it in a similar manner, in particular for the most accurately predicted categories for both populations.

Contextual effects on the neural predictivity

A natural next step was to look at the alignment between these neural and artificial features. Our findings suggest that the IT responses to isolated objects—those presented without any surrounding context—are most effectively predicted by deep CNN models' features. Figure 4.5.A illustrates the predictive capabilities of Inception-v3 across three markedly different context scenarios: congruent, incongruent, and no context. There is a noticeable difference between the prediction accuracy for images without context and those with the two other contextual variations. It is essential to note that this difference is not a consequence of limited data availability; as seen in the figure, the pattern persists when the sample size of images in the linear mapping is increased. This indicates that the context-dependant gap would not be bridged by increasing our image set, as perhaps the model learns these representations "slower". We presented one model, however, this inclination towards better prediction for no-context images is consistent across all tested models, indicating a universal trend among them. Figure 4.5.B shows the average explained variance distribution across all CNNs for the three context types. We see a clear rightward shift when these models predict the responses to

images with the context removed compared to keeping the congruent and incongruent context. Furthermore, for each evaluated model the neural data responding to images with no context or minimal context were always better predicted than the same images with the other contextual manipulations that include the background (see supplementary Figure A.7). This trend persisted when testing using other neural similarity metrics including RSA, CKA and "neural correlation" (see A.1), shown in the supplementary Figure A.5.

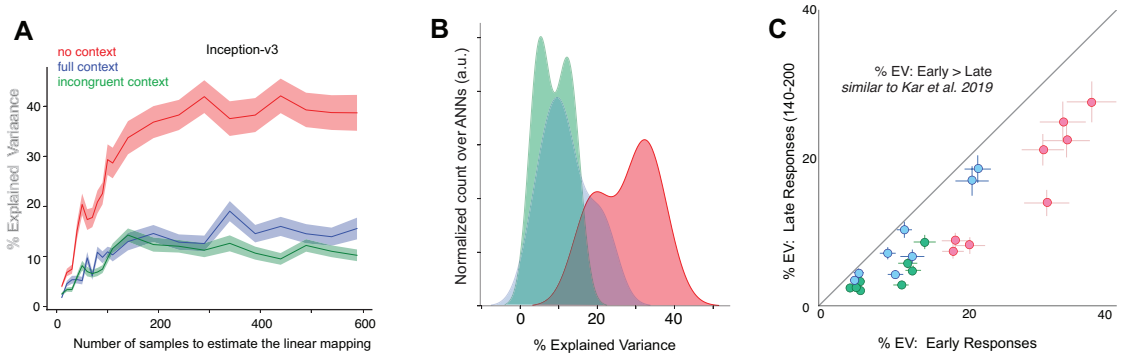

Figure 4.5: Comparison of neural predictivity for full, incongruent and no-context images. **A** Percentage of average explained variance and standard error for Inception-v3 as a function of the number of images used in the linear regression mapping. The colors indicate the different context types. **B** A KDE showing the distribution of the percentage of average explained variance for the models, categorized by context. **C** A scatterplot of the EV for 80-140 ms early and 140-200 ms averaged late neural responses. Each point is a model, the colors indicate the context types, the standard error is plotted on both axis respectively.

Next, we looked at the differences in explained variance for early and late neural response intervals. Based on the data peak response, the early interval was determined to be from 80 to 140 ms and the late from 140 to 200 ms, Figure 4.1.A. As highlighted in Figure 4.5.C, no context responses are consistently better predicted across both these time-frames. Our observations also confirm existing literature, as seen in Kar et al.[67], suggesting that these models can better explain earlier neural responses, compared to late. Kar et al.,[67] also highlighted that shallower recurrent CNNs outperformed standard feed-forward deep CNNs in predicting later neural responses. This might stem from the putative top-down mechanisms at play during later neural processing stages—mechanisms that these feed-forward models might struggle to capture. Over the past decade, studies on neural predictivity have used a range of image contextual manipulations. As shown in the supplementary Figure A.4, the context varies from no context to textured and incongruent context, across and within species (humans). Yet, our findings

underscore the significant influence of scene context (or its absence) on neural predictivity, suggesting that results from these studies may not be directly comparable. Our results consistently show that models struggle at aligning with neural data when presented images with (incongruent and congruent) context compared to the same images with the context removed. Additionally, these differences are there across contextual manipulations, notably, incongruent context is always predicted worse than full context (Figure A.7). We speculate that this is due to the need of more complex (feedback) mechanisms to capture the context (in particular incongruent context), which are not present in these one-directional models. A promising direction would be to assess the efficacy of recurrent neural networks in predicting these neural mechanisms.

4.2.2 Behavior

Moving from physiology, we looked at these contextual effects on behavior. We measured human and macaque behavior in a binary object discrimination task on the same contextually manipulated image set. To understand the neural processes behind the contextual influences, we require a more detailed examination of the neural networks involved. Rhesus macaques constitute an ideal animal model due to their similar visual processing circuits to humans[97,103]. Nevertheless, it is essential to first determine if macaques show comparable contextual effects. We observed reliable contextual modulation in the monkeys' behavioral accuracies, which were significantly correlated, and aligned with those observed in humans at the image-by-image level. Importantly, these changes could not be accounted for by low-level image features. To formulate hypotheses regarding the neural mechanisms driving these task performance patterns, we evaluated current deep neural network models of primate vision. Although many of these models sufficiently predicted the overall effects of context in primate behavioral performance (C1), they demonstrated a significant explanatory gap for image-level comparisons ($I1_n$) across specific contextual manipulations.

Behavioral effects of scene context on humans

Humans (90 participants on Amazon Mechanical Turk) participated in a binary object discrimination task (for details see, Kar et al., 2019[67]). Not surprisingly, our results show that

varying the context of the image significantly changed the performance of the human participants (Figure 4.6.A). The effect of contextual manipulations resulted in a consistent pattern of behavior (with a trial-split reliability of approximately 0.8, see figure inset). This was critical to ensure that such effects can be compared across other animals and ANN models. Incongruent context caused a significant drop in performance from full context, which supports previous research [6,41]. This decline was not solely due to the abrupt transition from the background to the object; even when this context/object boundary was blurred (termed as "blurred incongruent"), we observe the same effect . Predictably, the removing the object, retaining only its silhouette, also led to reduced accuracy. The blurring process itself seemed to have minimal influence on human responses, as the kernel size used was relatively small. Using a synthesized texture, which retained the visual attributes of the original context (generated using the Portilla and Simoncelli iterative technique, refer to 4.1.1), also adversely affected human behavior. Moreover, when the context was removed or minimized, there was again a decline in performance, indicating that humans also rely on the surrounding for object recognition. Our results aligned with extensive previous research on human behavior [132], which further validated our method and data collection.

Behavioral effects of scene context on macaques

Rhesus macaques have a visual processing circuit that is homologous to humans. However, it is critical to first ask whether macaques show similar contextual effects. To ensure that macaques are familiar with scene context, we first explicitly trained them with images in context (from the Microsoft COCO dataset). Macaques showed robust cross-validated accuracy during such training (Figure 4.2).

Contextual effects: human vs. macaques

Once the monkeys (n=2) were fully trained (i.e., reached ≥80% performance) in their homecages [110], we presented them with the same contextually manipulated images as humans. We recorded their responses and tested the internal consistency by calculating the split-half reliability across trial repetitions (for details, see Rajalingham et al., 2018 [103]) of the monkeys' behavioral accuracy as the context varied. We obtained a high correlation of ≈ 0.9, validating

the consistency of the monkey behavior. Figure 4.6.B shows that the contextual variations in

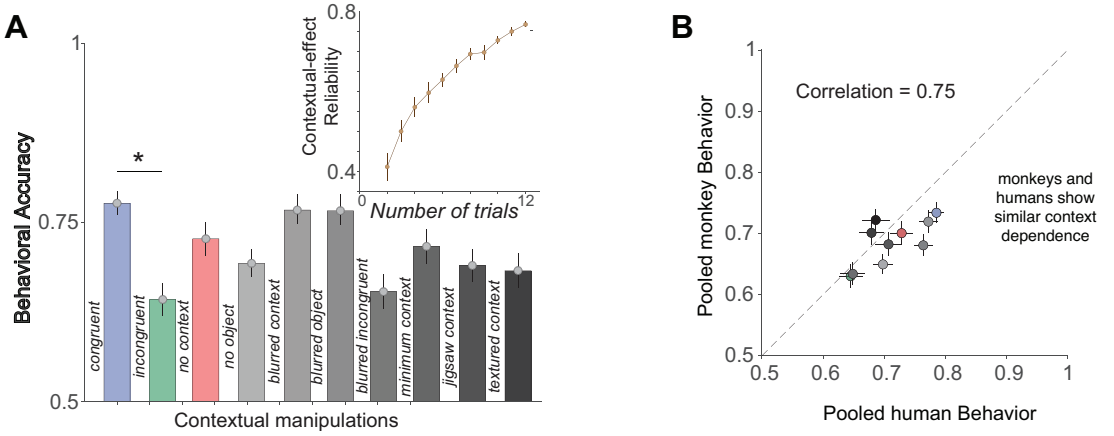

Figure 4.6: Effects of context on human and monkey behavior. **A** Contextual manipulations produce significant changes in human behavior. An example effect is shown between congruent and incongruent context. Inset shows that the pattern of contextual effects can be reliably estimated as total number of trials per image increases during human data collection (Pearson R ~0.8, for 24 repetitions per image) **B** Contextual effects in monkeys are correlated with those in humans (Pearson R = 0.75). The color indicates the context category from A

monkeys and humans were significantly correlated (noise-corrected Pearson R = 0.75). The pooled humans and macaques have very similar contextual recognition patterns. Humans, far more exposed to context, are more accurate for most context types, apart for jigsaw and textured context. We observed that these low-level features do not predict the context-level or image-level measured behavioral variance (see Figure 4.7). From these features, object size showed the most consistency at the image level, aligning with prior studies highlighting its significant influence on human behavior[132]. Its effect however is marginal, accounting for only 10.5% of the explained variance. As expected, the control Pixels model - using the raw image pixel values, did not reflect primate behavior at image nor context level. These observations establish monkeys as a good model of humans to further study the neural mechanisms of context during visual object recognition.

Comparison with deep neural network models

Next, we tested whether the current best models of primate vision, a family of deep convolutional neural networks (DCNNs), can predict the behavioral variance observed during con-

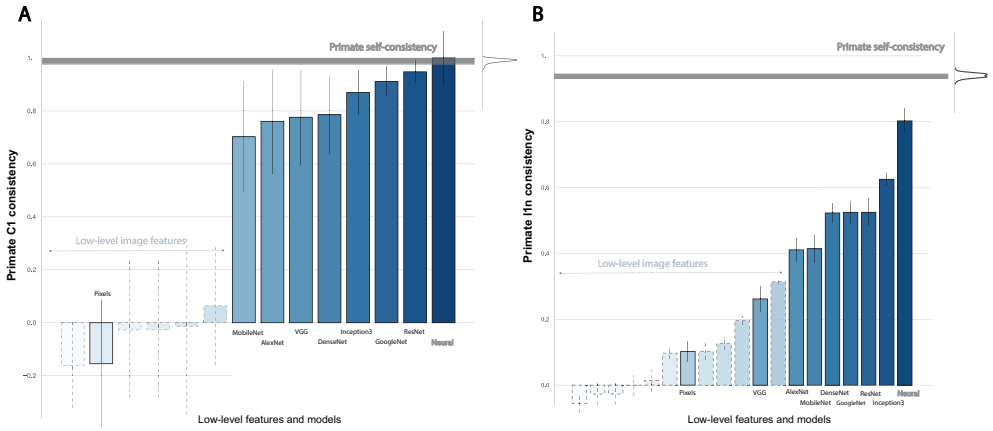

Figure 4.7: DCNNs and low-level image features primate behavioral consistency. **A** The C1 (context-level) primate consistency with the std (across image sub-sampling), for all models and low-level features. The primate self-consistency ceiling is shown in gray with its distribution across images on the right margin. Low-level image properties (from left: spectral mean, spectral std, contrast std, peak power, contrast mean) and Pixels (model of the flattened input image) do not capture the context-level primate accuracies. Most CNN models and the neural data model reach the primate consistency band for C1. **B** Same as A, but for $I1_n$, the context-corrected image-level primate consistency. Low-level image properties (from left: context category, spectral std, contrast std, contrast mean, object x-position, object y-position, spectral mean, peak power, object category, object size), and current ImageNet pretrained DCNNs do not capture the image-level behavioral accuracies of the primate behavior. The neural data model decreases the gap with primate consistency.

textual manipulations. These models also demonstrated sensitivity to contextual changes, with their accuracy varying significantly across different context types, as depicted in their C1 behavioral signature in Figure 4.3.A. Our results (Figure 4.7.A) comparing this contextual signature to that of primates suggest that most DCNNs are able to capture the primate context-level behavioral accuracy patterns (C1) and are within the primate consistency band (pooled across the human and monkey behavioral data). However, as it can be seen from the right panel, they do not fully explain the (context corrected) image-level accuracy patterns, $I1_n$, of the primates. This discrepancy arises because accuracy variations within both context and object category types are not consistently aligned between the primates and the artificial models (see supplementary Figure A.8, showing the consistency within each context type). This indicates that such models do not currently possess the mechanisms required to process scene context in a primate-like fashion.

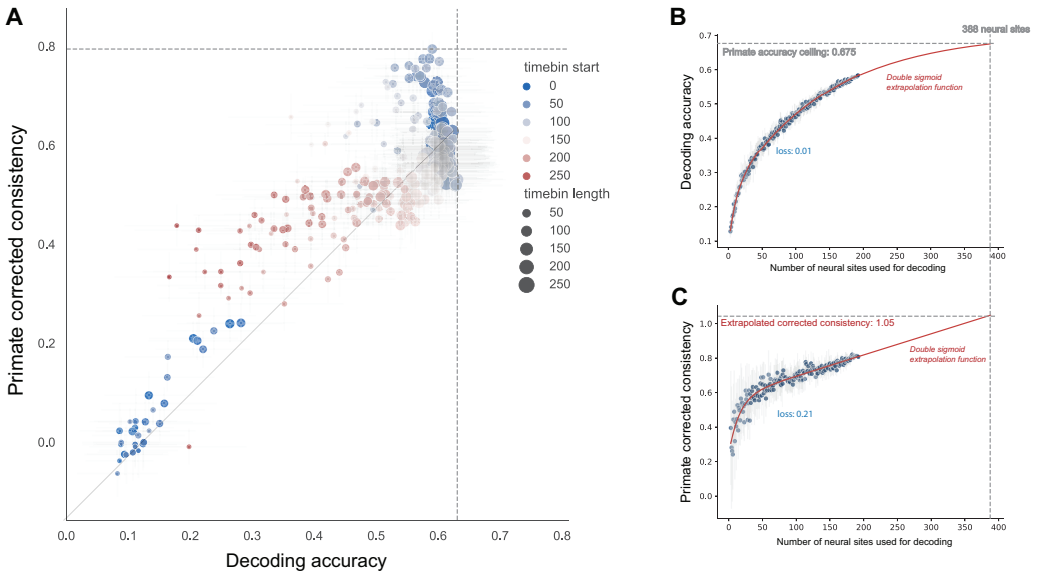

Figure 4.8: Extrapolation for the primate consistency of neural data. A Showing the corrected (by each time-bin's internal reliability) primate consistency as a function of the decoding accuracy for each time-bin of the neural data. We used all 192 neural sites to decode for every point. The color, from dark blue to dark red indicates the start and the size of the point indicates the length of the average time-frame used. We filtered the unreliable bins (internal reliability<0.2) that would drive up the primate corrected consistency beyond 0.5. The lights gray lines indicate the standard deviation for the decoding accuracy and primate consistency respectively, across different randomization of images used for training and testing the decoder (one-vs-all classifiers). **B** The decoding accuracy using the 70-170ms averaged time-frame, as a function of the number of neurons used for the accuracy decoding. A double sigmoid function (eq. 4.1) is used to fit the points with a loss of 0.01 (sum of squared residuals). Based on the extrapolation, 388 neural sites are needed for the neural data to reach primate accuracy of 0.675. The light gray lines indicate the standard deviation across different sub-sampling of neural sites used for decoding. **C** Similar as A, but extrapolating the neural data's corrected consistency with the primate accuracy. The same double sigmoid function is used (loss 0.21), to fit the points and extrapolate the consistency to 388 neural sites. The extrapolated value, 1.05 (higher that 1 due to the reliability correction), reaches the primate consistency. The light gray lines indicate the standard deviation across different sub-sampling of neural sites used for decoding. (Supplementary Figure A.9 shows the same plots without the standard deviation, for more clarity)

Comparison with neural data models

In an attempt to reduce this image-level consistency gap, we decoded the neural data from a context-naive monkey during passive fixation recorded in IT (responsible for core object recognition [25]) and compared this model's alignment with primate behavior. The neural data accurately predicted the context-level primate behavioral patterns, similar to most state-of-the-art models. Interestingly, the neural data model outperformed all CNN model features,

inching closer to closing the primate consistency gap. To factor in potential constraints arising from the number of neural sites, we extrapolated the primate consistency (corrected by the neural self-reliability) based on the pool of neural responses used. The extrapolation method, using a double sigmoid function (equation 4.1) is shown in Figure 4.8. We first found the optimal time interval to use: 70-170ms, by calculating the decoding accuracy and consistency for each possible interval of at least 10ms and up to 290 ms, then filtering unreliable frames and sorting based on the primate consistency. We then extrapolated the decoding accuracy using this averaged 70-170ms response, as a function of the number of neural sites. Figure 4.8.B shows the number of neural sites extrapolated to reach primate accuracy, found to be 388. Similarly, the primate consistency (for the same time-frame), was extrapolated for this number of neurons. This extrapolated value reached the primate ceiling, indicating that with a larger pool of neural data this model will likely mirror context-trained macaque behavior. This indicates a potential path forward: by aligning artificial models with neural data, we might foster a closer behavioral alignment with primates[21].

$$\sigma_{a,b,c}(x) = \frac{a}{1 + e^{-(x-b)/c}}$$
$$\sigma_{double}(x) = \sigma_{a1,b1,c1}(x) + \sigma_{a2,b2,c2}(x) + d \tag{4.1}$$

Our findings underscore the importance of context in real-world object recognition. We establish rhesus macaques as an appropriate animal model to study the effect of scene context in human visual object recognition, and lay the ground-work for further exploration of the neural mechanisms behind contextual modulation. We show that context has a strong effect on the neural predictivity of artificial models. Additionally, these deep CNNs show a substantial explanatory gap for image-level comparisons with primates across contextual manipulations. Our results highlight the necessity of refining deep neural network models to more accurately capture the intricacies of contextual influences on visual object recognition, which could potentially be achieved by creating more IT-aligned models.

5

IT aligned CNNs

While deep neural networks have demonstrated remarkable performance in computer vision tasks[78,119,115,45,28,87], they are very fragile in generalizing to simple image distortions[121,14,15,107,11]. Conversely, the visual system of primates displays exceptional resistance to various perturbations. The deep CNNs significantly differ from humans in their classification behavior towards images that have undergone human-imperceptible and non-random perturbations (typical adversarial attacks), as they can cause the models to misclassify images despite correctly classifying the unperturbed ones, resulting in poor robustness. By aligning the model features to the extracted neural responses of the primates when viewing that same image, we could force the model activations to become more 'brain-like' which could reduce this mismatch. Prior research has shown that aligning CNNs to the primary visual cortex (V1) improves their robustness[22,32,109,84]. These studies focused on the early-stage visual responses, whereas primate visual object recognition is critically supported by the late-stage visual processing region of the primate ventral stream - the IT region. Recently Dapello et al., 2022[21]

have been the first to explore this, by demonstrating that IT alignment improves adversarial robustness, neural predictivity across subjects and human behavioural similarity.
Previous research faced constraints, due to their limited range in diversity of visual stimuli and image distortion, and access to neural data. We adopted a data-intensive approach, mitigating these limitations by using an extensive collection of recordings from over 4300 neural sites across five macaques, coupled with approximately 700k images with a distribution similar to that one of ImageNet. We also tested a wide range of image distortions beyond only adversarial attacks. Furthermore, our method leverages a simplified architecture and neural loss function, which harmonizes more effectively with the linear alignment method. We introduce two models based on an ImageNet pre-trained AlexNet architecture. A "Neural model", fine-tuned to predict neural activity in macaque inferior temporal cortex (IT) in reaction to some natural stimuli, and a "Multitask model" simultaneously tuned to perform this neural prediction and the standard image classification. We assessed these models' out-of-distribution generalization capabilities and their resilience to common image distortions and adversarial attacks. We investigated the impact of augmenting the weight of the neural loss relative to the classification loss in the joined task, on both generalization and robustness. To ensure the observed effects were not simply a result of extended image classification training, we evaluated a "Control model", fine-tuned with the same amount of data, purely on ImageNet.

5.1 Data and methods

5.1.1 Macaque neural data collection and processing

The neural activity was collected from 5 adult macaques, in the span of many sessions, while the monkeys were passively fixating. The recordings were done in a larger timeframe, with the chronic array exact locations varying across different sessions, but always within IT. Animals were implanted with custom floating microelectrode arrays (32 channels, MicroProbes, Gaithersburg, MD or 128 channels, NeuroNexus, Ann Arbor, MI) or microwire bundles (64 channels; MicroProbes). Neural signals were amplified and sampled at 40 kHz using a data acquisition system (OmniPlex, Plexon, Dallas, TX). Multi-unit spiking activity was detected using a threshold-crossing criterion. Channels containing separable waveforms were sorted online using a template-matching algorithm. All procedures were approved by the Har-

vard Medical School Institutional Animal Care and Use Committee, and conformed to NIH guidelines provided in the Guide for the Care and Use of Laboratory Animals.
Each image was presented an average of 3-4 times, we discarded the single presentation recordings and we filtered only the most consistent IT neural sites, using the average for 70-170ms (see [91]) based on their split-half reliability (amount of internal consistency) with a threshold of 0.6.

Monkey	Mean reliability	Total # reliable neurons	Total # of images	# train images	# test images
Bo	0.457±0.224	1447	240 397	50 000	160 000
Fr	0.459±0.206	1116	133 055	50 000	60 000
Lo	0.282±0.208	548	91 501	50 000	X
Re	0.517±0.176	1945	213 619	X	213 619
Pa	0.230±0.240	278	112 263	X	112 263

Table 5.1: Summary of the acquired data for each macaque. The first column shows the monkey name (ID), followed by the mean initial reliability across all sessions and the standard deviation. The next column indicates the total number of filtered reliable neural sites (self-consistency>0.6) that were used in the analysis. The next column shows the total number of images these monkeys were exposed to while recording neural activity. We used the neural responses to 50k images from the first three macaques for training. The last column shows the number of images in the testing sessions for all macaques used for evaluation.

5.1.2 Model architecture

Baseline model

All of our experiments were based on the AlexNet architecture, due to its good balance of performance and complexity. While AlexNet is simpler than newer models, it still performs exceptionally well on many tasks. The standard Alexnet architecture is shown in Figure 5.1.A.

Neural model

To create a model capable of predicting the neural data, henceforth termed the "Neural model", we modified this baseline architecture. We incorporated two linear layers subsequent to the feature extraction layer - avgpool, chosen based on the BrainScore benchmark [112,113]. The initial linear layer served to reduce the dimensionality, channeling all 9216 avgpool features into a 3k dimensional space. These features were used for all further neural predictivity analyses. The subsequent linear layer represented the output corresponding to each recorded IT

site across the three macaques. Every neural site incorporated in the training was distinctly mapped one-to-one with a specific output unit in this layer. The loss was only propagated to the lower (convolutional) layers, causing a misalignment of its convolutional and classification layers, leading us to remove the latter from this new architecture. To maintain this model's capability for image classification, we incorporated a linear support vector machine[20] (SVM) after avgpool, more precisely, after the newly added linear layer for dimensionality reduction. This allowed us to perform multiclass classification using the extracted features. The SVM was trained using the CIFAR10[76] training dataset, containing 50k images spanning 10 categories, and subsequently assessed on the 10k CIFAR10 test set.

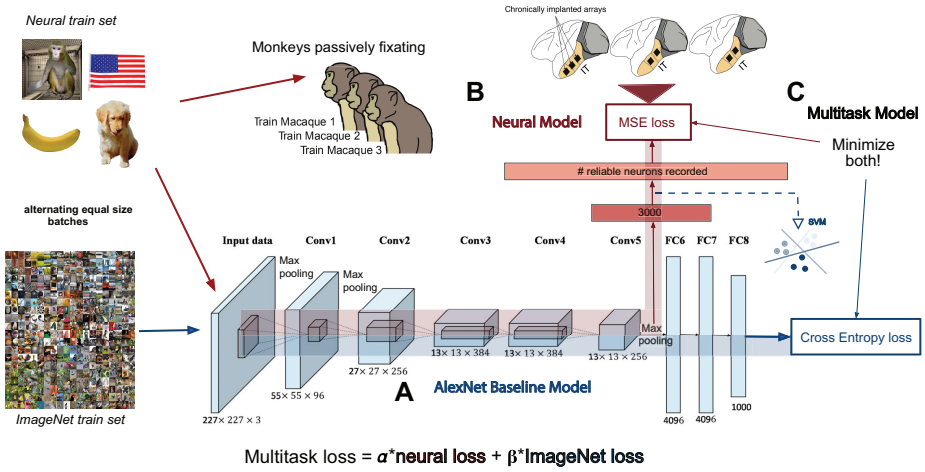

Figure 5.1: Multitask model architecture. **A** The baseline model - the standard AlexNet architecture shown in blue. It consists of eight layers, including five convolutional layers and three fully connected layers, with all convolutional layers followed by rectified linear unit (ReLU) activation functions. The output of the last convolutional layer goes through an average pooling layer (avgpool) and is fed into three fully connected layers, each with 4096 neurons. The last fully connected layer has an output size 1000, corresponding to the number of classes in the ImageNet dataset. The Control model has the same architecture as the Baseline. The light blue band shows the layers used for this architecture. **B** The Neural model, changes done following the baseline model convolutional layers are in red. The output of avgpool was fed to two new linear layers, the first for dimensionality reduction to 3k features and an output layer for the neural predictions. An SVM was added after the first linear layer, using the extracted features for CIFAR10 multiclass classification. The light blue red shows the layers used for this architecture. **C** The multitask model, using the full AlexNet architecture with the "neural enhancement". This model was jointly trained to reduce the neural and ImageNet classification loss, with their respective weights on the learning given by the α and β parameters. The loss was back-propagated on both branches shown by the light blue (L_{CE}) and light red band (L_{MSE}).

Multitask model

Our aim is to enhance the model's resilience and generalization without sacrificing performance. The neural model with its classification layers removed, is not comparable with the baseline AlexNet in terms of ImageNet classification. To address this, we adopted a multi-task learning approach, for stability, also training the model on a subset of the ImageNet train set. This multi-task model retained the architecture of the neural model but also incorporated the classification layers found in the baseline model. To preserve ImageNet accuracy and prevent catastrophic forgetting, we fine-tuned the model using both the neural data and ImageNet training data. We generated different variations of this model by changing the weight of the neural loss relative to the classification loss (parameter α in equation 5.1).

Control model

As a control we fine-tuned the baseline model with the same amount of images as in the joint task, but without co-training on the neural data. This was done to ensure that the effects observed with the joint training were not due to more training on the classification task. In simple terms, we extended the standard ImageNet training previously done for this pre-trained model for more epochs with 150k images.

5.1.3 Fine tuning

We kept the standard Cross-Entropy (L_{CE}) loss for the classification and used the Mean Squared Error (L_{MSE}) for the neural predictivity task. While training the multitask model, both tasks were alternately optimized using batches of 64 images from each reshuffled dataset images during each epoch, aggregating 300k images in total. This included 150k images from the ImageNet validation set and 150k from the neural predictivity dataset - 50k for each of the three monkeys used for training. The combined loss was the weighted sum of the L_{CE} and L_{MSE} , given by:

$$L_{total} = \alpha * L_{MSE} + \beta * L_{CE} \tag{5.1}$$

Where both parameters, α and β are set to 1 in the balanced loss scenario (standard Multitask model), $\beta = 0$ for the Neural model and $\alpha = 0$ for the Control model. For optimization, we

used Stochastic Gradient Descent (SGD) with a learning rate set to $5 * 10^{-1}$. Additionally, a learning rate scheduler was applied, with a start and end factor of 1 and 0.5 respectively, over 60 epochs.
The multiclass linear SVM underwent typical training with ten one-versus-all classifiers, incorporating squared hinge loss, a L2 norm penalty, and a regularization parameter set to one. The neural model was trained using identical parameters but was exclusively fed the 150k images from the neural set each epoch, optimizing solely through the neural MSE loss.

5.1.4 Image manipulations

To assess the robustness of our models we performed a series of common image alterations on 10k images from the ImageNet validation set, including blurring, additive noise, random image rotation and elastic transformation. Gaussian blurring was applied using a kernel of size 3 (minimal blur) and 13 (maximal blur). As part of the noise alteration we randomly flipped half of the image pixels to be 1 (salt) or 0 (pepper) with a balanced ratio of both (salt and pepper noise). We also tested the robustness by adding Gaussian-distributed additive noise. The random image rotations spanned from zero to 180 degrees.

5.1.5 Adversarial attacks

Adversarial attacks exploit machine models' vulnerabilities to small, intentional perturbations of their inputs. Despite being barely noticeable to humans, these modifications can result in the model misclassifying the input with high confidence. Projected Gradient Descent[80] (PGD) is a form of adversarial attack that has demonstrated remarkable efficacy in deceiving machine learning models[91]. PGD entails the repetitive distortion of an input in the direction of the gradient of the loss function concerning the input. Additionally, this attack constrains the distortions to remain within a designated epsilon radius (attack budget). All models including the SVM were subjected to untargeted PGD attacks with a step size of 0.1 at each iteration, using L_∞ and a maximum perturbation $\varepsilon = 0.2$ that the attacker can introduce. We used the Adversarial Robustness Toolkit[96] to perform these attacks.

5.2 Results and discussion

Our goal was to test whether aligning late ANN activations to the IT neural population, by training with a large neural dataset will improve their robustness and generalization. To evaluate the out-of-distribution neural predictivity generalization we measured the predictivity on new neural data from the same training monkeys and from completely different macaques. We then tested the robustness, by subjecting the models to a series of image alterations and attacks described above.

5.2.1 Neural data alignment improves generalization

First, we investigated whether IT aligned models would retain their neural explainability when applied to data from the same training monkeys which was left out from the fine-tuning (see Table 5.1 for details). Figure 5.2.A shows the average variance explained for both models for the left-out sessions from one training monkey. Both the Neural and Multitask model demonstrate an improvement in neural predictivity. The Neural model, tailored solely around neural data, as expected, achieves higher predictivity on the left out set, plateauing after a few epochs potentially due to the simple AlexNet architecture overfitting beyond that. We show the effects of increasing the neural loss weight (α=1, 2, 10) when training the model, linked to an increase in IT-alignment, the higher neural loss weight leads to higher explained variance of new sessions. Interestingly, the Control model, shows that further classification training with the same set caused a slight decrease in IT-likeliness.

We next investigated how these models' IT alignment affected generalization across new monkey subjects. The predictivity of our models was tested against over 325k new images (see table 5.1 for details) with neural responses from two different macaques that had not been used in the training phase. Figure 5.2.B shows the average EV for the models for a test macaque as a function of the number of training epochs. It is evident that even after a handful of epochs, there is a notable increase in the average explained variance for the IT-aligned models. During the first 5 epochs, we can see the same trend as observed before for the left-out sessions from the training monkeys, namely increasing α increases the explained variance. Yet, with further fine-tuning, it is again apparent that the models relying heavily on neural data representations start to overfit to the training sessions, leading to a drop in generalization. The

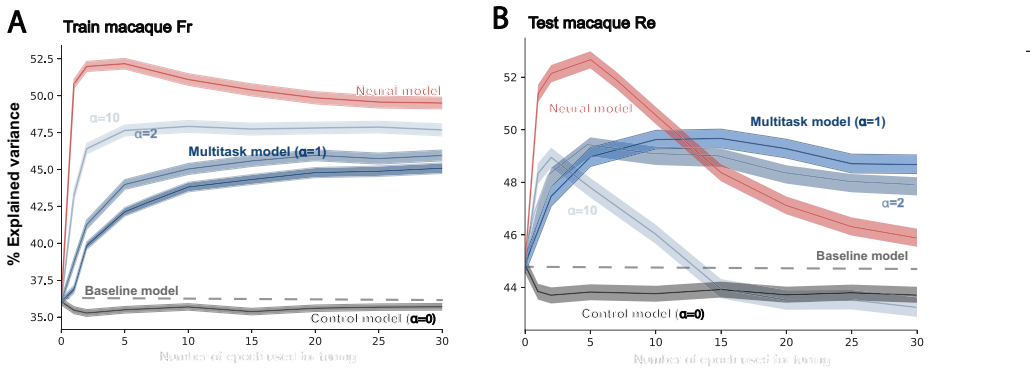

Figure 5.2: Generalization on new samples and subjects. A The percent of explained variance with the standard error across epochs of training for left out sessions of one macaque used for fine-tuning. The Baseline model performance is shown with the grey dashed line. The average EV for the multitask model is depicted in blue, with the color opacity indicating a higher weight to the neural loss. The Neural and the Control model are shown in red and black respectively. **B** Same as A but for a left out macaque not used in training.

multitask model, while not matching the performance of the neural model, shows a more gradual EV decline. This suggests that the joint training process inherent in the multitask approach acts as a mitigating factor, reducing the propensity for the model to overfit. These effects can also be seen on the other left-out subject (supplementary Figure A.11.B). Here, the trend of generalization is present with lower overall performance increase for all models. The more neurally aligned multitask model (α=10) fails to generalize and performs worse than the baseline model. Nevertheless, since this macaque's data is far less reliable (mean consistency 0.23, Table 5.1) with only 278 reliable neural sites, about 15% of the ones in Re, it poses a more challenging model to fit.

Taken together, these results demonstrate that our IT tuning method increases generalization by improving the IT-likeness in our models for held out sessions and completely new subjects.

5.2.2 Neural data alignment improves robustness

We evaluated our Multitask model on a series of image manipulations on the ImageNet validation set.

As depicted in Figure 5.3, our model showed improvement in ImageNet top-1, 5, and 10 accuracy scores across a majority of these manipulations as well as on the non-altered test set. Notable improvements include an increase of 27.67% when applying elastic transformations

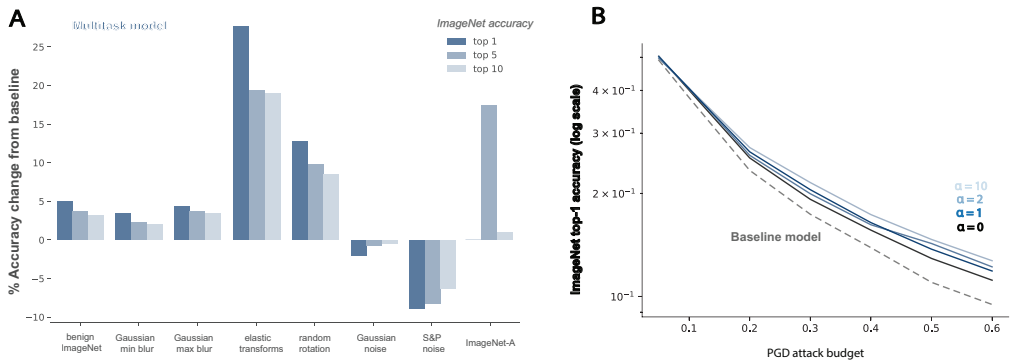

Figure 5.3: Performance of the multitask model on ImageNet manipulations and adversarial attacks. **A** We show the percent of ImageNet top-1, top-5 and top-10 accuracy change from the Baseline model for the Multitask model tuned to 15 epochs. Results for the benign (non-altered) ImageNet 10k validation subset, the same set after applying Gaussian blurring with a kernel of size 3(min) and 13 (max), elastic transformation, random rotation of the images, adding Gaussian noise, adding Salt and Pepper noise and results for ImageNet-A, a subset of ImageNet where state-of-the-art models perform very poorly. **B** ImageNet top-1 accuracy as a function of the strength of the PGD adversarial attack. The range of the attack budget (ε) is shown on the x-axis. The color scheme from darker to lighter blue indicates models with increasing α parameter value. All models were taken after 15 epochs of fine-tuning.

and 12.74% for random image rotations in top 1 accuracy. However, when compared with the baseline pre-trained AlexNet, our multitask model was more sensitive to noise interference. This is seen in the minor dip in performance upon the introduction of Gaussian noise and a more pronounced decrease when subjected to S&P noise. However, when we evaluated the purely Neural model on the CIFAR10 noisy altered test-set[47], this model outperformed the baseline for all noise distortion severity for both shot and Gaussian noise (supplementary Figure A.13). The difference in noise robustness for the Multitask and Neural model might stem from the relative simplicity of the CIFAR10 dataset compared to ImageNet, or it could be attributed to the variances in classification methodologies, given that SVM classification isn't in line with traditional AlexNet classification techniques. We also tested the Multitask model performance on ImageNet-A[48], a similar dataset as ImageNet, but far more challenging for existing models, designed to expose their weaknesses. The IT-aligned model outperforms the baseline for his dataset, with an increase of 17.4 % top-5 performance, there was no change in top-1 accuracy.

We then performed a series of white-box adversarial attacks on the models to evaluate their adversarial robustness. The performances of the Neural model facing an untargeted PGD

attack directed at the SVM, is shown in Figure 5.4.A. The SVM was trained on CIFAR10 (see 5.1.2) and evaluated both on the benign (non-altered) CIFAR10 tested and the PGD modified version. Remarkably, our Neural model, exhibiting better IT alignment, is far less susceptible to this adversarial attack than the baseline (shown in Figure 5.4 with the dashed line), in fact its performance stays very close to the one on the benign set while the baseline model is severely affected and falls at 20%. After plotting the performance as a function of the attack strength (shown in supplementary Figure A.12), we can see a more linear, far slower drop for the Neural model compared to the exponential decay for the Baseline model.

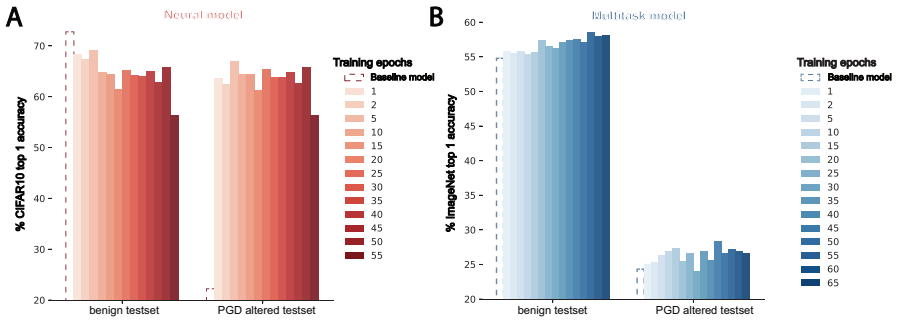

Figure 5.4: Models show improved robustness against adversarial attacks.A Top-1 performance of the Neural model on the CIFAR10 benign and PGD altered testset, across tuning epochs. The dashed bar shows the performance of the Baseline model for both datasets. The PGD attack was done on the SVM used for classification.**B** Similar as A, but for the Multitask model and its performance on the ImageNet 10k validation set. The PGD attack was done on the model itself.

Additionally, we performed a PGD attack directly on the multitask model when evaluating it on a 10k ImageNet validation subset. The results in Figure 5.4.B show that aligning the model to fit the IT data boosts its performance, evident both on the benign as well as the PGD-altered ImageNet subset. We also tested the performance of this model with varying IT-likeness (based on the parameter α), while increasing the PGD attack budget. Figure 5.3 shows that our model is more robust than baseline and the Control model, and for higher perturbation ($\varepsilon > 0.2$), we see a trend of increased robustness linked to the IT-similarity intensity. These findings reinforce the idea that neural data alignment can improve models, making them more robust to image distortions and adversarial attacks.

Our work highlights several directions for future study. The neural dataset we used was relatively limited in terms of average trials per image. It would be insightful to see the effects on

the model performance of increasing it, by adding more trials, neurons, images and different subjects . It is also worth noting that we have not yet fully explored how different, more complex model structures, training data, or learning techniques might affect the model's ability to mirror macaque IT responses. It is important to point out that while our method does increase robustness, the actual extent of improvement is relatively small. A promising direction for improving robustness is having a model of the full primate ventral stream, by also aligning the earlier and intermediate model features to the primary visual cortex (V1), V2 and V4. Although these models are still far behind the generalization and robustness power of the primate visual system, our work is a conceptual step towards bridging the gap between artificial and biological vision.

6

Putting it all together

The work presented in this thesis was motivated by the broad scientific goal of discovering models that quantitatively explain the neuronal mechanisms underlying primate invariant object recognition behavior.

To this end, previous research has shown that deep convolutional models are able to explain neural representations of the early and late ventral visual stream. Such models are often evaluated with neurobehavioral datasets where the stimuli are presented in the subjects' central field of view (FOV). However, the exact visual angle often varies widely across studies. A unified model of the primate visual system cannot have a varying FOV. Similarly, the type of images used for model evaluation varies across studies, ranging from objects embedded in randomized contexts to objects with no and textured contexts. Here we systematically tested how the predictivity of macaque inferior temporal (IT) neurons by DCNNs depends on the FOV and the image-context, as well as methods to increase this alignment and all the improvements that follow.

In chapter 3 we evaluated current state of the art CNNs to estimate their optimal FOV. We performed large-scale recordings in one macaque (≈90 IT sites) while the monkey passively fixated on images presented at a large 20 degree visual angle. To estimate the most neurally aligned visual angle for the DCNNs, we compared the DCNN IT predictivity at varying image crop sizes. We observed that 7-11 visual degree center crops produced the strongest DCNN IT predictions across all models, suggesting a critical need for constraints on the visual angle when predicting animal neural data. Some future directions include evaluating these effects using a broader and varied set of stimuli and assessing the influence of the FOV on a behavioural level. Additionally, it would be beneficial to asses artificial models that emulate foveal-peripheral vision [89], drawing inspiration from the retino-cortical mapping observed in primates. Such models might be immune to the visual angle effects shown for current state-of-the-art DCNNs.

Next, in chapter 4.2.1 we looked at the effects of image-context on the DCNN neural alignment. To test this, we generated a dataset of ten contextual variations including full-context, no-context, and incongruent-context. The DCNN's IT predictivity was significantly higher for no-context compared to the full/incongruent-context images. This raises the question of whether there are more complex, putative recurrence signals used for contextual processing in IT which cannot be captured by these feed-forward models. These strong effects provide critical constraints within the experimental design to guide the development of more brain-aligned artificial models. A natural future step would be to evaluate recurrent models, as well as novel advanced models like vision transformers [28], given their remarkable capacity for global context understanding, image segmentation and state-of-the-art object recognition performance.

Moving from physiology, in chapter 4.2.2 we investigated these models' primate behavioural consistencies affected by scene-context in visual object recognition. Our findings reveal a consistent contextual modulation in the behavioral accuracies of both humans and monkeys, with a significant correlation between the two species observed at the image-level. Notably, these results cannot be attributed to low-level image features. Furthermore, current deep neural network models do not adequately predict the overall effects of context on primate behavioral performance. Models fall within the primate self-consistency zone for context-level similarities, however they show a substantial explanatory gap for image-level comparisons across contextual manipulations. Our research highlights the importance of understanding contex-

tual influences on visual object recognition and the need to refine current models to accurately capture the critical role of contextual reasoning during real-world vision.

Finally, in chapter 5, we adopt a data-driven approach to further align these models to the macaque Inferior Temporal (IT) Cortex, in hopes to achieve a more robust, primate-like behavior. We model spiking activity in IT using two architectural approaches: one that is exhaustively fine-tuned using neural data, and a multitask model trained simultaneously on the classification task to preserve accuracy on object recognition tasks. Our evaluation on out-of-sample non-human primates (macaques) validates the generalization of our fine-tuning approach, showing enhanced IT representational similarity. After a series of adversarial attacks and image manipulations, we provide evidence for increased robustness due to this neural alignment, bringing these models closer to primate object recognition behaviour. Moving forward, this method should be assessed on more complex models and a broader scope of tasks, including generalization challenges such as one-shot learning. Moreover, it would be insightful to also incorporate novel techniques, such as "neural harmonization[33]", by aligning ANNs with human visual strategies , which could lead to more human-like neural and behavioural patterns.

While modern artificial neural networks continue to excel in computer vision tasks such as object recognition, they seem to diverge further from their biological counterparts. Aligning these two networks leads to mutual benefits — creating more resilient and universal artificial intelligence, and enriching our understanding of visual processing in the brain. Taken together this work is a step forward in bridging the gap between primate and artificial vision. It lays the foundation for further exploration in building more primate-like, generalizable and robust artificial models, offering deeper insights into the complexities of the brain.

A

Supporting Information

A.1 Alternative neural similarity metrics

A.1.1 Representational Similarity Analysis

The representational similarity analysis (RSA)[75] is another metric for assessing the alignment between the model features and neural population. This approach begins by building an Representation Dissimilarity Matrix (RDM) for each population, detailing the variability in responses to each image pair. This matrix effectively captures the unique representational characteristics of the population. The degree of similarity between two distinct populations is subsequently obtained by correlating their respective variability matrices.

A.1.2 Centered Kernel Alignment

Centered Kernel Alignment (CKA) provides a technique to assess the similarity between high-dimensional representations. Leveraging kernel methods, CKA captures the alignment between two datasets in their respective feature spaces. This makes it an effective tool for comparing not only neural network layers but also a variety of datasets where a comparison of underlying structures or patterns is needed.

A.1.3 Single-Unit Neural Correlation

Another metric measuring the alignment between artificial and biological models of vision, pioneered in 2018, is the neural correlation. It is defined as the "correspondence at the level of the population code: [where] stimulus category can be partially decoded from real neural responses using a classifier trained purely on a matched population of artificial units in a model"[4]. This metric essentially evaluates the one-to-one(or one-to-few) mappings between single units in a deep neural network model and neurons in the brain.

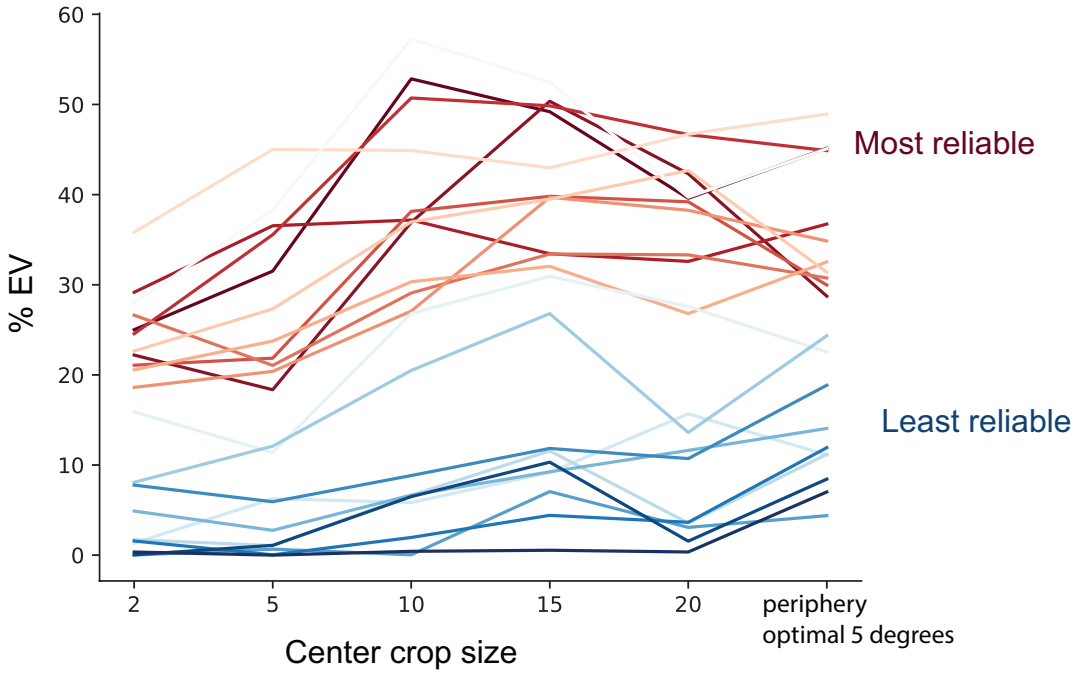

Figure A.1: Showing the explained variance for each of the "periphery preferring" neurons, neurons whose responses were better predicted when the model was presented 5 degree periphery crops. The x-axis shows the center crop size in degrees of visual angle, with the last tick showing the optimal periphery 5 degree value for that neuron. The color indicates self-reliability for the neural site from dark blue to dark red, least to most reliable.

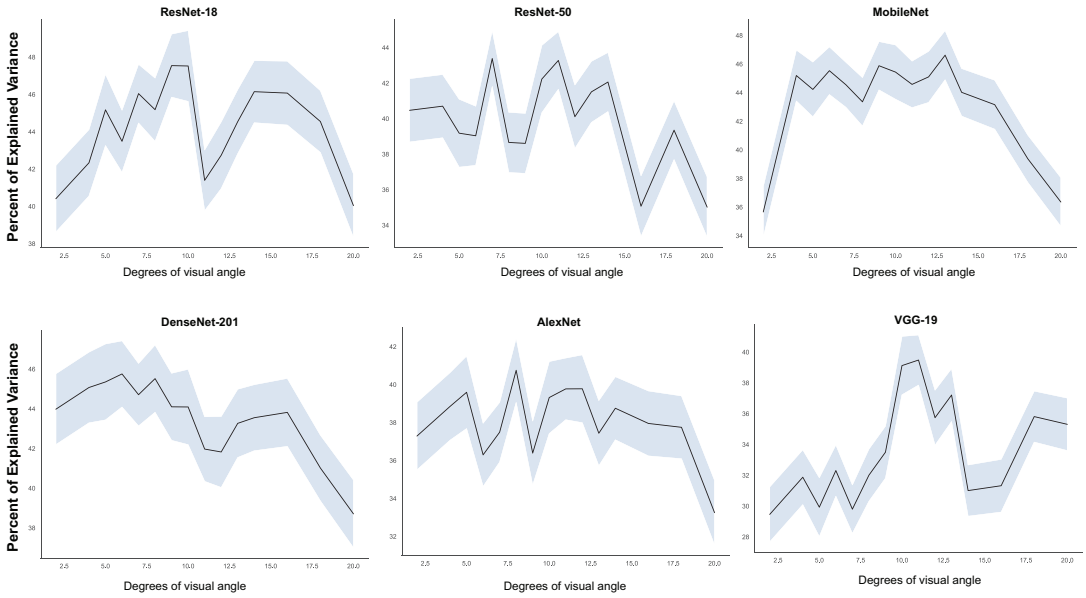

Figure A.2: Explained variance of all models as a function of increasing degrees of visual angle. Shows the average percent of EV per model, when changing the input image center crop size. The standard error is shown with a blue band.

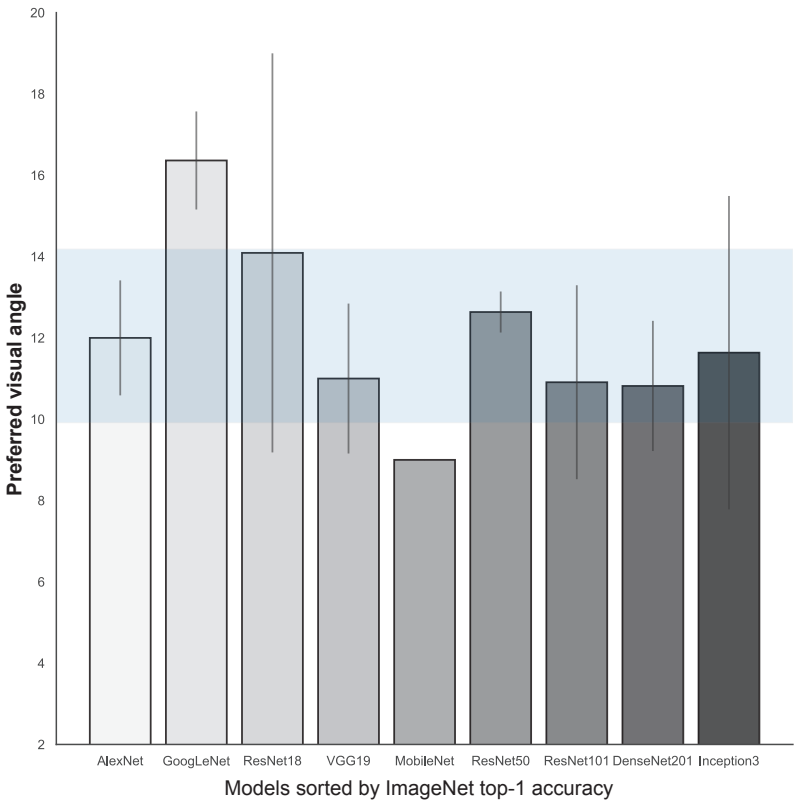

Figure A.3: Similar as Figure 3.3, but showing the "optimal" visual angle per model when explaining neural responses to full context images. The mean angle across models is 12.01 ± 2.13 degrees.

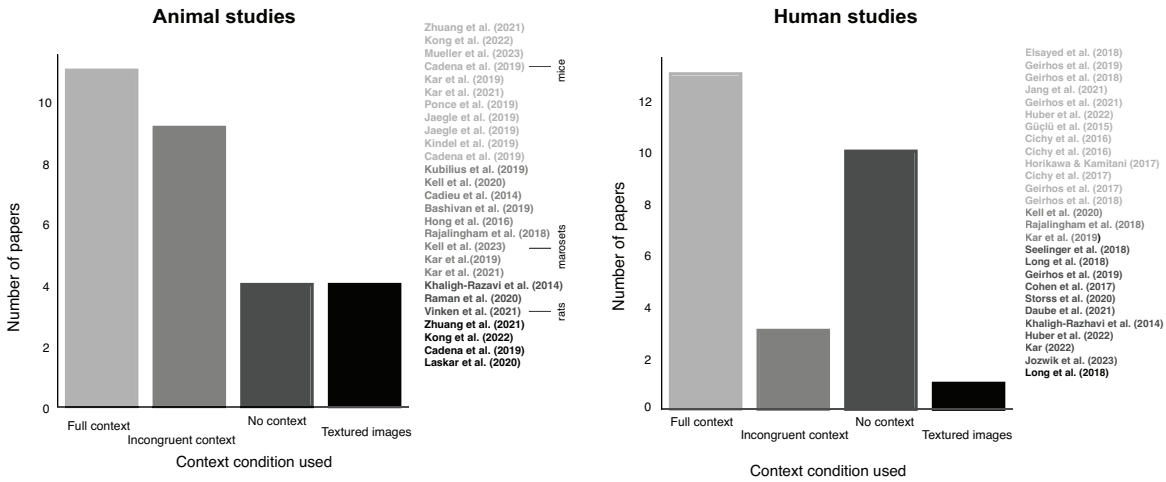

Figure A.4: Survey of studies evaluating the alignment between ANNs and empirical data. A review of studies done over the past decade predicting animal neural or behavioural responses to a visual stimuli. For each study we show the image-context reported during stimulus presentation. The left panel shows non-human animal studies and the right panel shows studies that used human subjects. The varying context is present for studies across species as well as within. The same studies were used as for Figure 3.1.

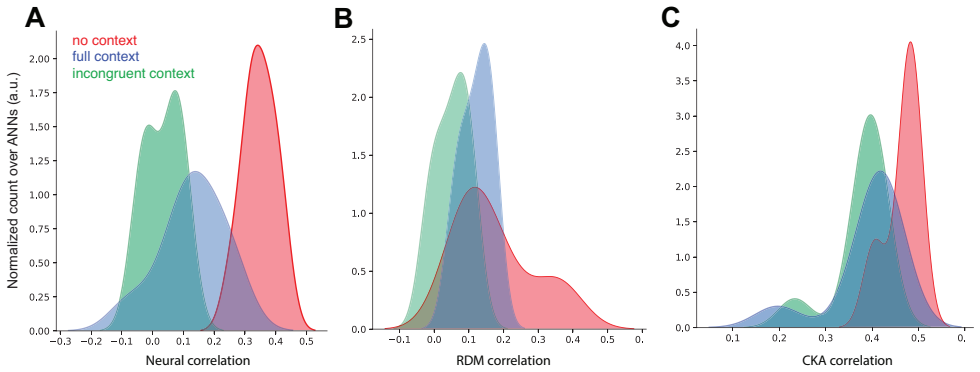

Figure A.5: **A** Neural correlation - finding most correlated artificial units for each site on a train set of images and evaluating this correlation on a left-out test set in a cross validated way. Shows the distribution of the test correlation for all models, grouped by context type. **B** Similar as A, shows the correlation distribution for all models, for the RSA metric - correlation of the neural and artificial units' Representation Dissimilarity Matrix (RDM) matrices. **C** Similar, but showing the Centered Kernel Alignment (CKA) correlation, the results shown are using Radial Basis Function (RBF) kernel, similar results where obtained with a linear kernel.

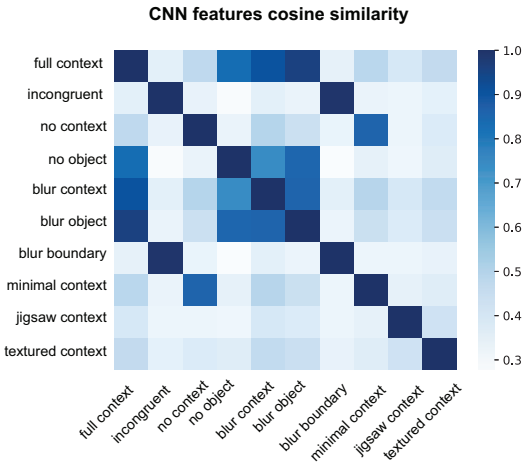

Figure A.6: Cosine similarity for DCNN features exposed to context. The cosine similarity matrix for the extracted model features compared pairwise for each context type for each image. The result shown is the mean across all images and all models.

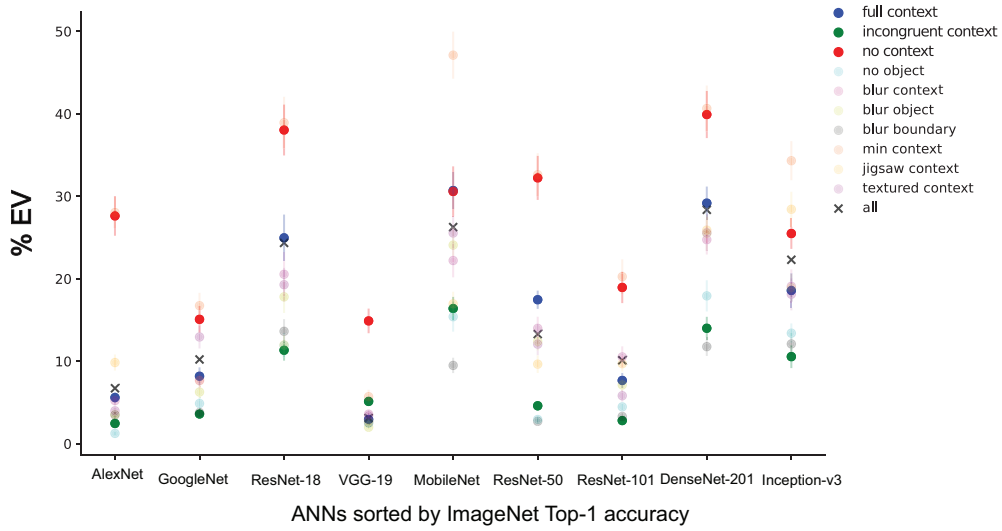

Figure A.7: Shows the mean explained variance with standard deviation across images for every evaluated model, across different context categories. Every point indicates the averaged predictive power (neural population EV) of that model, on all 60 images of the specific context indicated by the color of the point.

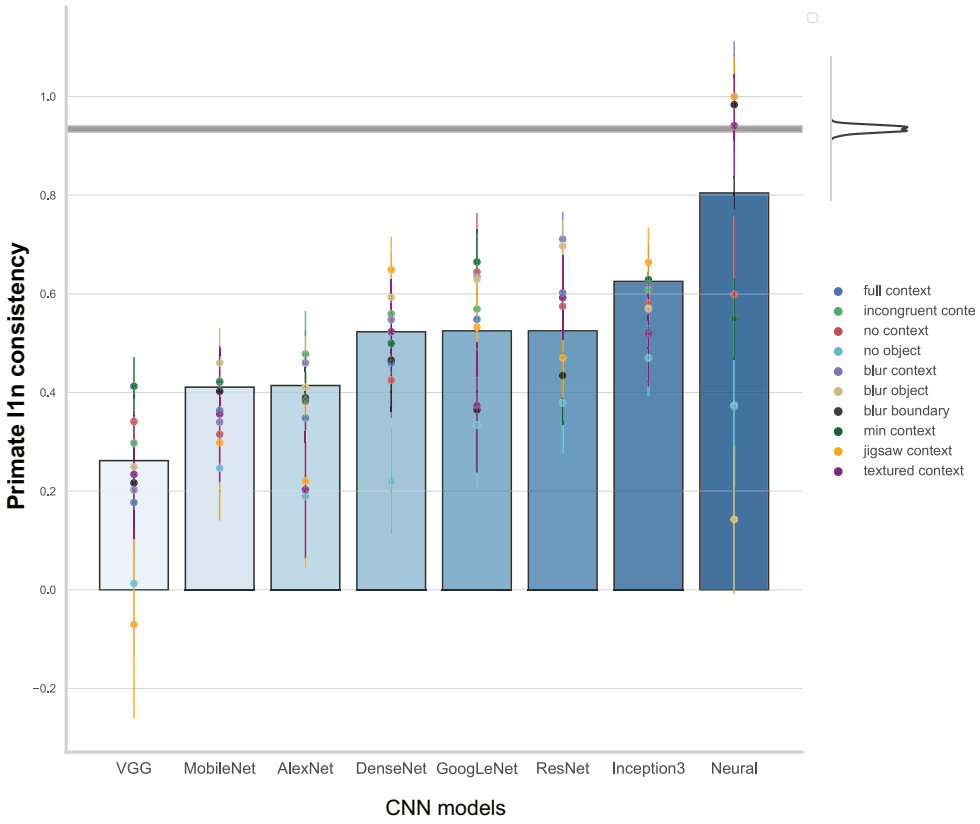

Figure A.8: Shows the image-level ($I1_n$) primate consistency for all models. The mean consistency with standard deviation when using all images is shown by the bars, similar to Figure 4.7.B. The points (color coded for context categories) show the mean consistency within context types, each one indicating the mean primate consistency of the model only for the images of that context type.

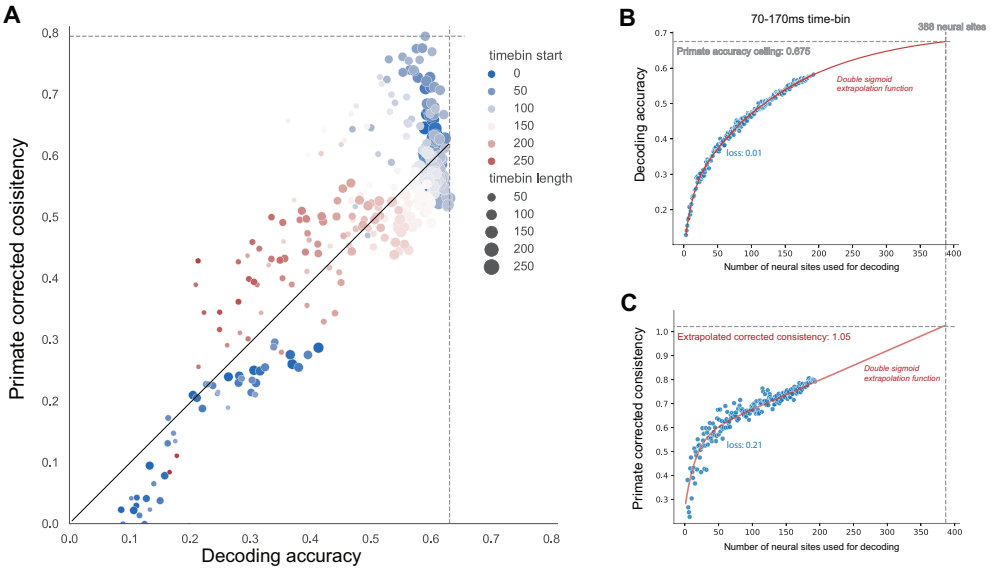

Figure A.9: Extrapolation for the primate consistency of neural data (without std). Same as Figure 4.8 but without the standard deviation, for more clarity.

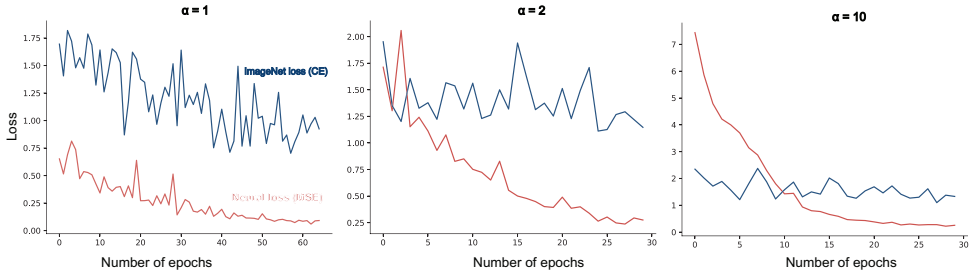

Figure A.10: Loss during the multi-task training. Shows the training loss on both tasks, neural, L_{MSE}. and classification L_{CE}. The different panels show the loss when changing the α value used as a weight parameter on the neural loss.

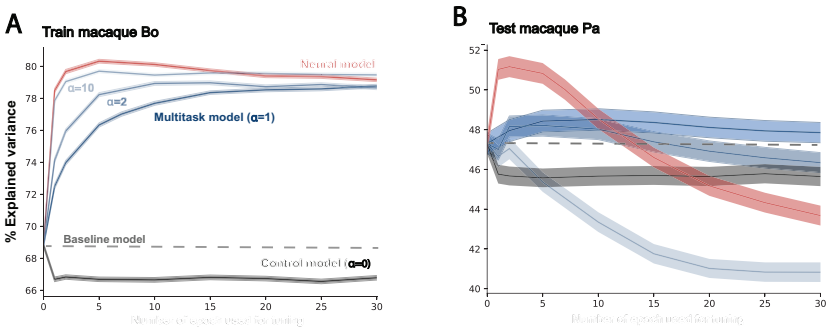

Figure A.11: Model neural predictivity for another train and test macaque. A shows the EV (and standard error) change as a function of neural alignment with the training data (number of fine tuning epochs) for left-out session of a macaque used for training (Bo). The color indicates the model type evaluated. **B** Similar but for a new subject (Pa) not used during training.

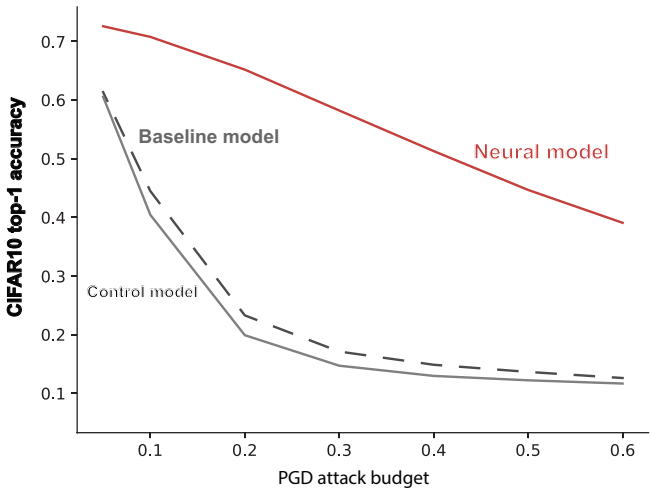

Figure A.12: Model performance as a function of the PDG attack strength on the SVM. CIFAR10 accuracy as a function of the attack budget on the SVM for the baseline (dashed line) and neural model(red).

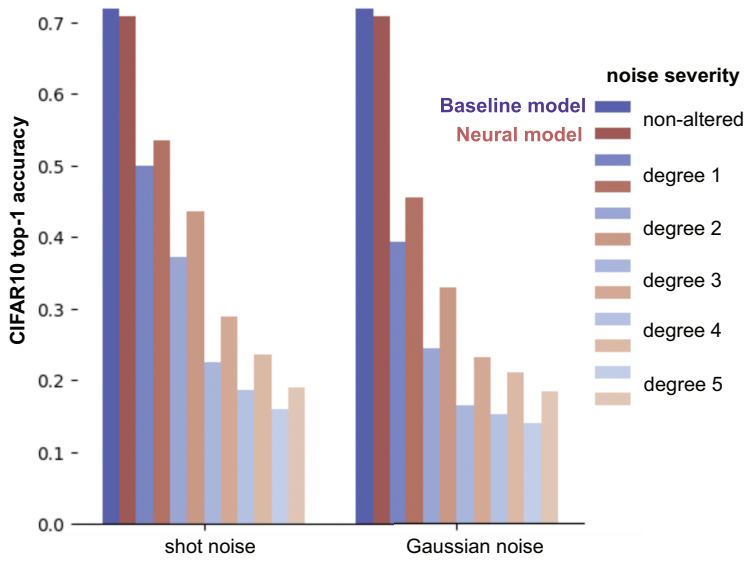

Figure A.13: CIFAR10 top-1 performance of the Neural and Baseline model for Gaussian and shot noise. The top-1 performance on the CIFAR10 testset, altered by adding Gaussian or shot noise, the severity of the noise is indicated by the opacity descending. The Baseline model performance is shown in blue and the Neural model in red. The altered CIFAR10 test-set was obtained from Hendrycks et al. [47].

References

[1] Alitto, H. J. & Usrey, W. M. (2015). Surround suppression and temporal processing of visual signals. *Journal of neurophysiology*, 113(7), 2605–2617.

[2] Amit, D. J., Bernacchia, A., & Yakovlev, V. (2003). Multiple-object working memory—a model for behavioral performance. *Cerebral Cortex*, 13(5), 435–443.

[3] Anand, A., Sen, S., & Roy, K. (2021). Quantifying the brain predictivity of artificial neural networks with nonlinear response mapping. *Frontiers in Computational Neuroscience*, 15, 609721.

[4] Arend, L., Han, Y., Schrimpf, M., Bashivan, P., Kar, K., Poggio, T., DiCarlo, J. J., & Boix, X. (2018). *Single units in a deep neural network functionally correspond with neurons in the brain: preliminary results*. Technical report, Center for Brains, Minds and Machines (CBMM).

[5] Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, 5(8), 617–629.

[6] Bar, M. & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron*, 38(2), 347–358.

[7] Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439), eaav9436.

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

[9] Bengio, Y. (1993). A connectionist approach to speech recognition. *International journal of pattern recognition and artificial intelligence*, 7(04), 647–667.

[10] Bracewell, R. N. & Bracewell, R. N. (1986). *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York.

[11] Brendel, W., Rauber, J., Kümmerer, M., Ustyuzhaninov, I., & Bethge, M. (2019). Accurate, reliable and fast robustness evaluation. *Advances in neural information processing systems*, 32.

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

[13] Cadena, S. A., Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walker, E. Y., Reimer, J., Bethge, M., Tolias, A., & Ecker, A. S. (2019b). How well do deep neural networks trained on object recognition characterize the mouse visual system? In *Real Neurons {\&} Hidden Units: Future directions at the intersection of neuroscience and artificial intelligence@ NeurIPS 2019*.

[14] Carlini, N. & Wagner, D. (2016). Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*.

[15] Chen, P.-Y., Sharma, Y., Zhang, H., Yi, J., & Hsieh, C.-J. (2018). Ead: elastic-net attacks to deep neural networks via adversarial examples. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

[16] Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 153, 346–358.

[17] Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016a). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1), 27755.

[18] Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016b). Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. *arXiv preprint arXiv:1601.02970*.

[19] Cohen, M. A., Alvarez, G. A., Nakayama, K., & Konkle, T. (2017). Visual search for object categories is predicted by the representational architecture of high-level visual cortex. *Journal of neurophysiology*, 117(1), 388–402.

[20] Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20, 273–297.

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

[22] 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. *Advances in Neural Information Processing Systems*, 33, 13073–13087.

[23] Daube, C., Xu, T., Zhan, J., Webb, A., Ince, R. A., Garrod, O. G., & Schyns, P. G. (2021). Grounding deep neural network predictions of human categorization behavior in understandable functional features: The case of face identity. *Patterns*, 2(10).

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

[25] DiCarlo, J. J. & Johnson, K. O. (1999). Velocity invariance of receptive field structure in somatosensory cortical area 3b of the alert monkey. *Journal of Neuroscience*, 19(1), 401–419.

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

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

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

[29] Dvornik, N., Mairal, J., & Schmid, C. (2018). Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 364–380).

[30] Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152, 184–194.

[31] Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, 31.

[32] Federer, C., Xu, H., Fyshe, A., & Zylberberg, J. (2020). Improved object recognition using neural networks trained to mimic the brain's statistical properties. *Neural Networks*, 131, 103–114.

[33] Fel, T., Rodriguez Rodriguez, I. F., Linsley, D., & Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems*, 35, 9432–9446.

[34] 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.

[35] Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*.

[36] 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.

[37] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018a). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

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

[39] Glaser, J. I., Benjamin, A. S., Chowdhury, R. H., Perich, M. G., Miller, L. E., & Kording, K. P. (2020). Machine learning for neural decoding. *Eneuro*, 7(4).

[40] Goertzel, B. (2014). Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1), 1.

[41] Goh, J. O., Siong, S. C., Park, D., Gutchess, A., Hebrank, A., & Chee, M. W. (2004). Cortical areas involved in object, background, and object-background processing revealed with functional magnetic resonance adaptation. *Journal of Neuroscience*, 24(45), 10223–10228.

[42] 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.

[43] Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, 95(2), 245–258.

[44] Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3), 416–434.

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

[46] Helland, I. S. (1990). Partial least squares regression and statistical models. *Scandinavian journal of statistics*, (pp. 97–114).

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

[48] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15262–15271).

[49] Henry, C. A., Joshi, S., Xing, D., Shapley, R. M., & Hawken, M. J. (2013). Functional characterization of the extraclassical receptive field in macaque v1: contrast, orientation, and temporal dynamics. *Journal of Neuroscience*, 33(14), 6230–6242.

[50] Herzog, M. H. & Clarke, A. M. (2014). Why vision is not both hierarchical and feedforward. *Frontiers in computational neuroscience*, 8, 135.

[51] Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.

[52] Horikawa, T. & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1), 15037.

[53] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861.

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

[55] Hubel, D. & Wiesel, T. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160, 106–154.

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

[57] Huber, L. S., Geirhos, R., & Wichmann, F. A. (2022). The developmental trajectory of object recognition robustness: children are like small adults but unlike big deep neural networks. *arXiv preprint arXiv:2205.10144*.

[58] Hubel, D. H. & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1), 106–154.2.

[59] Hubel, D. H. & Wiesel, T. N. (1963). Shape and arrangement of columns in cat's striate cortex. *The Journal of Physiology*, 165(3), 559–568.2.

[60] 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.

[61] Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., & Rust, N. (2019). Population response magnitude variation in inferotemporal cortex predicts image memorability. *Elife*, 8, e47596.

[62] Jang, H., McCormack, D., & Tong, F. (2021). Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images. *PLoS biology*, 19(12), e3001418.

[63] Januszewski, M., Kornfeld, J., Li, P. H., Pope, A., Blakely, T., Lindsey, L., Maitin-Shepard, J., Tyka, M., Denk, W., & Jain, V. (2018). High-precision automated reconstruction of neurons with flood-filling networks. *Nature methods*, 15(8), 605–610.

[64] Johnson, K. O., Hsiao, S. S., & Yoshioka, T. (2002). Neural coding and the basic law of psychophysics. *The Neuroscientist*, 8(2), 111–121.

[65] Jozwik, K. M., Kietzmann, T. C., Cichy, R. M., Kriegeskorte, N., & Mur, M. (2023). Deep neural networks and visuo-semantic models explain complementary components of human ventral-stream representational dynamics. *Journal of Neuroscience*, 43(10), 1731–1741.

[66] Kar, K. & DiCarlo, J. J. (2021). Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, 109(1), 164–176.

[67] Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature neuroscience*, 22(6), 974–983.

[68] Kell, A., Bokor, S., Jeon, Y., Toosi, T., & Issa, E. (2020). Brain organization, not size alone, as key to high-level vision: Evidence from marmoset monkeys.

[69] Kell, A. J., Bokor, S. L., Jeon, Y.-N., Toosi, T., & Issa, E. B. (2023). Marmoset core visual object recognition behavior is comparable to that of macaques and humans. *Iscience*, 26(1).

[70] Kelley, H. J. (1960). Gradient theory of optimal flight paths. *Ars Journal*, 30(10), 947–954.

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

[72] Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of neurophysiology*, 97(6), 4296–4309.

[73] Kindel, W. F., Christensen, E. D., & Zylberberg, J. (2019). Using deep learning to probe the neural code for images in primary visual cortex. *Journal of vision*, 19(4), 29–29.

[74] 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.

[75] Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, (pp.4).

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

[77] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012a). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

[78] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012b). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

[79] 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.

[80] Kurakin, A., Goodfellow, I., Bengio, S., et al. (2016). Adversarial examples in the physical world.

[81] Laskar, M. N. U., Giraldo, L. G. S., & Schwartz, O. (2020). Deep neural networks capture texture sensitivity in v2. *Journal of vision*, 20(7), 21–1.

[82] 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.

[83] Li, N., Cox, D. D., Zoccolan, D., & DiCarlo, J. J. (2009). What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? *Journal of neurophysiology*, 102(1), 360–376.

[84] Li, Z., Brendel, W., Walker, E., Cobos, E., Muhammad, T., Reimer, J., Bethge, M., Sinz, F., Pitkow, Z., & Tolias, A. (2019). Learning from brains how to regularize machines. *Advances in neural information processing systems*, 32.

[85] Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Doll'a r, P., & Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

[86] 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.

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

[88] 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*, 115(38), E9015–E9024.

[89] Lukanov, H., König, P., & Pipa, G. (2021). Biologically inspired deep learning model for efficient foveal-peripheral vision. *Frontiers in Computational Neuroscience*, 15, 746204.

[90] Macpherson, T., Churchland, A., Sejnowski, T., DiCarlo, J., Kamitani, Y., Takahashi, H., & Hikida, T. (2021). Natural and artificial intelligence: A brief introduction to the interplay between ai and neuroscience research. *Neural Networks*, 144, 603–613.

[91] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

[92] Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, 35(39), 13402–13418.

[93] Merel, J., Botvinick, M., & Wayne, G. (2019). Hierarchical motor control in mammals and machines. *Nature communications*, 10(1), 5489.

[94] Motter, B. C. (1994). Neural correlates of attentive selection for color or luminance in extrastriate area v4. *Journal of Neuroscience*, 14(4), 2178–2189.

[95] Mueller, K. N., Carter, M. C., Kansupada, J. A., & Ponce, C. R. (2023). Macaques recognize features in synthetic images derived from ventral stream neurons. *Proceedings of the National Academy of Sciences*, 120(10), e2213034120.

[96] Nicolae, M.-I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., et al. (2018). Adversarial robustness toolbox v1. 0.0. *arXiv preprint arXiv:1807.01069*.

[97] Orban, G. A., Van Essen, D., & Vanduffel, W. (2004). Comparative mapping of higher visual areas in monkeys and humans. *Trends in cognitive sciences*, 8(7), 315–324.

[98] Pasupathy, A. & Connor, C. E. (2001). Shape representation in area v4: position-specific tuning for boundary conformation. *Journal of neurophysiology*.

[99] Peterhans, E. & von der Heydt, R. (1989). Mechanisms of contour perception in monkey visual cortex. ii. contours bridging gaps. *Journal of Neuroscience*, 9(5), 1749–1763.

[100] Pfeiffer, M. & Pfeil, T. (2018). Deep learning with spiking neurons: Opportunities and challenges. *Frontiers in neuroscience*, 12, 774.

[101] Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, 177(4), 999–1009.

[102] Portilla, J. & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, 40, 49–70.

[103] 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.

[104] Rajalingham, R., Schmidt, K., & DiCarlo, J. J. (2015). Comparison of object recognition behavior in human and monkey. *Journal of Neuroscience*, 35(35), 12127–12136.

[105] Raman, R. & Hosoya, H. (2020). Convolutional neural networks explain tuning properties of anterior, but not middle, face-processing areas in macaque inferotemporal cortex. *Communications biology*, 3(1), 221.

[106] Rawat, W. & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29(9), 2352–2449.

[107] Rony, J., Hafemann, L. G., Oliveira, L. S., Ayed, I. B., Sabourin, R., & Granger, E. (2019). Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4322–4330).

[108] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115, 211–252.

[109] Safarani, S., Nix, A., Willeke, K., Cadena, S., Restivo, K., Denfield, G., Tolias, A., & Sinz, F. (2021). Towards robust vision by multi-task learning on monkey visual cortex. *Advances in Neural Information Processing Systems*, 34, 739–751.

[110] Sanghavi, S. & Kar, K. (2023). Distinct roles of putative excitatory and inhibitory neurons in the macaque inferior temporal cortex in core object recognition behavior. *bioRxiv*, (pp. 2023–08).

[111] Schein, S. J. & Desimone, R. (1990). Spectral properties of v4 neurons in the macaque. *Journal of Neuroscience*, 10(10), 3369–3389.

[112] Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*.

[113] Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*.

[114] 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.

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

[116] Storrs, K., Kietzmann, T., Walther, A., Mehrer, J., & Kriegeskorte, N. (2020). Diverse deep neural networks all predict human it well, after training and fitting. biorxiv. *Preprint*, 10(2020.05), 07–082743.

[117] Surianarayanan, C., Lawrence, J. J., Chelliah, P. R., Prakash, E., & Hewage, C. (2023). Convergence of artificial intelligence and neuroscience towards the diagnosis of neurological disorders—a scoping review. *Sensors*, 23(6), 3062.

[118] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

[119] Szegedy, C., Reed, S., Erhan, D., Anguelov, D., & Ioffe, S. (2014). Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441*.

[120] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

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

[122] Tacchetti, A., Isik, L., & Poggio, T. (2017). Invariant recognition drives neural representations of action sequences. *PLoS computational biology*, 13(12), e1005859.

[123] Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual review of neuroscience*, 19(1), 109–139.

[124] Ullman, S. (2019). Using neuroscience to develop artificial intelligence. *Science*, 363(6428), 692–693.

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

[126] Vinken, K. & Op de Beeck, H. (2021). Using deep neural networks to evaluate object vision tasks in rats. *PLoS computational biology*, 17(3), e1008714.

[127] Vogels, R. & Orban, G. A. (1996). Coding of stimulus invariances by inferior temporal neurons. *Progress in brain research*, 112, 195–211.

[128] Wichmann, F. A., Janssen, D. H., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., & Bethge, M. (2017). Methods and measurements to compare men against machines. *Electronic Imaging*, 2017(14), 36–45.

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

[130] Yamins, D. L., 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.

[131] Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

[132] Zhang, M., Tseng, C., & Kreiman, G. (2020a). Putting visual object recognition in context. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12985–12994).

[133] Zhang, X., Ma, Z., Zheng, H., Li, T., Chen, K., Wang, X., Liu, C., Xu, L., Wu, X., Lin, D., et al. (2020b). The combination of brain-computer interfaces and artificial intelligence: applications and challenges. *Annals of translational medicine*, 8(11).

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