Exploring the robust nature of human visual object recognition through comparisons with convolutional neural networks

By

Hojin Jang

Dissertation

Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in
Psychology

August 31, 2021

Nashville, Tennessee

Approved:

Frank Tong, Ph.D.

Randolph Blake, Ph.D.

Talia Konkle, Ph.D.

Thomas J. Palmeri, Ph.D.

Acknowledgments

I appreciate all my dissertation committee: Drs. Frank Tong, Randolph Blake, Talia Konkle, and Thomas Palmeri. Your insightful feedback has helped me strengthen my scientific skills throughout the process. Your encouragement has been valuable and kept me continuing to pursue science.

I would like to pay my special regards to my advisor, Dr. Frank Tong, for his consistent support, guidance, and patience in every step throughout the projects. He has always provided me with valuable guidance and professional advice with his kindness and enthusiasm. Every discussion with him has been inspiring and intellectually enriched my academic background.

I would like to acknowledge my colleagues and the Tong lab, without whom I would not have been able to complete the thesis projects in any way.

Finally, I would like to thank my family and friends. I could not have completed this dissertation without their support.

Table of Contents

1. General introduction

1.1 Computational models of visual object recognition

In everyday life, people recognize visual objects quickly and accurately, often without even realizing that they are doing it. It is remarkable that, despite the continuous flux of visual information, our visual system is able to maintain reliable recognition virtually all of the time. For decades, vision scientists have sought to discover the mechanisms underlying object recognition using a wide range of methodologies including psychophysics, neurophysiology, and neuroimaging. Among them, computational models offer a unique advantage in that a theory can be tested in explicit terms, providing a detailed algorithmic level of understanding and a causal account of how an object might be processed in the brain. Particularly because behavioral and neural data are often hard to analyze by themselves due to the complex nature of the brain, the modeling approach can be complementary by providing a conceivable explanation for them.

That said, developing a computational model of visual object recognition is a hard problem, because a model needs to satisfy both sensitivity and invariance in the representations of objects, which are often trapped in a trade-off relationship (Palmeri and Gauthier, 2004; Peissig and Tarr, 2007; Tsao and Livingstone, 2008; DiCarlo et al., 2012; Tong, 2018). To be specific, the recognition model needs to be sensitive enough to distinguish an object from another when they are visually similar (e.g., face recognition). At the same time, the recognition model needs to be robust enough to reliably identify a single object across variations in viewing angles, positions, or lighting conditions.

Early models of object recognition focused heavily on tackling the viewpoint invariance problem, namely, an object needs to be recognized as identical regardless of how it is viewed from multiple angles. Seminal theoretical work by Marr and Nishihara (1978) suggested a structural description model in which an object can be decomposed into volumetric primitive components, which are recognized in a hierarchically organized manner within an object-centered coordinate system. Because the primitive components are defined in 3D as volumetric representations, they are inherently viewpoint invariant. In addition, the model assumed a transformation from viewer-centered coordinates to object-centered ones, allowing an object to be recognized regardless of its viewpoint. Soon thereafter, Biederman (1985) proposed a recognition-by-components theory based upon similar ideas but his theory assumed a finite set of primitive component parts (i.e., so-called “geons”) inspired by a limited number of lexical elements in speech recognition such as phonetic alphabets. Biederman also provided empirical data to support the theory. These early computational models offered a conceptual basis for the understanding of object recognition. However, they were insufficient and often inadequate to explain real-world recognition behavior. For instance, they were mostly focused on the representational aspect of objects, while they did not provide much consideration for how objects are perceptually categorized (Palmeri and Gauthier, 2004).

More importantly, the notion that an object is recognized in a viewpoint invariant manner has been challenged by empirical findings. For example, Jolicoeur (1985) found that observers were slower to recognize a line drawing of an object if it was rotated away from a canonical upright orientation. However, this response time cost for misoriented objects diminished considerably after observers had the opportunity to view the object at other orientations for dozens of trials. Such findings contradict the idea that object recognition relies on viewpoint-independent representations, otherwise initial response times should have been equally fast across variations in orientation. The fact that practice mitigated the orientation effect further suggests that viewpoint-independent object recognition performance arises from perceptual learning of specific views of that object. In a later study, Tarr et al. (1998) tested whether or not line drawings of 10 geons were recognized equally well across 3 rotations (0°, 45°, 90°) in 3d space and similarly observed that the geon rotated further from the target needed more response time, indicating that object recognition performance varied in a viewpoint-dependent manner. In another study, Tarr and Pinker (1989) trained observers to name letter-like stimuli that were oriented in particular directions. Observers initially showed a linear increase in response times as a function of angular disparity from canonical orientation, but this response time function gradually flattened over time with practice. More interestingly, following extensive training, when the stimuli were presented with novel orientations, the response time increased proportionally with the difference from the nearest trained orientation. The authors conjectured that the initial orientation-specific representations were stored over training, obviating the need for mental rotation, and that those stored representations were used to identify the novel orientations. These results offer a potential solution to the problem faced by view-based models, that is, the notion that they might require storing an infinite number of object representations across variations in appearance. Instead, by storing a small number of canonical views of an object, novel views can be simply interpolated with respect to the previously stored representations (Bülthoff and Edelman, 1992). Accordingly, Poggio and Edelman (1990) proposed a viewpoint-dependent model in which a novel view of an object was recognized by reference to its previously trained viewpoints. To do so, they created a model by which positional coordinates as inputs were mapped to a latent space using radial basis functions whose centers and weights were optimized from given training samples. Those view-tuned radial basis functions were then used to predict a novel view. This model was able to capture human performance where an object was presented with a range of orientations away from the trained one.

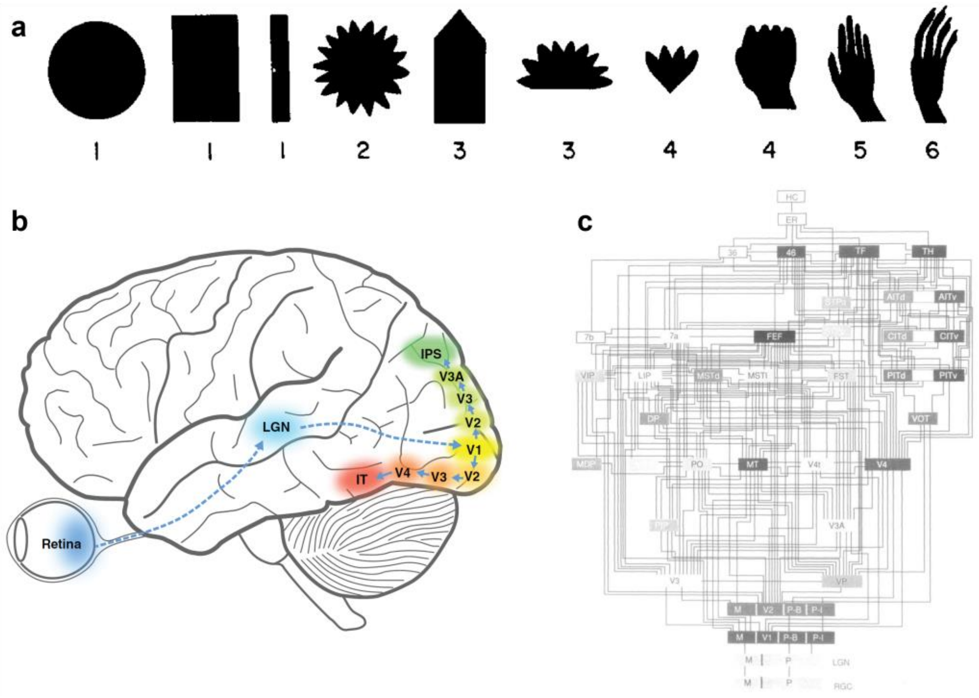

Figure 1. a Examples of stimuli used in Gross et al. (1972). The stimuli from left to right triggered stronger responses in IT neurons. **b** The ventral stream of the visual pathway, borrowed from Tong (2018). **c**

Hierarchical organization of the visual system on monkeys (Felleman and Van Essen, 1991). IT, inferotemporal.

Along with the behavioral and psychophysical methods that have provided insights into how visual objects may be perceived and recognized, the advent of non-invasive neuroimaging techniques has allowed researchers to examine the representations of objects at the neuronal population level. In particular, early neuroimaging studies have revealed evidence of modular representations of objects, showing how different categories of objects are mapped onto the brain. One of the most often studied and cited category-selective visual areas is the fusiform face area (FFA; Kanwisher et al., 1997), which strongly responds to faces over other objects. The response of the FFA has been shown to be consistent across human, animal, and even cartoon faces, and to be dependent on multiple viewpoints (Tong et al., 2000; Kietzmann et al., 2012). Subsequent neurophysiological research with macaque monkeys has revealed that 97% of the neurons in a face-selective area in the middle temporal lobe, separately defined by fMRI, exhibited strong face-selective responses (Tsao et al., 2006). Moreover, the pooled information conveyed by these neurons allowed for accurate discrimination of different human faces, implying that face-selective brain regions can support the identification of individual faces, going beyond simple face detection. This study provided a direct link between neurophysiology and neuroimaging to support the selective processing of visual objects. In addition to the FFA, other category-selective areas have been proposed, including the parahippocampal place area (PPA) which strongly responds places over other objects (Epstein & Kanwisher, 1998), the extrastriate body area (EBA) which responds strongly to body parts (Downing et al., 2001), and the lateral occipital complex (LOC) which responds more to intact objects than to scrambled ones (Grill-Spector et al., 2001). This body of research has contributed to a better understanding of how objects are processed in the brain and has allowed for a more detailed examination of the representation of specific categories of objects.

Further insights info the computational mechanisms underlying object recognition have emerged from neurophysiological studies. A landmark study by Gross et al. (1972) found that inferotemporal (IT) neurons in monkeys responded strongly to specific complex patterns such as hands or faces (**Figure 1a**). By contrast, these neurons showed poor responses to simple stimuli such as slits or rectangles. In conjunction with the earlier observation that V1 neurons had smaller receptive fields and greater sensitivity to simple visual stimuli such as oriented bars (Hubel and Wiesel, 1959; Hubel and Wiesel, 1962; Hubel and Wiesel, 1968), the authors conjectured that the IT cortex likely constituted a much later stage of hierarchical visual processing and may serve as a central locus for visual object recognition (Gross et al., 1972). Subsequent studies revealed that a subpopulation of neurons in the superior temporal sulcus selectively responded to faces over simple gratings or other objects (Perrett et al., 1982; Desimone et al., 1984), providing some support for the notion of modular organization of object representations. Furthermore, many of these neurons showed invariant responses to changes in the size or position of objects. However, the vast majority of these neurons appeared to be tuned in a viewpoint-selective manner, as they would respond maximally to a specific 3D view and their responses diminished as the object was rotated away in depth from that view (Perrett et al., 1987; Logothetis et al., 1995). These findings aligned with the notion of view-based models and with contemporary behavioral observations (Tarr and Pinker, 1989; Tarr et al., 1998).

Another important question concerns how complex objects are encoded in the IT cortex. One possibility is that an individual neuron represents a specific entity, such as a grandmother cell, a hypothetical neuron presumed to store the visual representation of one’s grandmother. An alternative view is that the representations of complex objects are encoded by populations of multiple neurons. A number of studies have supported the latter possibility (Pouget et al., 2000; Tsunoda et al., 2001; Hung et al., 2005). For instance, Hung et al. (2005) randomly sampled multiple IT neural sites and attempted to decode the object category and identity information using a regularization classifier. The authors found that categorization and identification performance increased as a function of the number of neural sites used for decoding, suggesting that the object information is distributed across multiple sites rather than specifically localized to single neurons. The authors further demonstrated that the same population of IT neurons was used for both categorization and identification and that decoding performance remained consistent across changes in the scale and position of the objects. This notion of population coding has also been supported by the studies of shape processing (Pasupathy and Connor, 2002; Brincat and Connor, 2004). The researchers measured neural responses in areas V4 and IT to parametric sets of 2D silhouette shapes and showed that the shape tuning function of a stimulus was successfully predicted by the population representation of neurons. These findings provide further support for the hierarchical organization of the visual system, with higher cortical visual areas serving to encode more complex features and object properties while exhibiting greater generalization capacity across changes in position and size. At the same time, such research has suggested that objects are processed in a parallel and are represented in a somewhat distributed manner.

This body of literature has offered important insights into how objects are processed in the brain, particularly underlining the fact that the visual system is composed of multiple areas connected in a hierarchical manner and that objects are processed along the ventral stream, which projects from the primary visual cortex (V1) to the temporal cortex (**Figures 1b-c**; Felleman and Van Essen, 1991; Desimone and Duncan, 1995). That is, the low-level features of objects are processed earlier in the hierarchy and those features are combined to process more complex representations of objects and their semantic information at later stages of visual processing. In later studies, this strategy has been suggested to be efficient in reducing the complexities of real-scene objects by disentangling or flattening object manifolds (DiCarlo and Cox, 2007; DiCarlo et al., 2012). An early model of invariant object processing was proposed by Wallis and Rolls (1997). This model consisted of a five-layer hierarchical structure, each layer optimized for achieving translation invariant representations, with the exception of the first layer which had a fixed set of parameters based on difference-of-Gaussians.

1.2 Robust object recognition under degraded viewing conditions

The main stream of research in object recognition has focused on how the visual system achieves invariance to changes in position, size, and viewpoint. However, another important problem in object recognition that has received comparatively less attention concerns how humans recognize objects under degraded viewing conditions, as can result from occlusion, clutter, blur, or noise. In an early study by Biederman (1985), when an object was decomposed into its components, the first few components were sufficient for human observers to achieve almost 90% of accuracy performance. Even when line drawings of objects were degraded by partial deletion of their contours up to 65%, human observers were still good at identifying objects. This suggests that the human object recognition system is fairly stable when only limited information is available. In addition, performance was affected by which parts of contours were deleted and how much time observers had to view the stimuli, suggesting that multiple factors may play a role in robust processing. The question of how we can achieve robust recognition performance with varying degrees of limited object features has not been fully answered until now.

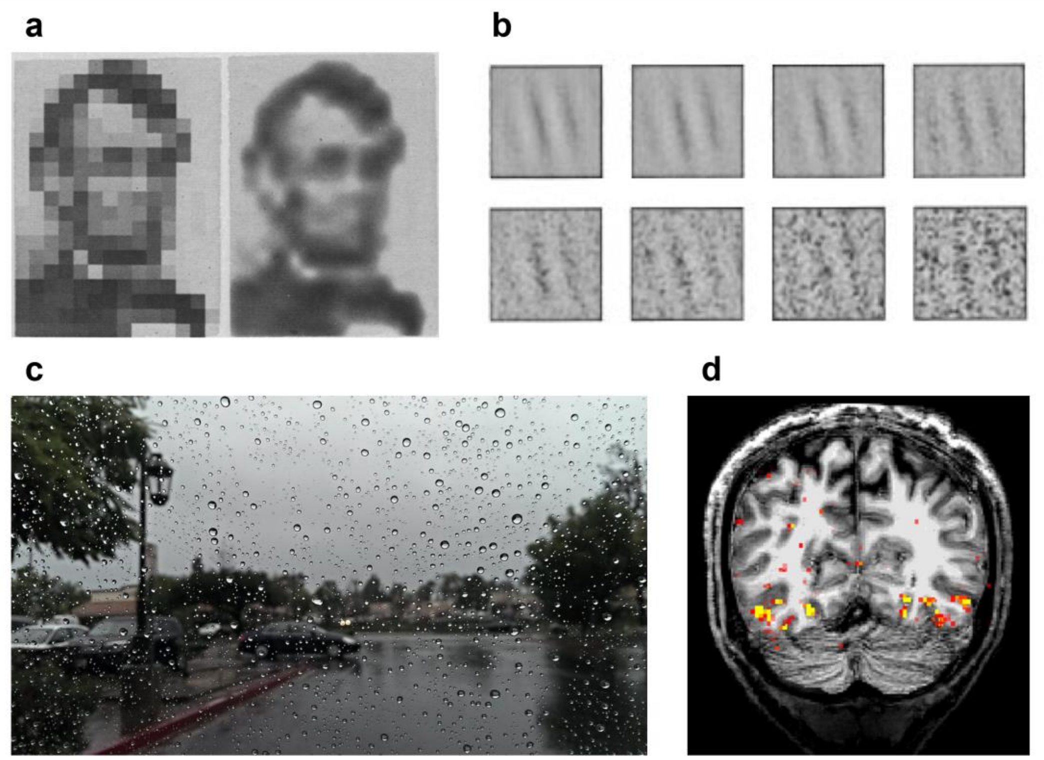

Figure 2. a Face stimuli used in Harmon and Julesz (1973). A face image was degraded by quantization (left) and low-pass filtering (right). **b** Orientation stimuli used in Dosher and Lu (1998) in the presence of random Gaussian noise. **c** Example of a rainy image through window. "Southern California Downpour" by NancyFry is licensed with CC BY 2.0. https://creativecommons.org/licenses/by/2.0/. **d** Anatomical (black and white) and functional (yellow and red) MRI images acquired by the author.

Studying object recognition under degraded viewing conditions is also useful because it could offer insights into different underlying mechanisms of visual processing, which can be revealed by different types of degradation. For example, Gosselin and Schyns (2001) devised a technique, so-called Bubbles, that uses multiple Gaussian windows placed at random locations on a face image to selectively reveal certain parts of the image. Given individual maps of Bubbles, observers' performance on resolving face gender, expression, and identity was measured. Based on their correct/wrong responses to the individual maps, it allowed a way to reveal task-specific diagnostic features for face processing. In a similar vein, blur has long been used in vision science with its unique characteristic of spectral selectivity (**Figure 2a**; Harmon and Julesz, 1973; Morrone et al., 1983), where the high spatial frequency components of images representing fine features are effectively removed. As such, blur has been applied to many aspects of vision research to reveal different aspects of recognition processes such as holistic processing in face perception (Farah et al., 1998; Goffaux and Rossion, 2006) or contextual scene understanding (Torralba, 2003; Oliva and Torralba, 2007). In some studies, multiple frequency bands were manipulated to create visual noise and various aspects of recognition were examined using the noise (Solomon and Pelli, 1994; Wichmann et al., 2006). On the other hand, random white noise containing equal intensity at different frequencies has been also extensively used, for example, by Dosher and Lu for studying the effect of perceptual learning on orientation discrimination tasks within the context of external and internal noise (**Figure 2b**; Dosher and Lu, 1998; Dosher and Lu, 2005; Dosher et al., 2013). In sum, understanding how we deal with various types of image degradation can help us better elucidate the robust nature of our visual object recognition system.

Attaining a better understanding of the mechanisms underlying the robust nature of human object recognition is of considerable importance, especially nowadays when a variety of computational recognition models are widely used in many applications. These computational models often have to resolve recognition problems under challenging viewing conditions due to environmental factors such as autonomous driving on rainy days (**Figure 2c**) or deep-sea exploration by robots/submarines. In other situations, the models may need to cope with the low-quality image data that is inherently noisy or images that contain artifacts. For example, medical images such as MRI or fMRI contain many artifacts possibly due to motion, magnetic field inhomogeneity, and so forth, which could adversely affect diagnostic decisions (**Figure 2d**).

1.3 Emergence of convolutional neural networks (CNNs)

A major branch of hierarchical recognition models has been primarily motivated by an early influential study by Hubel and Wiesel (1962). In the study, two types of cells in V1 were proposed, so-called simple and complex cells. Although both types of neurons showed orientation-specific preferences for bars of light, simple cells showed distinct excitatory and inhibitory zones within their receptive fields, whereas complex cells did not show such a division and responded quite well across positions within the receptive field. These findings suggested that complex cells achieved position invariance to some degree and were likely higher-order cells receiving afferents from simple cells. Expanding upon these ideas, Fukushima (1980) proposed Neocognitron, a neural network model that consisted of alternating layers of “S-cells” and “C-cells”, resembling the simple and complex cells in V1. Given an input pattern, S-cells extracted local features within their receptive fields, and C-cells received signals from a group of S-cells that shared the same feature but had slightly different receptive field locations and yielded consistent responses as long as at least one of the S-cells was activated. This hierarchical sequence of S-cells and C-cells was repeatedly stacked to form multiple layers and the network was trained to produce the latent representation of input patterns in an unsupervised learning manner. Through the hierarchy, the early set of cells (S1 and C1) responded to simple features, while the later cells (S3 and C3) responded to more complex patterns. Thereby, this architecture enabled the model to achieve more robust pattern recognition performance tolerant to slight positional shifts. The basis of the network later greatly inspired the initial conception of convolutional neural networks (LeCun et al., 1989; LeCun et al., 1990), as will be described below. This idea has continued to influence the later computational models for object recognition (Riesenhuber and Poggio, 1999; Zhu and Mumford, 2007).

In parallel with advances in perception and neuroscience, continuous efforts have been made in artificial intelligence to develop computational algorithms to emulate human cognitive abilities. In the 1950s, the “perceptron” was introduced as the first connectionist model that could learn the weights from inputs to elicit desired outputs (Rosenblatt, 1958). The weights of the model were adaptively updated based on the difference between the desired and actual outputs, similar to the delta learning rule (Widrow and Hoff, 1960), but the actual output response was thresholded in an all-or-none fashion. While the perceptron was successful in simple cognitive settings, it was challenged by the fact that a simple perceptron is not able to solve more complex or non-linear problems such as the exclusive-OR (XOR) problem. This led to a gradual decline in interest in neural networks over the following decades as artificial intelligence (AI) researchers shifted their focus to other approaches such as rule-based expert systems (Liao, 2005). After this so-called “AI winter” period, it was demonstrated that the limitations of the early perceptron model could be overcome by stacking up perceptrons into layers (i.e., multi-layer perceptrons) and training multi-layer networks via a back-propagation algorithm (Rumelhart et al., 1985;

Rumelhart et al., 1986). Soon after, LeCun and his colleagues demonstrated that the backpropagation algorithm could be successfully applied to a neural network model that leveraged a convolution operation to perform a handwritten digit recognition task, sometimes called MNIST (LeCun et al., 1989; LeCun et al., 1990). This study was subsequently extended to construct a more advanced neural network (called LeNet-5), which introduced the basic components of modern convolutional neural network architectures (LeCun et al., 1998). However, the training and implementation of these models still posed major technical challenges, including the vanishing gradient problem, overfitting, and slow training speed.

Over the past decade, these technical issues have been successfully resolved by advances in software and hardware technologies. For example, introducing rectified linear units as activation functions has turned out to be effective in mitigating the gradient vanishing problem (Glorot et al., 2011). Moreover, the powerful parallel computations performed by graphic processing units (GPUs) have enabled a huge speed boost in training these deep network models. With these technical advances, AlexNet, a deep CNN model was successfully trained on a massive natural image dataset (Russakovsky et al., 2015), and ultimately won the 2012 ImageNet challenge, defeating other computer vision algorithms that did not rely on deep neural networks by a large margin (Krizhevsky et al., 2012).

Since then, a series of newer CNN models have been proposed with major gains in recognition performance on a yearly basis (Chatfield et al., 2014; Simonyan and Zisserman, 2014; Szegedy et al., 2015; He et al., 2016). In 2015, CNNs have been suggested to even exceed human-level performance on the ImageNet classification task (He et al., 2015). The enormous success of CNNs has motivated many researchers to apply them to other vision tasks, such as object detection (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015; Redmon et al., 2016), object localization (Long et al., 2015; Noh et al., 2015; Ronneberger et al., 2015), and medical image perception (Shen et al., 2017; Litjens et al., 2017; Esteva et al., 2017).

1.4 CNNs as a model of biological visual systems

With the outstanding performance of CNNs in vision tasks, many neuroscientists have questioned whether CNNs process objects in a similar manner as do biological visual systems. Research by Yamins and DiCarlo examined how well V4 and IT neural responses could be predicted by a 3-layer CNN model trained on a categorization task, each layer of which consisted of convolutional filters, thresholding, local pooling, and normalization (Yamins et al., 2014; Yamins and DiCarlo, 2016). Their hierarchical CNN model not only outperformed shallow control models with respect to the explained variance of neural responses but also exhibited an interesting correspondence between individual layers of the CNN and neural sites. Specifically, the top layer of the CNN best predicted IT neural responses, whereas the intermediate layer best predicted V4 responses. Another neurophysiological study demonstrated that spiking activity in V1 of awake monkeys was also better predicted by CNNs than classical linear-nonlinear models or Gabor wavelet models (Cadena et al., 2019). A neuroimaging study used representation similarity analysis (RSA) to show that CNNs trained with millions of labeled images explained IT representations better than traditional computer vision models, with higher layers exhibiting higher correlations (Khaligh-Razavi and Kriegeskorte, 2014). A subsequent study nicely demonstrated that early visual areas and the lower layers of CNNs preferred low-level features, including edges and local contrast, while higher visual areas and higher layers of CNNs were more sensitive to high-level features such as object parts or entire objects (Güçlü and Gerven, 2015). These studies collectively suggest that CNNs resemble biological vision in processing visual objects across multiple stages of hierarchical processing.

Moreover, it has been suggested that CNNs appear to capture some of the key features of object recognition that are observed in biological systems. For example, the first-layer receptive field structures of the CNN appear to resemble those of V1 simple cells, exhibiting a broad range of orientation-selective preferences (Krizhevsky et al., 2012). Another study has demonstrated that even though CNNs were trained to be optimized for semantic categorization tasks, CNNs maintained above-chance performance when only silhouette images were displayed with color and texture cues removed, indicating that they had acquired a certain degree of shape sensitivity (Kubilius et al., 2016). Moreover, CNNs exhibit selectivity for boundary curvature in a manner that resembles mid-level stages of visual processing, such as curvature tuning as has been found in area V4 of the macaque monkey (Pospisil et al., 2018). A recent study used a reverse engineering technique to visualize the characteristics of V4 neurons, by showing monkeys synthetic images that were generated to maximally elicit the firing rates of V4 neural sites via a gradient descent algorithm, and indeed observed complex curvature patterns (Bashivan et al., 2019). Taken together, there is a sizeable and growing body of research to suggest that CNNs provide a promising and highly effective model for characterizing the mechanisms of object recognition in humans.

1.5 Discrepancies between CNNs and biological vision: Poor robustness of CNNs when faced with challenging viewing conditions

Although CNNs provide the best current computational model of biological vision thus far, they are still far from perfect. More important, recent studies have begun to reveal significant differences between CNNs and primate visual systems in various aspects. For instance, several studies have reported that CNNs can account for a sizeable percentage but not all of the variance of neural response data with respect to object recognition. In Bashivan et al.'s study (2019), the CNN model accounted for most of the explained variance for natural images (~89%) but was only able to predict 54% of the brain response to the synthetic images. A recent neuroimaging study has carefully examined a brain-CNN correspondence using an RSA approach compared with the noise ceiling of human-to-human performance (Xu and Vaziri-Pashkam, 2021), finding that the low-level representations of real-world objects from early visual areas were able to be fully accounted for by some CNNs but none of them reached the lower bound of the noise ceiling with respect to high-level visual areas. Another study has reported that although CNNs successfully accounted for human recognition performance at the object category level, they failed to predict individual image level behavioral patterns such as individual image difficulty and image-level confusions (Rajalingham et al., 2018), whereas monkeys were highly consistent with humans at both object and image levels.

Also, differences between CNNs and biological vision become more pronounced under certain experimental settings. In a visual search task, human observers tended to miss a target object when it appeared atypical in its size relative to the rest of the scene, while CNNs did show consistent performance regardless of the size of target objects (Eckstein et al., 2017). Another study reported that CNNs fail at simple visual reasoning problems such as the "same-different" task (Ricci et al., 2018), in which the models needed to determine whether two items appearing at different positions were identical or not. Also, despite the early suggestion that CNNs exhibited some degree of shape sensitivity in recognition (Kubilius et al., 2016), recent studies have shown that CNNs tend to strongly rely on texture rather than shape cues to classify objects (Geirhos et al., 2019; Baker et al., 2019). To be specific, Geirhos et al. (2019) created an artificial stimulus that contained the shape of an object covered with the texture pattern of another object (e.g., cat shape with elephant texture) and measured the recognition bias of human observers and CNNs. Strikingly, human observers demonstrated a strong bias to favor shape cues, whereas CNNs were largely biased by texture cues.

More crucially, CNNs are highly vulnerable to challenging viewing conditions. Several studies have shown that CNNs exhibit substantially degraded performance by a small amount of visual distortion (Vasiljevic et al., 2017; Dodge and Karam, 2017; Geirhos et al., 2018; Jang and Tong, 2018). For instance, Dodge and Karam (2017) compared the performance of humans and CNNs on a dog breed classification task in the presence of additive Gaussian noise and Gaussian blur. The authors found that the degraded but still fairly recognizable images to humans were severely detrimental to CNNs. Geirhos et al. (2018) evaluated the impact of various types of visual noise in humans and CNNs and similarly observed that humans outperformed CNNs in every condition. More strikingly, the authors claimed that directly training CNNs on one type of noise barely generalized to the other types, suggesting the poor generalization ability of CNNs. Concurrently, we also found that humans greatly outperform CNNs at recognition objects in visual noise, though they could be made more robust by training them to recognize objects in noise, as will be discussed in chapter 2 (Jang & Tong, 2018). These findings reveal that CNNs, especially those that are trained using standard image data sets, do not provide robust recognition under challenging viewing conditions, which makes them unsuitable for real-life applications that include dynamic and diverse environments. The poor robustness of CNNs is further accentuated by adversarial noise that is imperceptible to humans but causes devastating accuracy losses in deep learning models (Goodfellow et al., 2014; Szegedy et al., 2014). For example, Goodfellow et al. (2014) demonstrated that by adding an imperceptibly small amount of noise to a panda image that was driven by a gradient ascent algorithm, a CNN model was confused to classify it as a gibbon with 99.3% confidence. Brown et al. (2017) showed that a print-out of an image into which a small adversarial patch was embedded could even attack real-world applications of CNN classifiers. This series of observations suggests that CNNs are fragile and do not process objects in the same way that humans do.

1.6 Overall goal of this thesis

It is striking that CNNs can be fooled by fairly modest or subtle image manipulations. Such vulnerability is by no means acceptable for real-world vision problems. By contrast, the recognition ability of humans is highly robust and reliable across a variety of viewing conditions. The main goal of this thesis was to explore the robust nature of the human object recognition system under degraded viewing conditions by comparing human behavioral and neural data to those obtained from CNNs under the same conditions. This comparative approach could prove very useful for identifying the distinct features of the human visual system that mediate robust object recognition, when contrasted with the fragility of current CNNs. Furthermore, the discrepancies could provide insights into the design of future CNN models if the goal is to make CNNs more human-like.

Another goal of this thesis was to determine whether experiencing degraded viewing conditions may be beneficial or even necessary for achieving an acceptable level of robustness. Visual scenes in daily life are not always presented in optimal conditions; regardless, we need to solve these everyday vision tasks. We sought to examine whether suboptimal viewing conditions, even though they might occur quite rarely, might actually lead the visual system to be more robust. To this end, we leveraged CNNs to evaluate whether degraded visual conditions would improve their robustness and, furthermore, allow them to achieve a better match with human behavioral and neural performance.

1.7 Overview of projects

In Chapter 2, we examined how CNNs and human observers recognize objects in highly noisy viewing conditions. Their performance was compared across signal-to-noise ratios for objects shown in two types of visual noise, pixelated Gaussian noise and Fourier phase-scrambled noise. We found that CNNs demonstrated overall poorer robustness to noise than human observers, but more interestingly, that CNNs were more susceptible to pixelated Gaussian noise, whereas human observers showed worse performance when objects were presented with Fourier phase-scrambled noise. These findings provided initial evidence that CNNs process noisy objects in a different manner than humans do. We further examined whether training CNNs with noisy objects might provide a suitable way to mimic the robust recognition behavior of humans under noisy viewing conditions. We found that the noise-trained CNNs showed better correspondence to human observers than the control CNNs trained without noise, including better predictions of individual image difficulty levels, greater similarity in visual saliency maps, and better alignment of visual representations for noisy objects. This study suggests that standard CNNs do not have an adequate mechanism to deal with visual noise but training with noisy examples can allow CNNs to better mimic the human recognition system under noisy conditions.

We further explored perceptual learning with objects in noise in Chapter 3. We first asked whether human observers could further improve in their robustness via training even though they already seemed to be highly robust to noise. We observed that after extensive training with noisy objects, both human observers and CNNs showed significant improvements in robustness. Next, we investigated whether the effect of noise training would generalize to untrained categories by training humans and CNNs on either animate or inanimate objects in visual noise. Interestingly, human observers only showed improvement when tested on trained categories (i.e., category-specific effect). By comparison, CNNs showed a significant degree of improvement when tested on untrained categories but showed even greater benefit for trained categories (i.e., both category-specific and category-general effects). We further found that the category-general effect primarily involved changes to the early and middle layers of the CNNs, whereas the category-specific effect was more evident in the middle and higher layers. These findings provide an interesting implication that the robust nature of object recognition may involve multiple stages of a hierarchical visual system.

The study in Chapter 4 was inspired by the developmental literature that infant vision is initially very poor and coarse but gradually improves over the first year of life. We wondered if this developmental sequence of blurry to clear visual inputs would confer some ecological benefit, especially with respect to robust object recognition. We compared two versions of CNNs, one that was trained using clear images only and the other initially trained with blurry images followed by progressively clearer ones. Those two versions of CNNs were trained on either a face recognition task or an object recognition task. We observed a critical difference between the face- and object-trained networks: The face-trained CNNs successfully gained benefits from the sequence of blurry to clear training by achieving better robust performance to variations in spatial blur, whereas the object-trained CNNs did not show any noticeable improvement. A unit-level spatial frequency preference analysis revealed that the initial low spatial frequency components of faces were sufficient to achieve nearly perfect performance on face recognition. By contrast, the object-trained CNNs continued to require higher spatial frequency information for optimal performance on object recognition. This study accentuates the central role of low spatial frequencies in face processing, providing novel computational evidence that faces are processed in a more holistic manner than objects.

Lastly, the fact that CNNs process visual inputs differently from how we perceive the visual world motivated us to explore whether this may explain some of the previously reported discrepancies between CNNs and human vision. In particular, it is well known that a large proportion of the visual field is perceived as blurry and that the human visual system might have to deal with blurry objects in real-world situations, for example, searching for a target object that is out-of-focus and appears blurry. To illuminate this in Chapter 5, we evaluated whether CNNs trained with a mixture of clear and blurry images might provide a better model of human recognition under various viewing conditions. We first found that the CNNs trained on both clear and blurry images better predicted the behavioral and neural patterns of human observers than the control CNNs trained on clear images only. The effect of blur training appeared to be mainly on the earlier layers of the CNNs. More intriguingly, the CNNs trained on both clear and blurry images showed better correspondence to humans in non-blurry conditions, including greater shape bias, greater robustness to Gaussian noise, and more similar visual representations to those of humans across both noisy and clear viewing conditions. These observations suggest the possibility that modern CNN models typically trained on a predominant percentage of clear images may be biased to learn fine-scale representations of object features and thereby deviate from the biological visual system, which presumably relies on a wider range of spatial frequencies to attain robust object recognition.

2. Comparison of humans and convolutional neural networks in noise robustness

2.1 Introduction

Imagine driving in a downpour or searching for a friend’s face in a crowd. Because our visual world is intrinsically noisy and cluttered, sophisticated mechanisms are essential to enable us to accomplish stable and robust visual recognition. The robust nature of human vision is rather the consequence of multiple mechanisms, not one single process. For example, the simple principle of averaging can be useful in such a manner that the impact of random noise from independent sources will be mitigated by averaging (Faisal et al., 2008). Selective visual attention is another critical cognitive process that has been considered to enhance the signal-to-noise ratio via gain modulation and sharpening (e.g., Treue and Trujillo, 1999; Carrasco et al., 2004; Kamitani and Tong; 2005; Reynolds and Heeger, 2009).

The robust characteristics of human object recognition have been accentuated by the vulnerability of CNNs to image degradation. Although growing evidence has demonstrated that CNNs are currently the best computational model for nonhuman primate and human visual systems (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and van Gerven, 2015; Yamins and DiCarlo, 2016; Cadena et al., 2019), however, at the same time, recent studies have shown a critical limitation of CNNs that they are unacceptably poor at recognizing objects under challenging viewing conditions. Dodge and Karam (2017) evaluated the recognition performance of both humans and CNNs using stimuli from 10 different dog breeds degraded by Gaussian noise and blur. While the recognition performance of CNNs on clean stimuli was comparable to that of humans, their performance deviated progressively as more noise was added. Geirhos et al. (2018) reported a similar finding tested with more various types of visual noise. In early versions of my work, we have also observed that CNNs showed poor recognition performance with noise but also showed the opposite pattern of performance to that of humans while responding to different types of visual noise (Jang and Tong, 2018). More recently, a public benchmark providing nineteen different types of image corruptions has been proposed (Hendrycks and Dietterich, 2019), which has attracted further attention to this research problem.

This observation naturally leads to the following question: how can we improve the robustness of CNNs so they better generalize to noisy conditions? A simple but effective method that has also long been studied in the machine learning literature is directly adding noise into inputs. Injecting addictive noise into inputs has been formulated as a regularization term (Bishop, 1995). An (1996) has demonstrated that this regularization term constrained networks to be less sensitive to input variations, enabling better generalization. More recently, Vincent et al. (2010) proposed adding noise into the input of the autoencoder to construct a so-called denoising autoencoder, which is known to better capture robust implicit representations. Nowadays, adding noise into inputs has been widely used for the purpose of data augmentation, noise robustness, and overfitting control (Audhkhasi et al., 2016; Zheng et al., 2016; Geirhos et al., 2018; Jang and Tong, 2018; Rusak et al., 2020; Tong and Jang, 2021).

Despite the effectiveness of the method, there is a fundamental question that remains to be answered, whether the CNN trained with noise would process noisy objects in a similar manner as do humans. No single study has yet examined this question perhaps because that is merely presumed based on the improved accuracy performance. However, it should be noted that the increased accuracy does not indicate that they will necessarily use the same strategy to deal with noise as humans. It has been suggested that CNNs often adopt idiosyncratic strategies when solving vision problems (Geirhos et al., 2019; reviewed by Geirhos et al., 2020). Many previous studies reporting high human-CNN correspondence in their internal representations of objects were primarily based on clear and non-degraded images (Yamins et al., 2014; Yamins and DiCarlo, 2016; Güçlü and Gerven, 2015). Therefore, the systematic comparison of noise-trained CNNs and humans under noisy viewing conditions is still completely missing. In this study, we sought to carefully examine the recognition pattern of humans, standard CNNs, and noise-trained CNNs to gain a comprehensive understanding of the robust nature of object recognition systems.

We first determined two types of visual noise. The first noise was pixelated Gaussian noise which was randomly generated from a normal distribution. The second noise type was Fourier phase-scrambled noise by which the original power spectrum was maintained, while geometric features such as edge or contours were distorted by scrambling the phase contents (Wichmann et al., 2006). The advantage of using the two types of noise is that pixelated Gaussian noise represents unstructured and spatially independent noise, whereas Fourier phase-scrambled noise represents structured and spatially correlated noise.

Using the two types of noise, we measured both human and CNN recognition performance by parametrically manipulating the signal-to-noise ratio. As expected, CNNs showed worse performance than humans in both types of noise. More interestingly, we found the opposite pattern between them: CNNs were easily impaired by pixelated Gaussian noise, but humans were poorer at recognizing objects with Fourier phase-scrambled noise. This finding suggests that CNNs may process noisy objects in a qualitatively different manner than humans.

We next evaluated whether the CNNs trained with noisy examples might be better at predicting the response patterns of humans in noisy conditions. We expected higher recognition performance for the noise-trained CNNs when tested on noisy examples; however, again, it was not necessarily guaranteed if the noise-trained CNNs would internally process noisy objects similar to humans. We observed that, besides their improved robustness, the noise-trained CNNs yielded better predictions of human recognition thresholds on an image-by-image basis and more human-like diagnostic features under noisy conditions. A layer-specific analysis indicated that the effect of noise training emerged in the middle layers and was amplified as it ascended to higher layers. Moreover, the noised-trained CNNs showed a closer correspondence in their representations of noisy objects to human fMRI data in both early and higher visual areas. Finally, we showed that training with a certain type of noise could generalize to some of the untrained noise types, suggesting the possibility of designing a universal model robust to multiple noise types.

2.2 Material and methods

Participants

We recruited 23 participants in behavioral experiment 1 (18 females, 5 males), with 20 participants successfully completing both sessions of the study. A separate group of 23 participants were recruited in behavioral experiment 2 (14 females, 9 males), with 20 participants completing all 4 sessions of the study. Ages ranged from 19 to 33 years old.
An fMRI experiment was also carried out with a total of 11 participants (5 females), ages 21-49; data from 3 participants were excluded due to poor MR data quality. All participants reported having normal or corrected-to-normal visual acuity, and provided informed written consent using electronic consent forms (REDCap). The study was approved by the Institutional Review Board

of Vanderbilt University (IRB #040945). Participants were compensated monetarily or through a combination of course credit and monetary payment.

Visual stimuli

Object images were obtained from the ImageNet database (Russakovsky et al., 2015), which is commonly used to train and test convolutional neural networks on object classification. We selected images from 16 categories for our experiments, which included a mixture of animate and inanimate object categories that would be recognizable to participants (**Figure 3b**). The 16 categories were explained to participants before experiments. Both humans and CNNs were tested using images from the validation data set of ImageNet, with 50 images per category or 800 images in total. The test images were converted to grayscale to remove color cues that otherwise might boost the ability to recognize certain object categories in severe noise. CNNs were trained using images from the training set (1300 images per category), so the images used for testing were novel to both humans and CNNs.

In Experiment 1, objects were presented using two different types of visual noise: pixelated Gaussian noise and Fourier phase-scrambled noise (**Figure 3a**). To create each Gaussian noise image, the intensity of every pixel was randomly and independently drawn from a Gaussian distribution centered at 127.5, assuming that the range of possible pixel intensities (0 to 255) spanned ±3 standard deviations. For Fourier phase-scrambled noise, we calculated the average amplitude spectrum of the 800 images, generated a set of randomized phase values and performed the inverse Fourier transform to create each noise image. Such spatially correlated noise has some coherent structure that preserves the original power spectrum (close to a 1/F amplitude spectrum) but lacks strong co-aligned edges, due to the phase randomization, and can be described as having a cloud-like appearance. We avoided using the Fourier power spectrum of individual images to generate noise patterns, as residual category information could persist in this case, assuming that the categories differ to some extent in their overall power spectra.

To investigate the effect of noise on object visibility, we manipulated the proportion of object signal (w) contained in the object-plus-noise images. We describe the proportional weighting of this object information as the signal-to-signal-plus-noise ratio (SSNR), which has a lower bound of 0 when no object information is present (i.e., noise only) and an upper bound of 1 when the image consists of the source object only. SSNR differs from the more conventional measure of signal-to-noise ratio (SNR), which has no upper bound. Given a source object image defined by matrix $\boldsymbol{S}$ and a noise image $\boldsymbol{N}$, we can create a target image $\boldsymbol{T}$ with SSNR level of w as follows:

$$\boldsymbol{T} = w \cdot \boldsymbol{S} + (1 - w) \cdot \boldsymbol{N}$$

After the contrast-adjusted original image and the noise pattern were summed, any intensity values that fell beyond the 0-255 range were clipped. Clipping was modest as the standard deviation of the Gaussian noise distribution was 255/6.

Behavioral experiment 1

Participants were tested with either pixelated Gaussian noise or Fourier phase-scrambled noise, in two separate behavioral sessions. To control for order effects, half of the participants were presented with pixelated Gaussian noise in the first session and while the other half were first presented with Fourier phase-scrambled noise.

In each session, participants were briefly presented with each of 800 object images for 200ms at a specified SSNR level, and had to make a 16-alternative categorization response thereafter using a keyboard. Noisy object images were presented at 10 possible SSNR levels (0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.5, and 0.75). The highest SSNR level was informed by a pilot study that indicated that human accuracy reached ceiling levels of performance by an SSNR level of 0.75. Five images per category were assigned to each SSNR level, and image assignment across SSNR levels was counterbalanced across participants. The order of image presentation was randomized. The experiment was implemented using MATLAB and the Psychophysics Toolbox (http://psychtoolbox.org/).

Behavioral experiment 2

This study measured participants' SSNR thresholds for each of 800 object images over a series of 4 behavioral sessions. For this experiment, only pixelated Gaussian noise was evaluated. On each trial, a single noise image was generated and combined with a source object image, and the target image gradually increased in SSNR level by 0.025 every 400ms, until the participant felt confident enough to press a key on a number pad to halt the image sequence and then make a 16-alternative categorization response. Next, participants used a mouse pointer to "paint" the portions of the image that they found to be most informative for their recognition response.

After each trial, participants received visual feedback, based on a point scheme designed to encourage both fast and accurate responses. For correct responses, up to 200 points could be earned at the beginning of the image sequence (SSNR = 0), and this amount decreased with increasing SSNR, dropping to just 6 points at an SSNR level of 1. Incorrect responses were assigned 0 points. The participants received monetary payment scaled according to the total number of points earned across the 4 sessions.

MRI scanning parameters

MRI data were collected using a 7-Tesla Philips Achieva scanner with a 32-channel head coil at the Vanderbilt University Institute for Imaging Science. We collected fMRI data using single-shot T2*-weighted gradient echo echo-planar imaging at a 2mm isotropic voxel resolution (TR 2s; TE 25ms; flip angle 63°; SENSE acceleration factor 2.9, FOV 224×224 mm; 46 slices with no gap; phase-encoding in AP direction). To mitigate image distortions caused by inhomogeneity, an image-based shimming technique was used. A T1-weighted 3D-MPRAGE anatomical scan was collected in the same session at 1mm isotropic resolution. Separately, retinotopic data were acquired using a 3-Telsa Philips Intera Achieva MRI scanner equipped with a 32-channel head coil, with fMRI data acquired at 3mm isotropic resolution (TR 2s; TE 35ms; flip angle 80°; FOV 240×240 mm; 36 slices).

fMRI experiment

For the fMRI experiment, we selected 16 object images to characterize neural response patterns to objects in visual noise. The images included 2 examples drawn from 8 categories (bear, bison, elephant, hare, jeep, sports car, table lamp, teapot) whose difficulty levels were closely matched based on the reported SSNR levels from Experiment 2. Each object image was presented noise-free, embedded in pixelated Gaussian noise, and embedded in Fourier phase-scrambled noise. For the noise conditions, we chose an SSNR level of 0.4 as human performance dropped significantly by this noise level but was still accurate enough to be expected to lead to reliable neural responses. To control for potential order effects, the images were divided into two sets. In the first half of the experiment, one set was presented noise-free while the other set of object images appeared in each of the two types of noise. In the second half of the experiment, the assignment to noisy and noise-free conditions was reversed. Across participants, we counterbalanced how the objects were assigned to noisy and noise-free conditions across the two halves of the experiment. Each fMRI run consisted of 8 clean images, 8 Gaussian noise images, and 8 images of objects in Fourier phase-scrambled noise, presented in randomized order. On average, participants performed a total of 10 experimental runs with each image shown 5 times for a given condition.

Participants were instructed to maintain fixation on a central fixation point throughout each experimental run and to report whether each presented image was animate or inanimate using an MRI-compatible button box in the scanner. Each image from a stimulus set was centrally presented in a 9 × 9° window for 4 seconds, flashing on and off every 250ms, and followed by a 6-second fixation rest period. The order of the 24 images was randomized every run, and each run lasted approximately 4.4 minutes. We additionally ran 2 runs of a functional localizer, in which participants viewed blocked presentations of grayscale images of faces, objects, houses, and scrambled objects. A subset of the participants were scanned on a separate day for retinotopic mapping which used a standard phase-encoded measurement with rotating wedges and expanding rings (Engel et al., 1997).

fMRI data preprocessing and analysis

Data were preprocessed and analyzed using FSL, Freesurfer, and custom MATLAB scripts. The following standard preprocessing was applied: motion correction using MCFLIRT (Jenkinson et al., 2002), slice-time correction, and high-pass temporal filtering with a cutoff frequency of 0.01 Hz. No spatial smoothing was applied. Functional images were then registered to each participant's 3D-MPRAGE anatomical scan using Freesurfer's bbregister (Greve et al., 2009).

Boundaries between early visual areas V1-V4 were manually delineated from a separate retinotopic mapping session, using FSL and Freesurfer software. For those who did not perform retinotopic scanning, areas V1-V4 were predicted from the anatomically defined retinotopy template (Benson and Winawer, 2018). A general linear model analysis was used to identify visually responsive voxels corresponding to the stimulus location, as well as category-selective voxels. In conjunction with the retinotopic maps, a statistical map of the stimulus versus rest contrast of our functional localizer was used to define functionally active voxels in V1-V4 using a threshold of $t > 7$ uncorrected. The fusiform face area (FFA) was identified by contrasting faces versus all other stimulus conditions (objects, houses, scrambled stimuli) and identifying voxels in the fusiform gyrus that exceeded a threshold of $t > 3$ uncorrected. Similarly, the parahippocampal place area (PPA) consisted of voxels in the parahippocampal gyrus that responded more strongly to houses than to all other stimulus conditions ($t > 3$ uncorrected). Finally, the lateral occipital cortex (LOC) was defined by contrasting objects versus scrambled objects ($t > 3$ uncorrected).

Each voxel's time series was first converted to percent signal change, relative to the mean intensity across the run, and the averaged response of TRs 3 to 5 post-stimulus onset was used to estimate stimulus responses. Response amplitudes to each stimulus were then normalized by run to obtain an overall mean of 0 and standard deviation of 1. For each visual area, the multivariate response pattern to a given stimulus was converted into a data vector with associated category label, to be used for training or testing a classifier. We trained a multi-class linear SVM classifier to predict the object category of each stimulus, separately for each region of interest and viewing condition, using the LIBSVM MATLAB toolbox with the default parameter settings (Chang and Lin, 2011). The trained SVM was then tested on independent test runs, using a leave-one-run-out cross-validation procedure (Kamitani and Tong, 2005). (Note that the object category decoding analysis is sensitive to consistency of fMRI responses at both the image level and category level, and we confirmed that essentially the same pattern of classification results was found in early visual areas when decoding was performed to predict the specific image.) We required that classification accuracy for V1, the most reliable visual area for decoding, exceed a minimum of 20% (chance level 12.5%) when averaged across all 3 viewing conditions; otherwise, the data from that participant were excluded due to poor reliability. Data from three participants were excluded based on these criteria, and reported results are based on the data of 8 participants.

To compare the representations of CNNs to those in the human visual cortex, we analyzed the responses of all units within each layer of the CNN to each object image. The responses of a given unit to the set of object images were normalized and converted to z-scores. Next, we calculated the correlational similarity of the responses to all possible pairs of images by computing a 48 × 48 correlation matrix. After setting the main diagonal values to 0, the remaining values solely reflected the correlational similarity of responses to different object images for that layer. The representational structure of these object responses of the CNN could then be compared to the representational structure of object responses obtained from human visual areas by calculating the Pearson correlation coefficient between the correlation matrices. For statistical testing, the Fisher z-transform was applied to these correlation values obtained from each participant when comparing a visual area to a specific layer of a CNN, and t-tests were used to test for significant differences between Pearson correlation values.

Deep neural networks

We evaluated the performance of 8 pre-trained convolutional neural networks (CNNs) using the MatConvNet toolbox (Vedaldi and Lenc, 2015): AlexNet, VGG-F, VGG-M, VGG-S, VGG-16, VGG-19, GoogLeNet, and ResNet-152 (Krizhevsky et al., 2012; Simonyan et al., 2014; Szegedy et al., 2015; He et al., 2016). All networks were pre-trained on the ImageNet 1000-category classification task. Performance on the 16-category classification task was evaluated by determining which of the 16 categories had the highest softmax response to a given image. The training of CNNs with noisy object images was primarily performed using MatConvNet (version 1.0-beta25), with ancillary analyses performed using PyTorch (version 1.6.0). The majority of noise training experiments were performed using VGG-19, although we also confirmed that similar benefits of noise training were observed for AlexNet and ResNet-152.

For 16-category training, all CNNs were trained using stochastic gradient descent over a period of 20 epochs with a fixed learning rate of 0.001, batch size of 24, weight decay of 0.0005, and momentum of 0.9. All weights in all layers of the network were initialized from pre-trained models and were allowed to vary during the training, using backpropagation of the multinomial logistic loss across all 1000 classes. For our first set of analyses, pre-trained VGG-19 was trained with noisy object images presented at a single SSNR level (**Figure 5a**), using images from the 16 categories in the ImageNet training set (20,800 images in total). Separate networks were trained with either pixelated Gaussian noise or Fourier phase-scrambled noise. Training at a single SSNR level led to better performance for noisy object images but poorer performance for noise-free objects. Subsequently, we trained VGG-19 using a combination of noise-free and noisy images, typically using an SSNR level of 0.2 for most experiments. The VGG-19 model used to approximate human SSNR thresholds in Experiment 2 was trained with objects in pixelated Gaussian noise across a full range of SSNR levels from 0.2 to 1. The standard CNN used to fit human SSNR thresholds consisted of pre-trained VGG-19 that received the same number of training examples from the 16 categories using noise-free images only.

For training examples, we used the standard data augmentation pipeline provided by MatConvNet. Training images were derived from the original images by randomly cropping a rectangular region (with width-to-height aspect ratios that randomly varied from 66.67% to 150%) that subtended 87.5% of the length of the original image. The cropped image was resized to 224 × 224 pixels to fit most of the CNN models (except for AlexNet, which used 227 ×

227 pixels). Additionally, the intensity of each of the RGB channels was shifted by a small offset, randomly sampled from a Gaussian distribution with a standard deviation of about 3. The images were then converted to grayscale. Finally, after the SSNR manipulation was applied (as described in ***Visual stimuli***), the average pixel intensity across training samples was calculated and subtracted from each training image.

We trained a 1000-category version of VGG-19 with the full set of training images from ImageNet; these were presented either noise-free, with pixelated Gaussian noise (SSNR 0.2) or with Fourier phase-scrambled noise (SSNR 0.2). Color information from these images was preserved but the same achromatic noise pattern was added to all 3 RGB channels for noise training. The network was trained over 10 epochs using a batch size of 64. All other training parameters were the same as those used in training the 16-category-trained VGG-19.

We quantified the accuracy of standard and noise-trained CNNs at each of 20 SSNR levels (0.05, 0.1, 0.15, … 1). Unlike the human behavioral experiments, CNN performance could be repeatedly evaluated tested without concerns about potential effects of learning, as network weights were frozen during the test phase. The CNN was presented with all 800 object test images at every SSNR level to calculate the accuracy by SSNR performance curve. A 4-parameter logistic function was fitted to the accuracy by SSNR curve and the SSNR level at which accuracy reached 50% was identified as the SSNR threshold for Experiment 1.

For the layer-specific noise susceptibility analysis, we evaluated the stability of the activity patterns evoked by objects presented in progressively greater levels of noise, by calculating the Pearson correlation coefficient between responses to each noise-free test image and to that same image presented at varying SSNR levels. Analyses were performed on each convolutional layer after rectification, the fully connected layers and the softmax layer of VGG-19. A logistic function was fitted to the correlation by SSNR data for each layer, and the SSNR level at which the correlation strength reached 0.5 was identified as the SSNR threshold. If some positive correlation was still observed when SSNR level was 0, then the range of correlation values were linearly rescaled to span a range of 0 to 1, prior to calculating the SSNR threshold.

For the layer-specific classification analysis, multi-class support vector machines (SVM) were trained on the activity patterns evoked by noise-free objects from each of the 16 categories, using data obtained from individual layers of the CNN. After training, the SVMs were tested using the 800 novel test images presented at varying SSNR levels. The SSNR level at which classification accuracy reached 50% (chance level performance, 1/16 or 6.25%) was identified by fitting a logistic function, and served as the classification-based SSNR threshold.

Layer-wise relevance propagation

Layer-wise relevance propagation is a method that identifies diagnostic features that contribute to the prediction of a network (Bach et al., 2015). To do so, the method decomposes the network's output with respect to contributions of individual units, termed relevance scores R as defined below, and back-propagated the scores to the input layer:

$$R_i^{(l)} = \sum_j \frac{x_i w_{ij}}{\sum_i x_i w_{ij}} R_j^{(l+1)},$$

where $R_i^{(l)}$ is the relevance score of the unit i at layer l, x_i is the response of the unit i at layer l, and w_{ij} is the weight connecting the unit i at layer l to unit j at layer l+1. Layer-wise relevance propagation differs from other gradient-based methods in that it takes into account both gradients and unit activations, and may thereby better capture the set of features that are responsible for the network's classification response. In addition to the original implementation (i.e., LRP-0), several variants have been suggested including LRP-ε, LRP-γ, and LRP-zβ (Montavon et al., 2019). Following the guidance of Montavon et al. (2019), we implemented a VGG19-based custom PyTorch script as follows: LRP-0 from the 15th to 19th layers, LRP-ε (ε = 0.25) from the 9th to 14th layers, LRP-γ (γ = 0.05) from the 2nd and 8th layers, and LRP-zβ (lower bound = -1.99 and upper bound = 2.44) for the 1st layer. To create pixel-wise heatmaps, the relevance scores in the pixel space were summed over the rgb channels. Only positive values were taken into account in order to focus on the category-relevant features of a selected object.

2.3 Results

We first compared the recognition performance of 20 human participants and 8 ImageNet pretrained CNNs, including AlexNet, VGG-F, VGG-M, VGG-S, VGG-16, VGG-19, GoogLeNet, and ResNet-152, tested with pixelated Gaussian noise and Fourier phase-scrambled noise. Visual stimuli were obtained from a subset of ImageNet categories (8 animate and 8 inanimate objects; **Figure 3b**) and presented with varying levels of signal-to-signal-plus-noise ratio (SSNR; **Figure 3a**). The average recognition performance of 8 pretrained CNNs and 20 human observers are shown in **Figure 3c**. Although the CNNs achieved comparable performance to humans with noise-free images, their performance started to deviate as noise levels increased, showing poor stability of CNNs under noisy conditions. More interestingly, CNNs and humans showed opposite patterns of performance, that is, CNNs were more impaired by pixelated Gaussian noise, whereas humans were more disrupted by Fourier phase-scrambled noise. This pattern was highly consistent across different CNN architectures except for ResNet-152 (**Figure 3d**). To further analyze these performance differences, we measured the SSNR threshold that corresponded to 50% accuracy using a logistic function fitted to the performance data of individual human observers and CNNs. Human observers exhibited lower SSNR thresholds than CNNs for both types of noise (pixelated Gaussian noise, $t(26) = 15.94$, $p < 10^{-14}$; Fourier phase-scrambled noise, $t(26) = 12.29$, $p < 10^{-11}$). Moreover, humans showed lower SSNR thresholds with pixelated Gaussian noise than Fourier phase-scrambled noise (0.255 vs. 0.315; $t(19) = 13.41$, $p < 10^{-10}$), while CNNs showed higher SSNR thresholds with pixelated Gaussian noise than Fourier phase-scrambled noise (0.535 vs. 0.446; $t(7) = 3.81$, $p = 0.0066$). This finding suggests that CNNs may process noisy objects in a qualitatively different manner than humans.

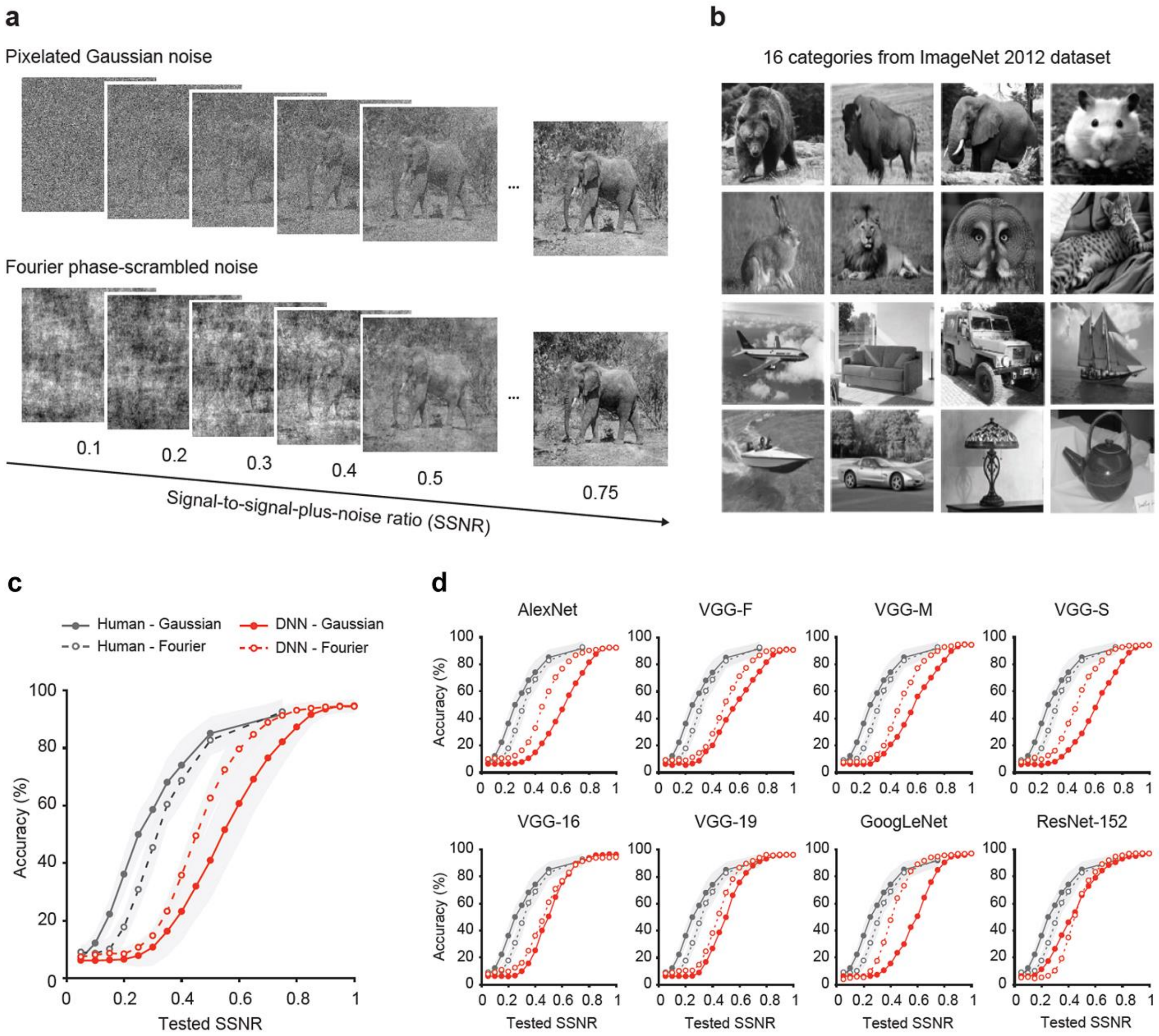

Figure 3. a Examples of an object image in pixelated Gaussian noise or Fourier phase-scrambled noise, shown at varying SSNR levels. **b** Example images from the 16 object categories used in this study. **c** Mean performance accuracy in a 16-alternative object classification task plotted as a function of SSNR level for human observers (black curves) and 8 standard pre-trained DNNs (red curves) with ± 1 standard deviation in performance indicated by the shaded area around each curve. Separate curves are plotted for pixelated Gaussian noise (solid lines with closed circles) and Fourier phase-scrambled noise (dashed lines with open circles). **d** Classification accuracy plotted as a function of SSNR level for individual pre-trained DNN models.

In addition to performance accuracy, we noticed that humans and CNNs exhibited different patterns of responses in their confusion matrices. Examples of their confusion matrices for both types of noise are shown in **Figure 4**. When SSNR levels were 0.75, both humans and CNNs performed nearly perfectly, showing a high frequency of responses along the diagonal. However, when SSNR declined to 0.2, CNN predictions were highly biased to a few categories such as “hare”, “cat”, and “owl”, whereas humans showed a moderate bias to “lion”.

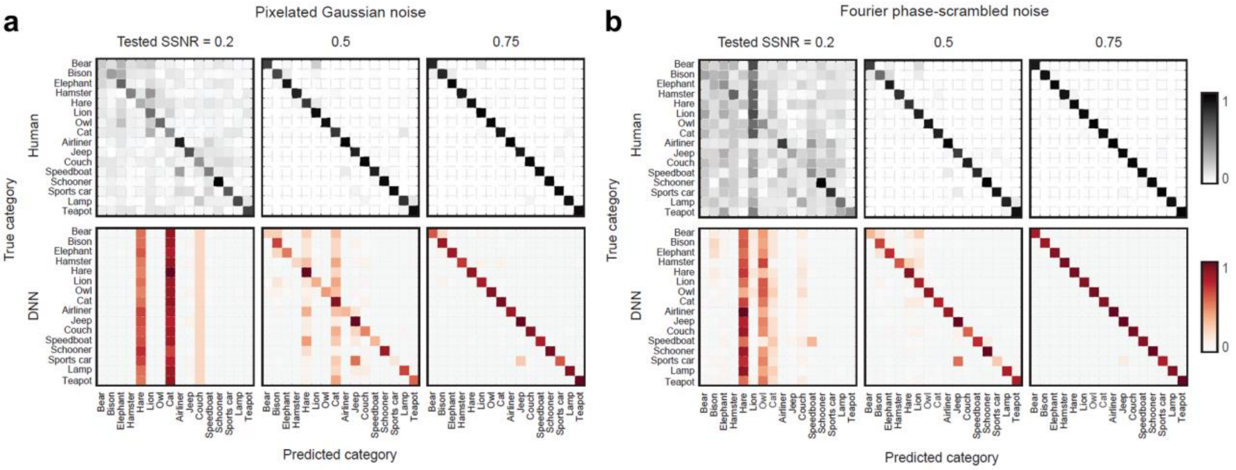

Figure 4. Confusion matrices of human observers and 8 standard pre-trained DNNs. Plots show the relative frequency of predicted category responses (columns) with true categories organized by rows. Confusion matrices are provided for 4 SSNR levels, separately for objects in pixelated Gaussian noise (**a**) and Fourier phase-scrambled noise (**b**).

We further investigated whether training with noisy examples might improve the robustness of pretrained CNNs (primarily using VGG-19) to better match human performance. Unexpectedly, however, we observed that when the CNNs were exclusively trained with noisy examples, their robustness was enhanced but this was accompanied by a loss of accuracy for clear object images (**Figure 5a**). This loss of accuracy for clear images became more prevalent when the training examples contained lower SSNR levels. Note that the CNNs were already pretrained by millions of clear images prior to training with noisy examples, which makes this observation somewhat surprising and provides evidence of their poor stability. Accordingly, we instead trained networks with a combination of noisy and noise-free images and found that this noise training procedure was highly successful, leading to enhanced robustness across a broad range of SSNRs (**Figure 5b**). We chose the network trained with a combination of 0.2 and 1.0 SSNRs for further analysis, as when the training SSNR was reduced to 0.1, the task seemed too challenging to promote stable learning.

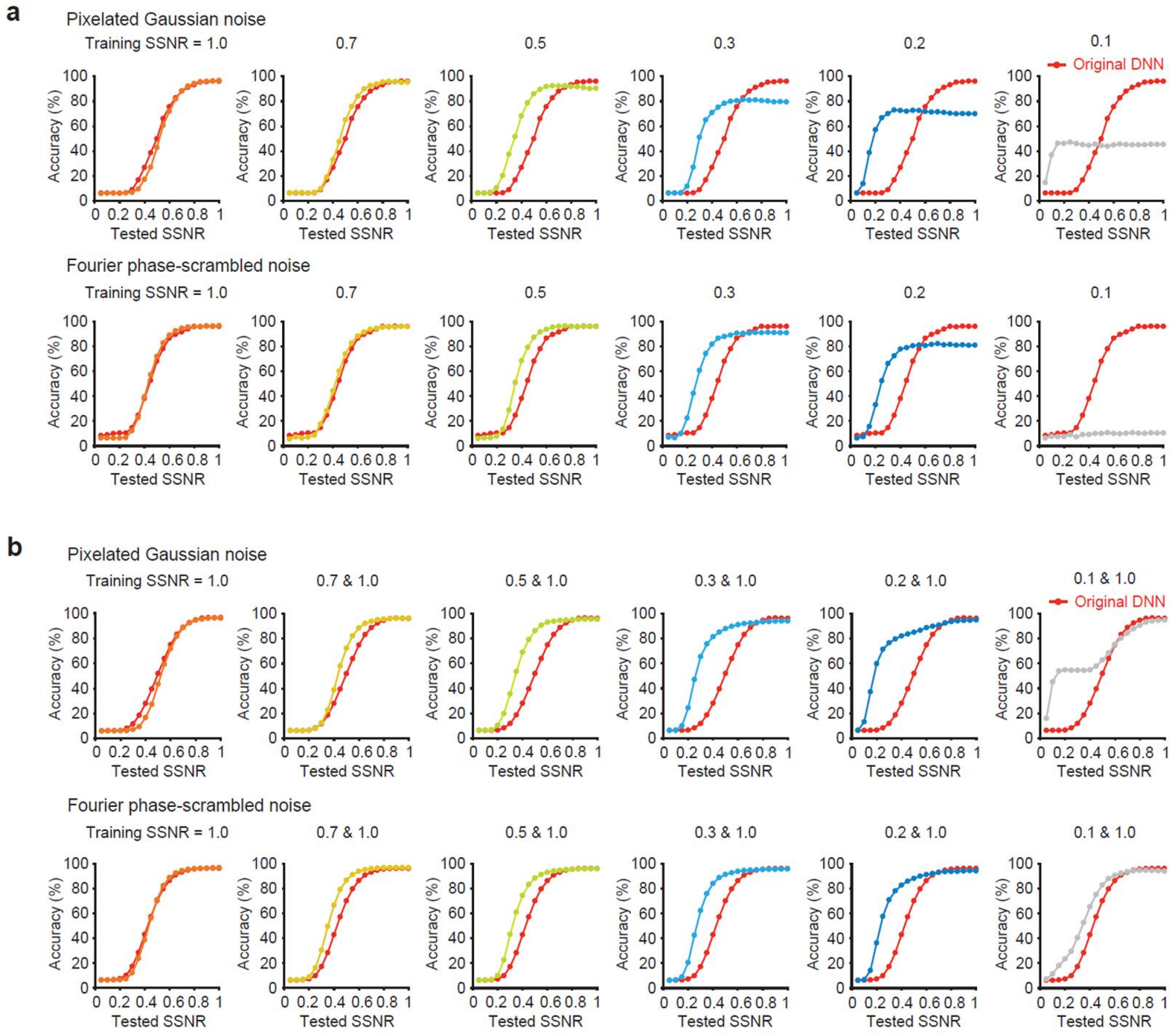

Figure 5. a Impact of training VGG-19 with object images presented at a single SSNR level (1.0, 0.7, 0.5, 0.3, 0.2, or 0.1) when evaluated with novel test images presented at multiple SSNR levels. Accuracy of pre-trained VGG-19 (red curve) serves as a reference in each plot. **b** Impact of training VGG-19 with a combination of noise-free images (SSNR 1.0) and noisy images at a specified SSNR level.

To compare directly the noise-trained and pretrained CNNs, **Figure 3c** is replotted with the performance of the noise-trained CNNs added (i.e., blue curves in **Figure 6a**). Noise training indeed improved robustness up to the human level. Moreover, noise training was more effective for pixelated Gaussian noise than for Fourier phase-scrambled noise, such that the pattern of performance of noise-trained CNNs now better matched that of human observers. We again fitted a logistic function to the performance accuracy of individual CNNs and human observers, and estimated their SSNR thresholds (**Figure 6b**). The histogram of SSNR thresholds demonstrated that the noise-trained CNNs even outperformed the best human observer in both pixelated Gaussian noise and Fourier phase-scrambled noise.

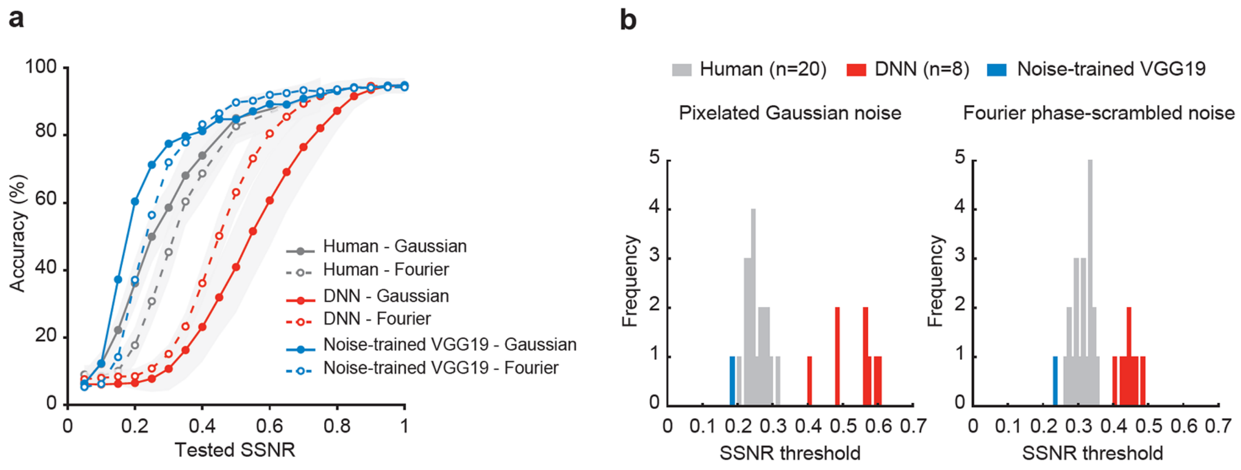

Figure 6. a Mean classification accuracy of noise-trained VGG-19 (blue), human observers (gray), and pre-trained DNNs (red) for objects in pixelated Gaussian noise (solid lines, closed circles) and Fourier phase-scrambled noise (dashed lines, open circles). **b** Frequency histograms comparing the SSNR thresholds of noise-trained VGG-19 (blue), individual human observers (gray), and 8 standard pre-trained DNNs (red).

To analyze the effect of noise training in more detail, we devised a layer-specific noise susceptibility analysis in which the correlation strength between the layer-specific activation patterns of a noise-free image and the noised versions of the same image was measured across a full range of SSNRs (**Figure 7a**). For each layer, the SSNR level required to reach 0.5 correlation was estimated, with a higher SSNR threshold signifying greater susceptibility to noise. The SSNR thresholds of the noise-trained and pretrained CNNs are illustrated in **Figure 7b**. For both pixelated Gaussian and Fourier phase-scrambled noise, the noise-trained CNNs exhibited lower SSNR thresholds overall than the pretrained CNNs, with the impact of noise training emerging at the fourth layer and becoming amplified across successive layers. Moreover, the SSNR thresholds of the noise-trained CNNs gradually decreased across layers, suggesting that the networks effectively enhanced object-related signals while mitigating noise across successive stages of visual processing. However, one could argue that this correlational analysis may not suffice to reveal the impact of training on recognition performance. We also measured classification-based SSNR thresholds using a support vector machine classifier applied to what, explain (see **Materials and Methods** for more details). We observed a similar trend as the impact of noise training became greater in higher layers for both types of noise (**Figure 7c**). This was also shown in the degree of change in the convolutional weights of the noise-trained CNNs. We calculated the average canonical correlation coefficient between the weights of the pretrained CNN and subsequent noise-trained CNN, and found greater changes in the higher layers except the last two fully connected layers (**Figure 7d**). Taken together, the results indicate that noise training alters the representations of the middle and higher layers of CNNs to achieve enhanced robustness to noise.

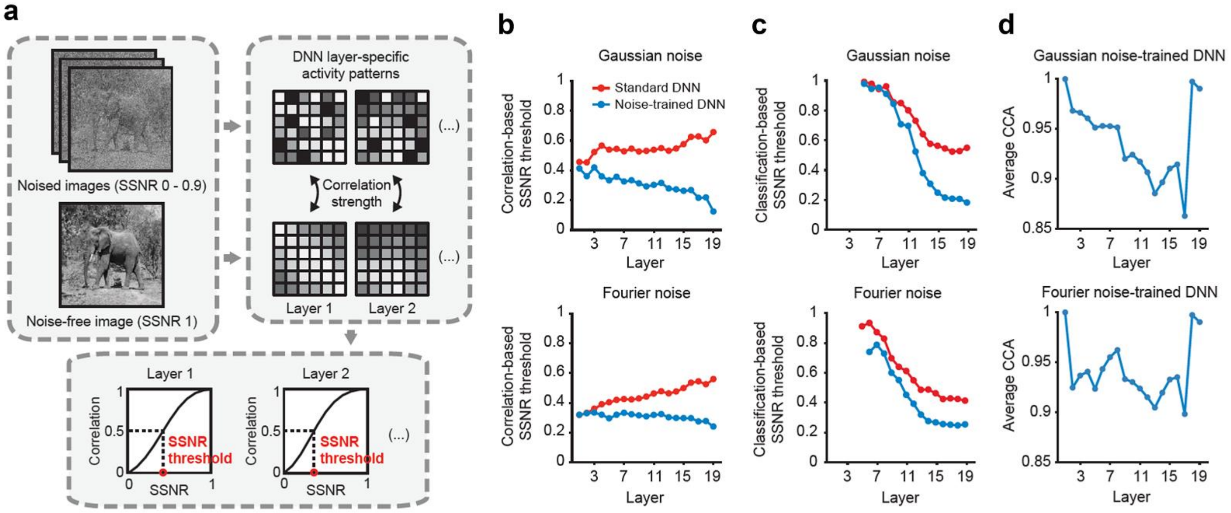

Figure 7. a Depiction of method used for layer-specific noise susceptibility analysis. **b** Correlation-based SSNR thresholds for pre-trained (red) and noise-trained (blue) versions of VGG-19 plotted by layer for objects shown in pixelated Gaussian noise or Fourier phase-scrambled noise. Higher SSNR thresholds indicate greater susceptibility to noise. **c** Classification-based SSNR thresholds plotted by layer for pre-trained and noise-trained networks. Multi-class support vector machines were used to predict object category from layer-specific activity patterns. **d** Similarity of feature representations for pre-trained and noise-trained versions of VGG-19, calculated using canonical correlation analysis (CCA).

So far, we have shown that noise-trained CNNs successfully acquired increased robustness and performed as well as human observers. That said, this does not inform us as to whether or not noise-trained CNNs process noisy objects in the same manner as humans do. They might have utilized a different strategy to handle noise, and if so, the claim that CNNs provide a viable model for human object recognition (Yamins et al., 2014; Yamins and DiCarlo, 2016) should then be only limited to undegraded conditions.

To examine this issue, we ran a second behavioral experiment in which, for each experimental trial a visual stimulus was displayed whose SSNR level gradually increased from 0 (noise only) to 1 (signal only). As soon as participants judged that a target object was recognizable, they pressed a spacebar to stop the process of increasing SSNRs and reported a category. Both accuracy and the SSNR level at which they stopped to report the category were recorded every trial. After they reported the category, participants viewed the stimulus image again with the SSNR level they stopped at and additionally performed a painting task to indicate the object features that appeared most informative by using a mouse pointer. This experiment allowed us to compare the similarity of humans and CNNs in their SSNR thresholds on an image-by-image basis and also the similarity in their diagnostic features for object recognition.

We found that not only did the noise-trained CNN show overall lower SSNR thresholds than the standard CNN, but also exhibited a better correlation with the human behavioral data on an image-by-image basis (**Figure 8a**; $r = 0.55$ vs. 0.27, $z = 6.50$, $p < 10^{-10}$). We should note that this was not necessarily guaranteed because it was also possible for the noise-trained CNN to enhance noise sensitivity to the same degree on every image with little change in the correlation. Despite its better correspondence to humans, the similarity among human observers ($r = 0.94$) was greater still, indicating that noise training alone cannot fully account for human behavioral performance at recognizing noisy objects. Next, we compared the diagnostic regions reported by human observers to the heatmaps obtained from CNNs using layer-wise relevance propagation (Bach et al., 2015; examples are shown in **Figure 8b**), by calculating their spatial correlation and overlap ratio. As shown in **Figure 8c**, the noise-trained CNN better captured the diagnostic regions by humans than the standard CNN, particularly under strong noisy conditions. These results support the notion that noise-trained CNNs do recognize noisy objects, similarly as do humans.

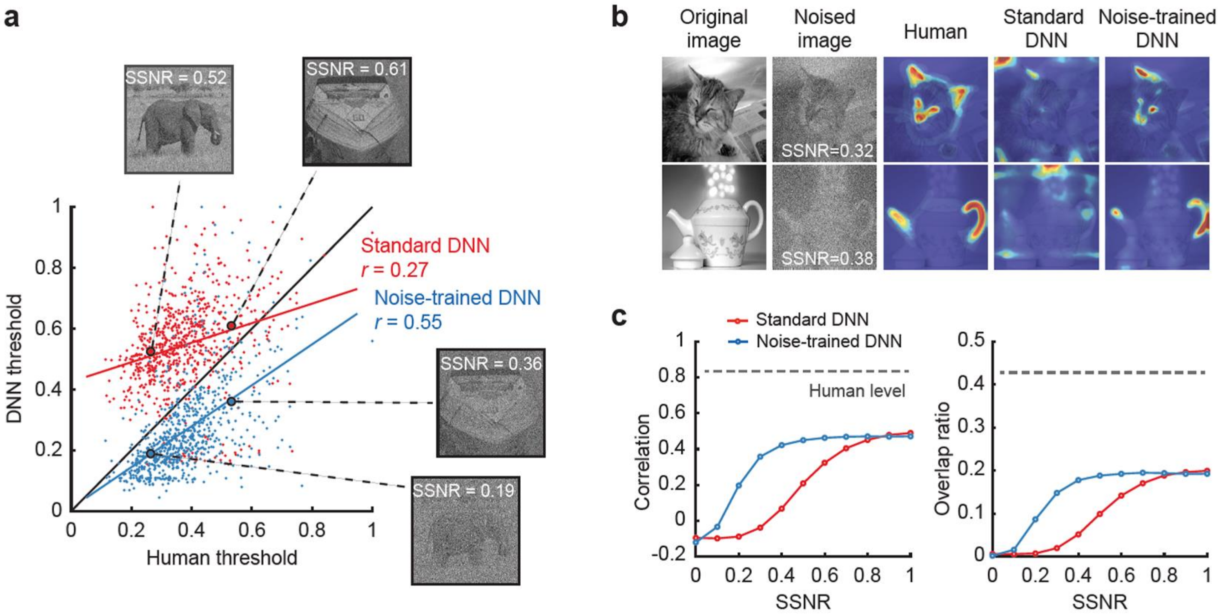

Figure 8. a Scatter plot comparing SSNR thresholds of human observers with the thresholds of standard VGG-19 (red) and noise-trained VGG-19 (blue). Each data point depicts the SSNR threshold for an individual object image. Examples of two object images, shown at the SSNR threshold obtained from standard or noise-trained networks. **b** Examples of diagnostic object features from human observers, standard VGG-19, and noise-trained VGG-19. The mean SSNR level at which human observers correctly recognized the objects is indicated. **c** Correlational similarity and overlap ratio of the spatial profile of diagnostic features reported by human observers and those measured in DNNs across a range of SSNR levels. Gray dashed lines indicate ceiling-level performance based on human-to-human correspondence.

Another way to assess whether noise-trained CNNs process noisy objects in a manner similar to humans is to compare the internal representations of CNNs and humans under noisy viewing conditions. To this end, we collected functional MRI (fMRI) data from 8 participants while they viewed 16 object images in each of 3 viewing conditions: noisy objects with 0.4 SSNR of pixelated Gaussian noise, noisy objects with 0.4 SSNR of Fourier phase-scrambled noise, and noise-free objects. We first assessed whether object processing in visual cortical areas was more disrupted by Fourier phase-scrambled noise than pixelated Gaussian noise, as was observed in our behavioral study. Indeed, decoding of fMRI activity in early visual areas V1-V4 revealed that classification performance for noise-free objects was significantly higher than that of noisy objects in pixelated Gaussian noise, followed by the most degraded performance with Fourier phase-scrambled noise ($t(7) > 4.7$ in all cases, $p < 0.0025$; **Figure 9a**). Interestingly, decoding accuracy for noisy objects in pixelated Gaussian noise did not significantly differ from that of noise-free objects in higher visual areas, suggesting that the impact of pixelated Gaussian noise was somehow lessened in higher visual areas. This may resemble the decreasing pattern of SSNR thresholds across the layers of the noise-trained CNN with pixelated Gaussian noise (**Figure 7b**). By comparison, classification accuracy for objects in pixelated Gaussian noise was significantly higher than Fourier phase-scrambled noise in the lateral occipital complex (LOC; (t(7) = 3.38, p < 0.025) and the parahippocampal place area (PPA; t(7) = 2.54, p < 0.05), but not in the fusiform face area (FFA; t(7) = 1.09, p = 0.31).

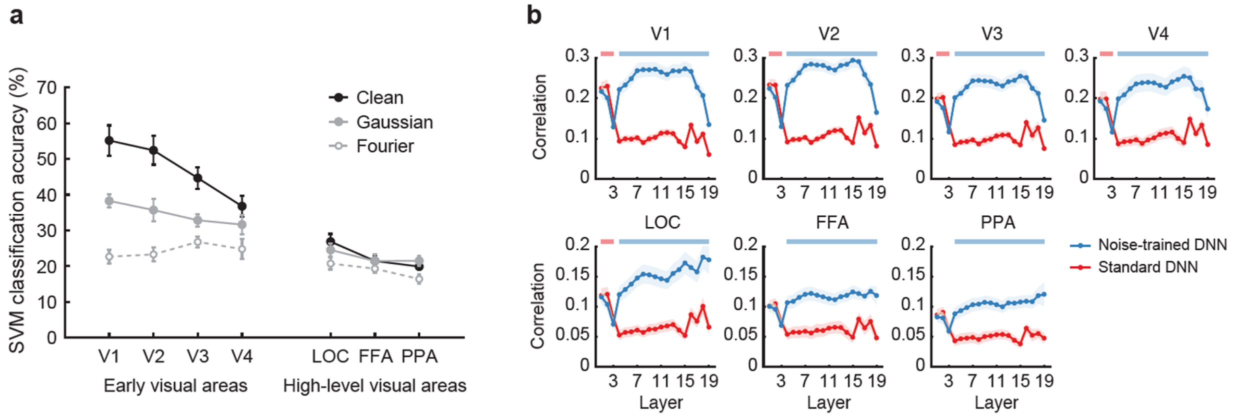

Figure 9. a Classification accuracy for fMRI responses in individual visual areas for clean objects (black filled circles), objects in pixelated Gaussian noise (gray filled circles) and Fourier phase-scrambled noise (gray open circles). Error bars indicate ± 1 standard error of the mean (n = 8). Chance-level performance is 12.5%. **b** Correlational similarity of object representations obtained from human visual areas and individual layers of DNNs when comparing standard versus noise-trained networks (red vs. blue, respectively). Color-coded horizontal lines at the top of each plot indicate a statistically significant advantage (p < .01 uncorrected) for a given DNN at predicting human neural representations of the object images.

Next, we performed representational similarity analysis using both cortical responses and CNN representations and compared these response patterns using the Pearson correlation coefficient (**Figure 9b**). We found that although standard CNNs showed slightly higher correlations than noise-trained CNNs in early layers, noise-trained CNNs exhibited a significant advantage in layers 4 and above. We also observed different patterns between early and high visual areas, such that the correlations at fully connected layers (layers 17-19) markedly dropped in V1-V4, while increased in LOC/FFA/PPA. Taken together, these findings provide neural evidence that noise-trained CNNs reliably account for human recognition behavior under noisy conditions.

Finally, we wondered whether the effect of noise training could generalize to untrained noise types, and if so, to what extent. A previous study claimed that CNNs showed extremely poor generalization capabilities by demonstrating that networks trained with uniform noise failed to show any advantage when tested with salt-and-pepper noise (Geirhos et al., 2018). This may be true as we found similar results in our noise types, namely, that training with pixelated Gaussian noise showed negligible benefits when tested with Fourier phase-scrambled noise and vice versa (**Figure 10a**). However, when we evaluated our noise-trained CNNs on the types of image distortions tested by Geirhos et al. (i.e., salt-and-pepper noise and low- and high-pass filtered images; examples are shown in **Figure 10b**), we observed some degree of successful generalization. The network trained with pixelated Gaussian noise generalized well to salt-and-pepper noise, while the network trained with Fourier phase-scrambled noise showed better performance at high-pass filtered images compared to the standard network (**Figure 10c**). Our findings are also consistent with a recent study that reported that additive Gaussian noise generalized well to unseen image corruption types (Rusak et al., 2020). We also observed that the network trained with both pixelated Gaussian noise and Fourier phase-scrambled noise consistently showed greater robustness across all noise types (**Figure 10c**). These results motivated us to ask whether this network could generalize to real-life noise that would appear different from the noise employed in the experiments.

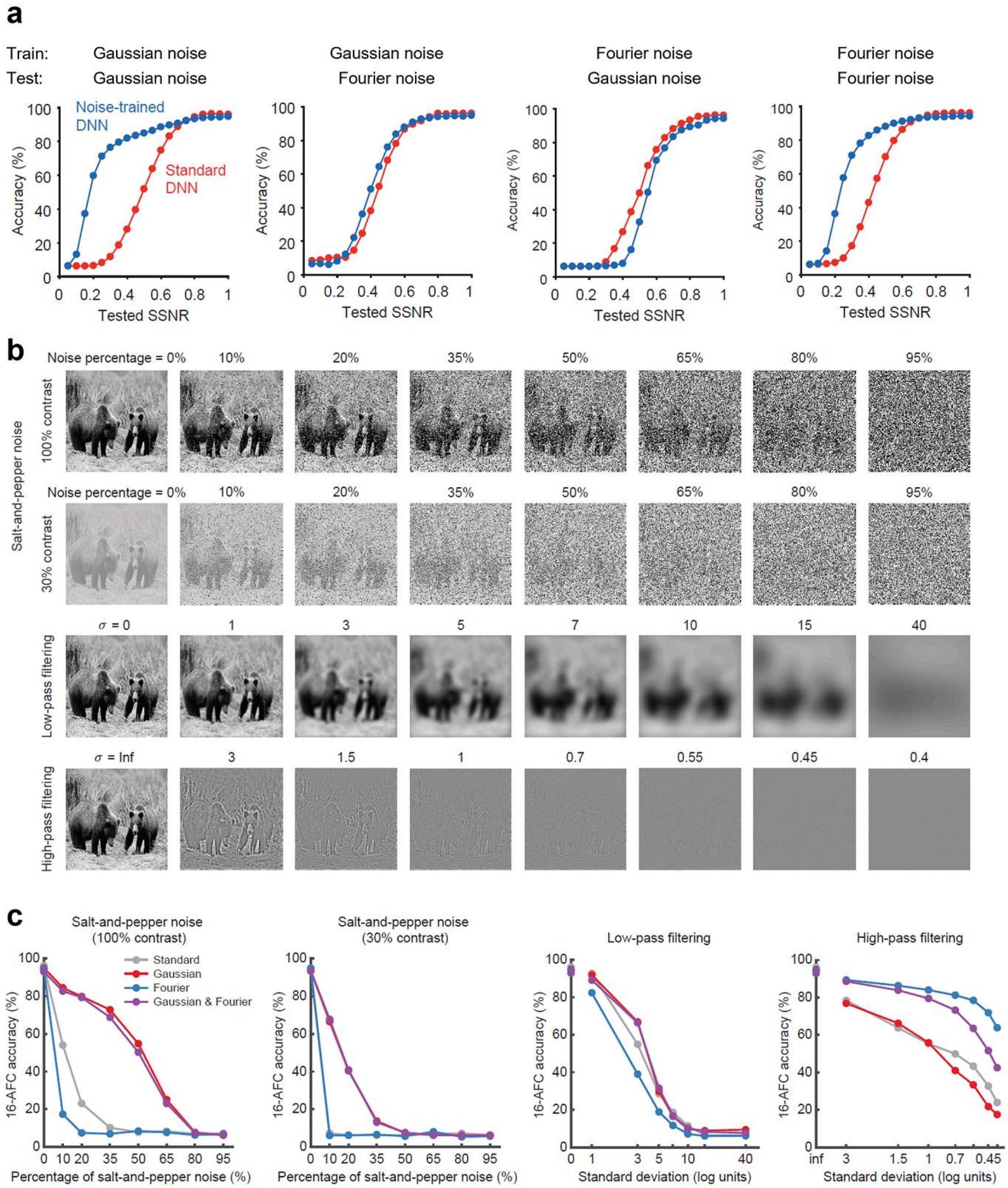

Figure 10. a Mean classification accuracy of noise-trained VGG-19 (blue) when trained with objects in either pixelated or Fourier phase-scrambled noise, and subsequently tested on either type of noise. Performance of standard VGG-19 (red), which lacked noisy image training, is provided for comparison. **b** Examples of images used to test the impact of salt-and-pepper noise, low-pass filtering and high-pass filtering on DNN performance. Image manipulations followed the methods described in Geirhos et al. (2018). **c** Performance accuracy of pre-trained and noise-trained versions of VGG-19 at recognizing images with different types and levels of image distortion.

We collected a dataset of vehicle images under either clear or bad weather conditions (e.g., snow, rain, or fog) and conducted a pilot experiment by asking three observers to rate the degree of noise present in the individual vehicle images. The image set ended up consisting of 102 noise-free and 102 noisy images, with examples shown in **Figure 11a.** These images were then tested by two versions of CNNs, pretrained and noise-trained CNNs with both pixelated Gaussian noise and Fourier phase-scrambled noise on 1000 ImageNet categories. We found that the noise-trained CNN significantly outperformed the standard CNN in noisy weather conditions (**Figure 11b**). Especially when the noise level was stronger, the performance of the noise-trained CNN was significantly higher than that of the standard CNN (**Figure 11c**). Altogether, contrary to the claims of Geirhos et al. (2018), our results demonstrate that CNNs can show some degree of generalization capability depending on which noise types are used for training. Our findings further suggest that it could be possible to build a universal CNN model that is robust to various types of image corruptions by training with a few selected noise types.

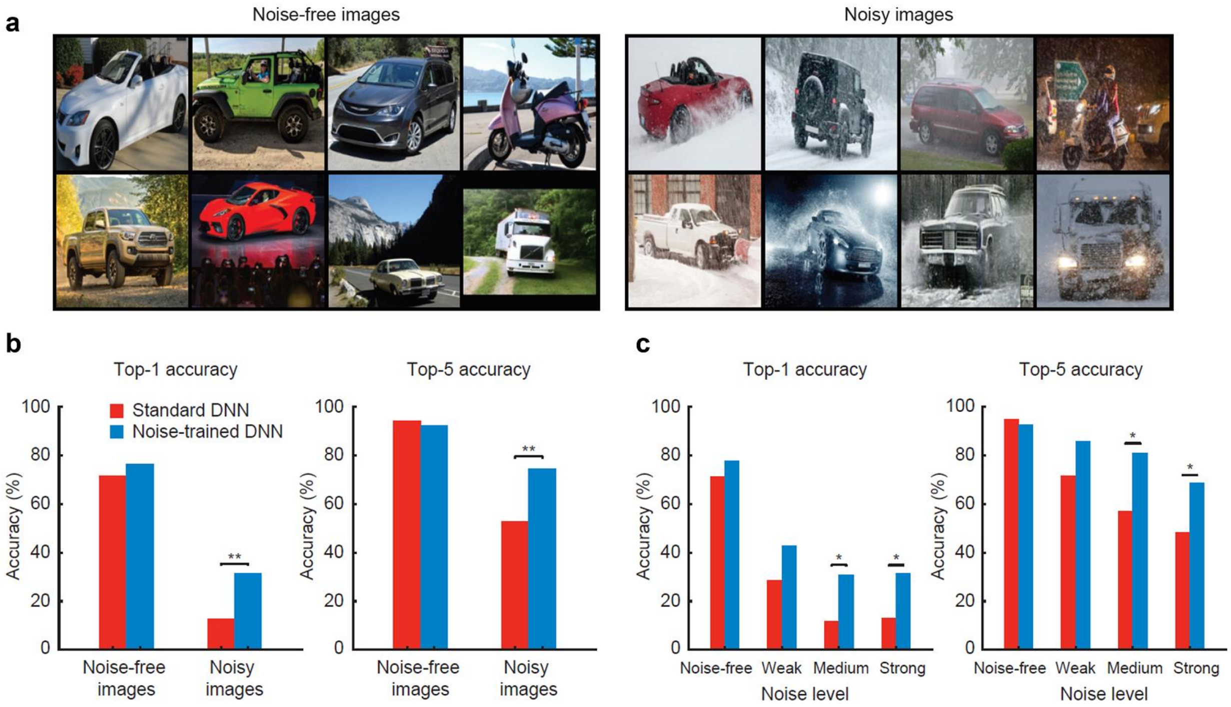

Figure 11. a Examples of real-world images of vehicles in noise-free and noisy conditions. Convertible, jeep, minivan, motor scooter, pick-up truck, sports car, station wagon, and trailer truck, from top-left to bottom-right. **b** Top1 and top5 accuracies of pre-trained VGG-19 (red) and noise-trained VGG-19 (blue) at classifying vehicles in noise-free or noisy weather conditions. Noise-trained VGG-19 outperformed pre-trained VGG-19 at recognizing noisy vehicle images (top1 accuracy, χ2 = 10.29, p = .0013; χ2 = 10.26, p = .0014). **c** Top1 and top5 accuracies sorted by noise-level rating. A statistical difference in performance was observed between models when the noise level was moderate or strong (χ2 > 4.5, p < .05 in all cases). Top5 accuracy means that the correct answer must come from one of the 5 categories with the 5 highest probabilities from a DNN model. Asterisks indicate * p < .05, ** p < .01.

2.4 Discussion

Visual noise is particularly detrimental to the recognition process of many CNN models (Dodge and Karam, 2017; Geirhos et al., 2018; Jang and Tong, 2018; Tong and Jang, 2021). In the present study, we show that CNNs not only fail to recognize noisy objects but also exhibit a pattern of responses to noise that qualitatively differs from that of human observers. Pixelated Gaussian noise which does not have a geometric structure itself tends to severely degrade CNN performance. By contrast, humans are poorer at recognizing objects corrupted by Fourier phase-scrambled noise. This disparity suggests that CNNs do not process objects in the same manner as humans do under noisy conditions.

We further evaluated whether noise-trained CNNs could provide a viable model to predict human behavior in noisy viewing conditions. We observed that those CNNs showed not only better performance at recognizing noisy objects than standard CNNs but also higher correspondence to human behavior in multiple aspects. For example, a previous study reported that CNNs failed to predict the patterns of human behavioral performance for individual image difficulty levels (Rajalingham et al., 2018), but we found that noise-trained CNNs fairly well predicted the recognition thresholds of humans on an image-by-image basis, significantly better than standard CNNs. Moreover, the noise-trained CNNs reliably captured the diagnostic features of objects obtained from human observers. Furthermore, we found that the brain responses to noisy objects in the lower and higher visual areas were better accounted for by noise-trained CNNs. This finding provides supportive evidence that CNNs still can provide a viable model for human vision in line with the previous studies (Yamins et al., 2014; Yamins and DiCarlo, 2016), even under degraded viewing conditions, but only if they are trained on noisy objects. Collectively, our findings provide multi-faceted evidence that noise training allows CNNs to learn noise-robust features that are presumably employed by humans.

While the present study focused on the capabilities of the standard and noise-trained CNNs in predicting human behavior with noisy objects, fundamental questions still remain open: what do the noise-robust features represent and how are they achieved via noise training? From a manifold learning perspective, the network seems to learn intrinsic mapping functions to project corrupted examples back onto the manifold (Vincent et al., 2008). A study with a single-layer denoising autoencoder has demonstrated that more distinctive and less local features were achieved as the amount of noise added into inputs was increased (Vincent et al., 2010). However, if noise levels become too severe, they can become harmful to learning as was found in **Figure 5**, suggesting that there may be an upper bound of noise training in the achievement of noise-robust features. More interestingly, our layer-specific noise susceptibility analysis reveals that noise training has a larger impact on the middle and higher layers of the network, while the low-level features remain mostly unchanged (**Figure 7**). This finding is of particular interest, because it tells us that the mitigated noise effect was not attributed to the low-level features simply better filtering out the noise, but involved with hierarchical processing of denoising. We will further explore the nature of the denoising process across layers in the following chapter.

Another natural question that could arise is whether the robust nature of object recognition in humans is achieved by, at least in part, real-life visual experience with noise. In the auditory system, humans may naturally learn to develop a robust system to auditory noise because the noise is highly prevalent in the real world (Kell et al., 2018; Kell et al., 2019). However, by comparison, visual noise in the real world (e.g., seeing through a glass window on a rainy day or seeing scenes through flakes of snow) is relatively rare. One possibility is that, though they rarely occur, they may impact significantly on the visual recognition system. Another possibility is that noise training may act as internal noise in biological visual systems (Faisal et al., 2008). For instance, the detection of individual photons by photoreceptors is known to follow a Poisson process, which creates some degree of uncertainty to visual systems (Pouget et al., 2000). Whether or how much visual noise contributes to developing robust biological systems is an unexplored question, probably because it is not readily feasible to design a control biological model that only transmits clean signals.

Finally, we have demonstrated that the effect of noise training could generalize to unseen noise types, depending on which types of noise are trained and tested (**Figure 10**). This is somewhat contradictory to the recent claim by Geirhos et al. (2018), who reported extremely poor generalization abilities of CNNs to untrained noise types. Even when CNNs were trained with salt-and-pepper noise and tested with uniform noise, they showed almost zero generalization performance, while they seemed visually relevant. The main difference we could identify is that we performed noise training based on pretrained models, while Geirhos et al. (2018) did not use any pretrained models for noise training, which possibly led to over-fitting due to the small number of training samples. We also found that a network trained with both pixelated Gaussian noise and Fourier phase-scrambled noise showed a meaningful benefit at recognizing vehicles in real-life noisy conditions (**Figure 11**). These findings suggest that it may be not necessary to train all types of noise but rather a well-chosen subset to build a universal noise-robust model, potentially useful for certain applications such as autonomous driving. It will be of future interest to identify what are the representative noise categories needed to be trained for the general purpose of designing a universal robust model.

Acknowledgments

It should be acknowledged that the noise training methods for deep neural networks described in this chapter have led to a patent issued in June 2021: U.S. Patent No. 11030487; Inventors, Tong and Jang; Applicant, Vanderbilt University.

3. Exploring the hierarchical nature of the noise-robust object recognition system

3.1 Introduction

Object recognition in real-world environments involves varying degrees of uncertainty, for example, detecting an object occluded by clutter or recognizing a friend's face wearing a mask. As complete information is not always available, inferring hidden patterns given limited visibility is often needed to perform daily visual tasks. Despite its inherent challenge, human object recognition is surprisingly stable across variations in the amount of information provided. To understand the robust nature of the human object recognition system, vision scientists have traditionally employed visual noise (Harmon and Julesz, 1973; Morrone et al., 1983), as it enables a systematic evaluation of performance as a function of the ratio of signal to noise.

Research has demonstrated that human observers can further improve in their robustness to noise via extensive training procedures (Dosher and Lu, 1998; Gold et al., 1999; Gold et al., 2004; Dosher and Lu, 2005; Dosher et al., 2013). The underlying mechanisms of this learning-induced robustness enhancement have been characterized by an observer model in which external noise and addictive or multiplicative internal noise are defined. Based upon signal detection theory (Tanner and Swets, 1954), d-prime can be experimentally quantified by measuring contrast thresholds with varying degrees of external noise across training sessions. Previous studies have concluded that learning increases the efficiency of perceptual templates in extracting useful signal from noise (Gold et al., 1999; Gold et al., 2004) and may possibly reduce internal additive noise (Dosher and Lu, 1998). Neurophysiological studies have provided additional insights into the underlying mechanisms of enhanced robustness. Rainer and his colleagues trained monkeys to recognize objects with varying degrees of noise and examined learning-related changes in area V4 and prefrontal cortex (Rainer and Miller, 2000; Rainer et al., 2004). They found that V4 neurons exhibited selective enhancement of neural activity at intermediate levels of noise (i.e., an inverted U-shape profile as a function of noise degradation; Rainer et al., 2004), whereas prefrontal neurons demonstrated a leftward shift in the neural response curve of signal intensity (Rainer and Miller, 2000).

The results above may seem surprising given that humans already appear to have a highly robust recognition system against noise even without training (Jang and Tong, 2018). According to the literature (Gold et al., 1999; Gold et al., 2004; Dosher and Lu, 1998), this improvement can be accounted for by at least two mechanisms; one, that humans rarely experience visual noise in real life and thus extensive training with noise effectively alters hard-wired visual channels better filtering out noise, or, two, that training with noise leads humans to be better sensitive to a particular set of stimuli by selectively transmitting the relevant signals from noise. These two effects do not necessarily coincide, because the first implies that the effect of training would be specific to noise, whereas the second implies that the effect of training would be specific to the stimulus. Over the rest of the project, we aimed to assess these two effects and examine the nature of noise training.

As we have shown that CNNs can successfully improve their robustness via noise training in Chapter 2, another important question is whether humans and CNNs would benefit from noise training in a similar manner. Given that CNNs have been suggested to share commonalities with biological visual systems (Yamins et al., 2014; Yamins and DiCarlo, 2016; Güçlü and Gerven, 2015), they may benefit similarly. On the other hand, Chapter 2 has revealed that CNNs appear to lack a basic mechanism for noise robustness, so they may benefit differently to achieve robustness.

To address these questions, we developed a learning paradigm where human observers were trained on an object recognition task with natural images in the presence of random Gaussian noise. We adopted a pretest-posttest experimental design, where individual observers were evaluated on their robustness to noise before and after training. This was quantified as the threshold of noise level at which recognition performance reached 57% of accuracy and was estimated by a QUEST staircase procedure. In each trial of the training procedure, human observers were displayed by a noised image where its signal-to-noise ratio became gradually higher. The observers reported an object category out of 16 choices (8 animate and 8 inanimate) when it was recognizable and received accuracy feedback that incentivized them to respond faster and more accurately. To actively engage them in learning, they were asked to identify the most informative features in noise by annotating with a mouse pointer.

It should be noted that although previous studies have shown that human observers could improve in their robustness to noise via training, the studies either tested on a simple recognition task such as orientation discrimination (Dosher and Lu, 1998; Dosher and Lu, 2005; Dosher et al., 2013) or used identical stimuli for training and testing (Gold et al., 1999; Gold et al., 2004; Rainer and Miller, 2000; Rainer et al., 2004). Therefore, it was not determined whether those findings would generalize to our task which used complex object stimuli in natural backgrounds and tested on a novel image set. Moreover, we have seen in Chapter 2 that noise training greatly enhanced the robustness of CNNs, but this may not be necessarily the case for human observers.

Next, we sought to disentangle the two potential effects of noise training in a second experiment. We used the same learning paradigm as the first experiment but divided observers into two separate groups: One group was trained by animate categories and the other group was trained by inanimate categories. Both groups were evaluated on their robustness to noise before and after training using all animate and inanimate categories. There are two possible outcomes from the experiment. If their noise robustness increased even when they were tested on untrained categories, this would indicate that noise training has a category-general effect. If their noise robustness increased only when they were tested on trained categories, this would mean that noise training has a category-specific effect. Similarly, two separate CNNs were trained, one on noisy objects from animate categories and the other on noisy objects from inanimate categories.

We first found that both human observers and CNNs successfully improved their robustness and generalized well across objects within the category. However, when the categories that were not trained with noise were tested, human observers failed to demonstrate a significant benefit of training. By contrast, CNNs showed a modest level of generalization to the categories that were not trained with noise (i.e., category-general effects), though not as much as when tested on the categories trained with noise (i.e., category-specific effects). Furthermore, we applied the layerwise noise susceptibility analysis, similarly as in Chapter 2, to elucidate the effects of noise training across the layers of CNNs. It revealed that the category-general effect was most pronounced at the early and middle layers of the CNN, whereas the category-specific effect was mainly observed at the middle and higher layers. These findings suggest that the robust nature of our object recognition may be accomplished by a multi-stage process along the visual hierarchy.

3.2 Materials and methods

Participants

We recruited 20 participants in behavioral experiment 1 and another group of 32 participants in behavioral experiment 2. None of them participated in both experiments. All participants reported having normal or corrected-to-normal visual acuity, and provided informed written consent using electronic consent forms (REDCap). Participants received monetary compensation or credits for a course. All aspects of this study followed the guidelines of and were approved by the Vanderbilt University Institutional Review Board.

Visual stimuli

Visual stimuli were collected from the ILSVRC-2012 dataset (Russakovsky et al., 2015). We selected 16 categories out of 1,000 which consisted of 8 animate and 8 inanimate objects including bear, bison, elephant, hamster, hare, lion, owl, tabby cat, airliner, couch, jeep, schooner, speedboat, sports car, table lamp, teapot. All test images were grayscaled to remove strong color cues but only preserving local spatial information.

The level of presented noise was manipulated by varying the signal-to-noise ratio of images, which followed the method in Chapter 2. To briefly recap, we introduced a signal-to-signal-plus-noise ratio (SSNR) as w in $\boldsymbol{T} = w \cdot \boldsymbol{S} + (1-w) \cdot \boldsymbol{N}$, where $\boldsymbol{S}$ is a source image, $\boldsymbol{N}$ is noise, and $\boldsymbol{T}$ is a target image where noise is added to the original source image. The full range of SSNR levels can vary from noise only (SSNR = 0) to signal only (SSNR = 1). The range of pixel intensities for $\boldsymbol{S}$ and $\boldsymbol{N}$ were all set to 0-1. For visual noise $\boldsymbol{N}$, we generated random noise sampled from a Gaussian distribution centered at 0.5 with a standard deviation of 1/3 so 99.97% of pixel intensities were in the range of 0-1. Pixels exceeding this range were clipped to fall within the 0-1 range.

Behavioral experiment 1

In the first experiment, we sought to test whether human observers could improve on their robustness to noise after extensive training of a recognition task with noisy objects. To do so, we developed a learning paradigm based on a pretest-posttest experimental design. For the training procedure, 800 images from the ImageNet validation set were displayed throughout 4 sessions. On each trial, participants viewed an image where a target object gradually emerged from noise as the SSNR level increased from 0 to 1 by 0.025 every 400ms. Participants paused at the SSNR level where the target was recognizable and performed a sixteen-alternative forced-choice categorization task. They were rewarded for responding faster and more accurately by monetary incentives and received accuracy feedback every trial. In addition, participants were asked to identify the most informative features in noise by annotating region(s) using a mouse pointer, which potentially helped them better engage in learning. This annotation task occurred in-between after the category response and before feedback. Participants performed practice trials prior to the main experiment to get familiar with the whole procedure using another novel image set.

Before and after the training procedure, all participants were evaluated on their robustness to noise using a QUEST staircase procedure (Watson and Pelli, 1983). In QUEST, the accuracy-SSNR curve of an observer was assumed to follow a Weibull psychometric function, and prior knowledge was continuously updated on a trial-by-trial basis based upon the observer's response and used to determine the SSNR level of the next stimulus. The parameters of the Weibull function were pre-estimated based on the average recognition accuracy by SSNR curve acquired in Chapter 2: beta (4.07), delta (0.15), and gamma (0.0625). The noise robustness was determined by an SSNR threshold level with a 57% performance criterion (inflection point) in a 48-trial run, where the prior threshold estimate and standard deviation were initially set to 0.5 and 0.4, respectively.

The 96 test images (6 examples per category) used in the pretest and posttest sessions were collected from the ImageNet training dataset. We recruited 4 independent raters (including one of the authors) to determine the noise thresholds of all 96 images by using the same procedure as in the main experiment but without the annotation task. Based on the reported noise thresholds, two separate image sets (48 for each) whose difficulty levels were closely matched were assigned to the pretest and posttest, with the assignment mapping counterbalanced across participants. To minimize potential demand characteristics, experimenters did not mention to participants that the purpose of the study was related to learning.

Behavioral experiment 2

The follow-up experiment was designed to answer the question: Is training with noise category-specific or category-general? A total of 32 participants were assigned to one of two groups. Half of the participants (odd numbers) underwent training with animate categories and the other half (even numbers) underwent training with inanimate categories. For each group, 400 images were shown over 3 sessions. Most participants completed all sessions within a week. The training procedure was identical to Experiment 1.

Similarly, the QUEST procedure was used to measure the noise robustness of individual observers before and after training. Regardless of the category types on which they were trained, all participants were tested with both animate and inanimate categories. Three raters were recruited to report the noise thresholds of 192 images obtained from the ImageNet training set in the same manner as Experiment 1. Based on the reported noise thresholds, 4 sets of images (48 for each) were created with their average difficulty levels closely matched. Two of them were used for pretest and the other two were used for posttest. The assignment order was counterbalanced across participants. Two QUEST runs were carried out at the pretest or posttest to measure the noise robustness with animate and inanimate categories, separately. Whether animate or inanimate categories were tested first was also counterbalanced across participants within an animate/inanimate training group. All other parameters and settings for QUEST were identical to Experiment 1.

Training of convolutional neural networks

In this study, we evaluated 6 CNNs including AlexNet, VGG16, VGG19, GoogLeNet, ResNet50, and Inception-v3 (Krizhevsky et al., 2012; Simonyan et al., 2014; Szegedy et al., 2015; He et al., 2016; Szegedy et al., 2016), and primarily relied on AlexNet and VGG19 for layerwise analyses. All networks were initialized with ImageNet pretrained weights and further trained on noisy objects from 16 categories. Noisy training images were generated by randomly sampling object images for each training batch and determining the SSNR level for each object by randomly sampling from a uniform distribution ranging from 0.2 to 1. We adopted the standard data augmentation methods of PyTorch that are commonly used in image classification, i.e., RandomResizedCrop and RandomHorizontalFlip. Images were resized to 224 × 224 grayscale images, and then converted to RGB by concatenation to fit them to the pretrained models. Individual images were normalized by the mean (0.449) and standard deviation (0.226) of the ImageNet training samples. The networks were trained using a stochastic gradient descent optimizer with a fixed learning rate of 0.01 and a weight decay of 0.0001 for 100 epochs. For our control CNNs, the same pretrained models were trained on noise-free images from 16 categories using the same procedures described above. All training experiments were conducted using PyTorch.

Layerwise noise susceptibility analysis

For the layerwise noise susceptibility analysis, we first evaluated the consistency of the activity patterns evoked by objects presented in progressively greater levels of noise via correlation. To be specific, we calculated the Pearson correlation coefficient between the responses to an individual noise-free image and to the same image presented at varying SSNR levels. We assumed that, if the network maintained high robustness to noise, the responses to noisy images would be quite similar to those evoked by noise-free images, and indeed we observed a monotonic increase in correlation strength with increasing SSNR levels (data not shown). The analysis was performed on all layers of VGG19 including 16 convolutional layers and 3 fully connected layers. The correlation by SSNR curve was rescaled to 0-1 to get rid of any remaining offsets at 0 SSNR. A logistic function was fitted to the correlation by SSNR curve for each layer, and the SSNR level at which the correlation strength reached 0.5 was identified as the SSNR threshold.

In addition to the correlation-based SSNR threshold, we trained 16-way support vector machines on the layerwise activity patterns evoked by noise-free objects obtained from the ImageNet training set (1300 images per category). The trained classifiers were then tested on the 800 ImageNet validation images presented at varying SSNR levels. The SSNR level at which classification accuracy reached 50% was identified by fitting a logistic function and served as the classification-based SSNR threshold.

3.3 Results

We first sought to determine whether human observers would achieve better robustness to noise via noise training. This was previously observed for CNNs in Chapter 2 (**Figure 6**) but was not evaluated in humans. To this end, 20 human observers participated in training sessions where they performed a 16-way classification task under stress-test visual noisy conditions with trial-by-trial feedback. Their SSNR thresholds which corresponded to 57% of accuracy performance were estimated by a novel image set from the same 16 categories using a QUEST procedure before and after noise training. Correspondingly, six ImageNet pretrained CNN models (Krizhevsky et al., 2012; Simonyan et al., 2014; Szegedy et al., 2015; He et al., 2016; Szegedy et al., 2016), including AlexNet, VGG16, VGG19, GoogLeNet, ResNet50, and Inception3, were further trained on the same 16 categories with a wide range of noise levels included (i.e., from 0.2 to 1.0 SSNR). The SSNR thresholds of the 6 CNNs were similarly estimated by a novel image set before and after noise training. **Figure 12a** shows the SSNR thresholds of human observers and CNNs before and after noise training. Both significantly improved in their robustness to noise. In particular, CNNs initially showed poor robustness with approximately an SSNR threshold of 0.6 but achieved near human-level performance after training. This result suggests that the effect of noise training can successfully generalize at least across different objects within the same category.

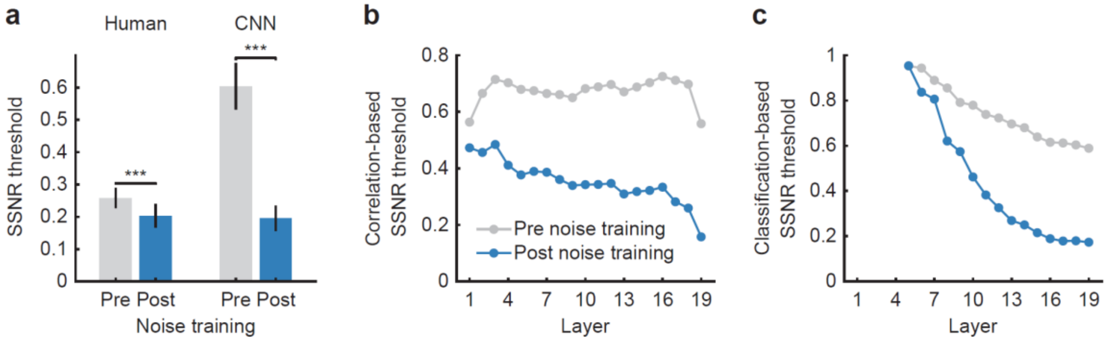

Figure 12. a SSNR thresholds of human observers and CNNs before (gray) and after (blue) noise training. **b** Correlation-based SSNR thresholds of VGG19 before (gray) and after (blue) noise training. Higher SSNR thresholds indicate greater susceptibility to noise. **c** Classification-based SSNR thresholds of VGG19 before (gray) and after (blue) noise training.

A layerwise noise susceptibility analysis was performed based on either Pearson correlation (**Figure 12b**) or an SVM classifier (**Figure 12b**) in order to assess the effect of training across layers of the CNN. VGG19 was selected for analysis. The results were identical to those in Chapter 2 (**Figures 7b-c**), as they only differed in which deep learning package was used for training. Both demonstrated that noise training successfully decreased the SSNR thresholds across all layers and the effect of training appeared to be amplified as the layer increased.

Next, we tested whether observers would still benefit from noise training if the categories we tested were different from the ones we trained with noise. If an observer showed improvement in robustness even when novel categories that were never trained with noise were tested, it would indicate that noise training has a category-general effect; otherwise, noise training has only a category-specific effect. We divided a total of 32 human observers into 2 groups, one trained with animate objects presented in noise and the other trained with inanimate objects presented in noise. All participants were tested on their SSNR thresholds for both animate and inanimate objects before and after training. Correspondingly, two separate CNNs trained on either noisy animate or noisy inanimate objects were tested on their SSNR thresholds for both object categories before and after noise training. Interestingly, human observers and CNNs demonstrated different patterns of improvement in noise robustness (**Figure 13a**). The recognition performance of human observers significantly improved when the trained and tested categories were identical, as consistent with **Figure 12a**, whereas little improvement was seen when the trained and tested categories were different. By contrast, CNNs showed significant improvement regardless of the type of categories, though they exhibited greater benefit when the trained and tested categories were identical.

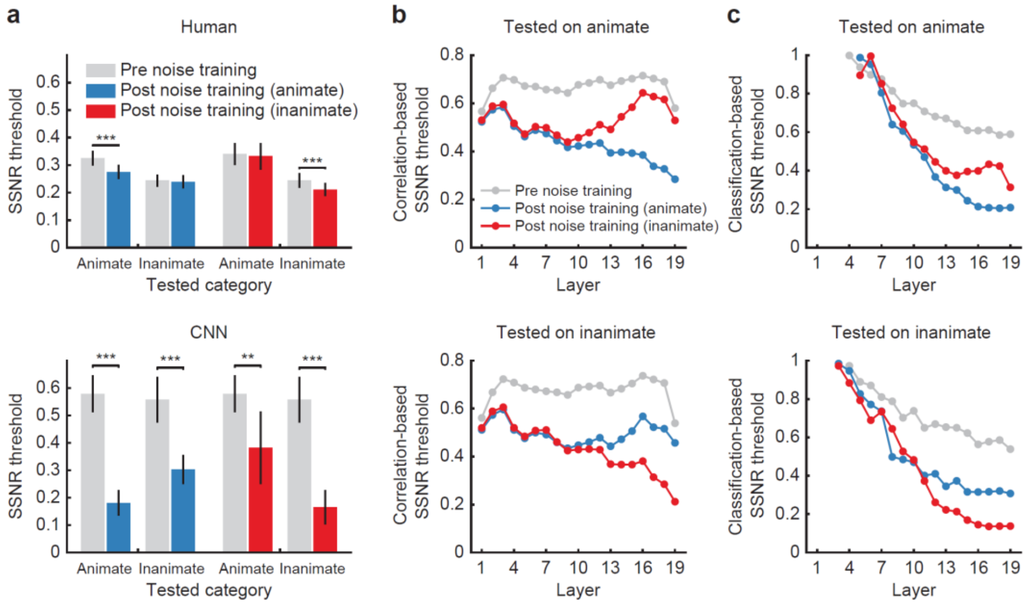

Figure 13. a SSNR thresholds of human observers and CNNs before (gray) and after (blue, trained on animate categories; red, trained on inanimate categories) noise training. **b** Correlation-based SSNR thresholds of VGG19 before and after noise training. Higher SSNR thresholds indicate greater susceptibility to noise. **c** Classification-based SSNR thresholds of VGG19 before and after noise training.

In the same manner, the layerwise noise susceptibility analysis was performed in two scenarios: one under the condition where the network trained on animate categories with noise were tested on animate categories and the other under the condition where the network trained on animate categories with noise were tested on inanimate categories, or vice versa. Strikingly, the effect of noise training appeared to differ at different levels of the CNN. The correlation-based SSNR thresholds substantially decreased in early layers, compared to the control network, regardless of whether that category was trained with noise or not (**Figure 13b**), indicating the category-general effect. However, the difference between the two conditions (i.e., trained and tested on the same categories vs. trained and tested on different categories) became pronounced at the middle layers and larger in the higher layers. When the network was tested on trained categories, the SSNR thresholds continued to decrease, whereas the network was tested on untrained categories, the SSNR thresholds rather increased again. This suggests that the higher layers of the CNN mainly benefit from noise training only when the objects from trained categories were processed, indicating the category-specific effect. The same pattern was observed by the classification-based SSNR thresholds (**Figure 13c**).

The observation that human observers did not show category-general effects led us to think that humans may already possess general noise-robust mechanisms allowed to handle certain levels of noise. If so, would the CNN pretrained on weak noise levels better capture the robustness pattern of human observers? In other words, what if the network is initially trained on both animate and inanimate categories by weak noise levels (i.e., category-general training by 0.5 to 1.0 SSNRs) and further trained on either one of the two categories by strong noise levels (i.e., category-specific training by 0.2 to 1.0 SSNRs)? We found that the category-specific effect was still observed, while the category-general effect in CNNs disappeared (**Figure 14a**). Accordingly, the robustness pattern by human observers was now better captured by those CNNs; however, interestingly, training with inanimate categories presented in strong noise rather harmed the performance when tested on animate categories compared to initial training with both categories by weak noise. This pattern was not observed in humans. This was further elucidated by the layerwise noise susceptibility analysis (**Figures 14b-c**). The category-general effect marginally existed in the early layers, whereas the category-specific effect rather detrimentally influenced the middle and higher layers of the network.

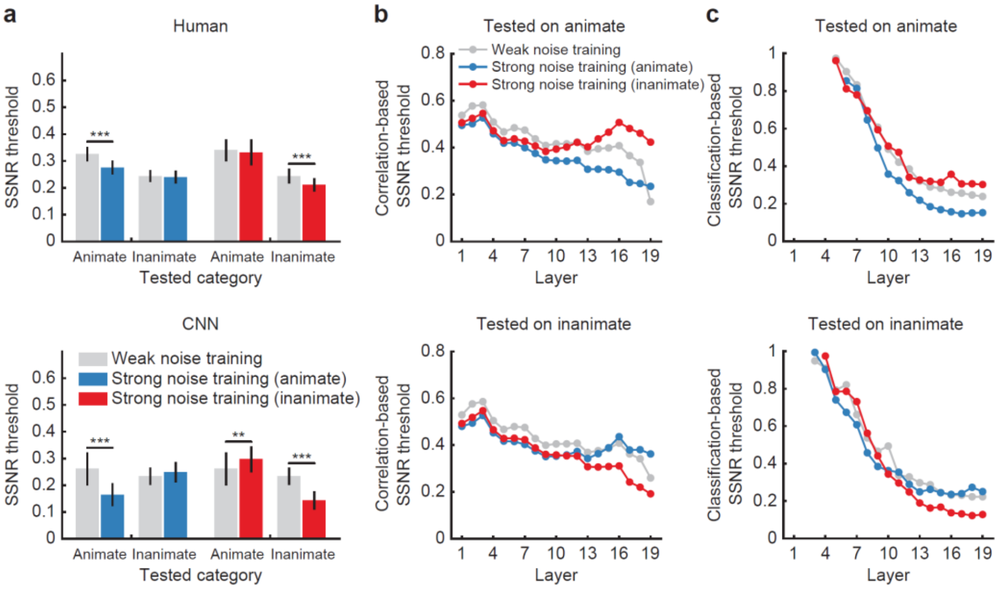

Figure 14. a SSNR thresholds of human observers (identical to **Figure 13a**, but displayed for comparison) and CNNs after category-general training by weak noise levels (gray) and after category-specific training by strong noise levels (blue, trained on animate categories; red, trained on inanimate categories). **b** Correlation-based SSNR thresholds of VGG19 after category-general training by weak noise levels and after category-specific training by strong noise levels. **c** Classification-based SSNR thresholds of VGG19 after category-general training by weak noise levels and after category-specific training by strong noise levels.

3.4 Discussion

In the present work, we examined the robust nature of object recognition by comparing human observers and CNNs in their learning capabilities of noise robustness. Though the effect of noise training was well generalized to novel images within the same categories in both human observers and CNNs, they exhibited a marked difference when tested on noisy examples from the categories that were never trained with noise. Human observers were only able to improve their robustness when tested on the categories they were trained on with noise, whereas CNNs demonstrated some degree of generalization to untrained categories. This result again supports the idea that CNNs would have a fundamentally different system in a way of coping with noise compared to humans. That is, CNNs do not seem to possess any basic mechanisms for robust object processing, as contrasted to humans exhibiting only category-specific learning effects.

We also found that different hierarchical levels of CNNs appeared to involve robust object processing in different manners. Our layerwise susceptibility analysis showed that the category-general noise training effect emerged as early as the second or third layer of the CNN, suggesting that the early level of the visual system needs to be changed to acquire category-independent robustness against noise. By comparison, the category-specific effect was more prominent at the middle and higher layers. Although the principles of learning and attention would differ from each other, our findings remind the previous neuroimaging study by Pratte et al. (2013), demonstrating that top-down attention would act as an external noise filter in the early stage of visual processing (as it may correspond to the category-general and noise-specific effect of noise training) and act as a target-relevant signal amplifier in the later stage (as it may correspond to the category-specific effect of noise training).

A few fundamental questions still remain to be clarified. How do the models achieve stronger robustness to noise? In other words, what are the mechanisms underlying the category-general and category-specific effects of noise training? One possible strategy is the principle of averaging to mitigate the detrimental impact of noise such that different neurons transmit the same signal so it can increase signal power while mitigating noise. This would unavoidably lead to an increased redundancy in signal processing, but would lead a model to be better robust to noise. Stringer et al. (2019) has demonstrated that biological neural networks may be optimized on the verge of satisfying both efficiency and robustness, which is possibly lacked in CNN models.

4. Fundamental differences in face and object processing with a developmental sequence of blurry to clear image training

4.1 Introduction

The symptom of blurry vision is one of the most common problems in optometry and ophthalmology. It can be caused by refractive errors such as myopia, hyperopia, and astigmatism, or by retinal deficits such as glaucoma and cataract. Often neglected, however, is the fact that a person who has normal or corrected-to-normal visual acuity also experiences blurry vision in daily life. For instance, objects that appear farther away from the central field of vision look less distinct because of the larger receptive fields of ganglion cells in the periphery of the retina (Strasburger et al., 2011). In addition, when a person fixates a particular object is focused, objects which are either nearer or further than the current focus can appear blurred on the retina (Sprague et al., 2016). The ability to recognize such blurry objects is often important for our daily visual tasks such as navigating through dense crowds or crossing streets where the traffic is heavy. From an evolutionary perspective, rapid predator detection in the periphery is a life-and-death necessity.

Behavioral studies have demonstrated that humans are surprisingly good at recognizing blurry objects. Thorpe et al. (2001) examined if human observers were able to determine whether a stimulus contains an animate or inanimate object by varying the location of the stimulus across the horizontal visual field and observed well-above chance performance at the extreme eccentricity of 70.5º. Dodge and Karam (2017) have shown that human observers exceeded chance-level performance in a 10-alternative forced-choice dog breed categorization task when the stimulus was highly blurred. Given that blurry objects inevitably entail the loss of fine-detailed information, it is surprising how humans well can recognize objects across a range of blur levels. Would this robust nature of the human recognition system be attributed to a particular mechanism in the visual system or learned via experience?

Though many possible explanations could exist, the observation that visual acuity in infants is very poor at birth but rapidly improves over the first year of life (Dobson and Teller, 1978) particularly motivated us to explore whether early experiences with blurry vision during infancy may confer some ecological benefit in developing robust recognition systems. This hypothesis has been originally suggested by clinical studies of patients born with congenital cataracts. The patients who lost the opportunity of experiencing a developmental period of blurry to clear vision later exhibited severe deficits in configural face processing even though they received treatment (e.g., days of deprivations were 53 to 586 in Geldart et al., 2002) and had at least 9 years of experience with clear vision (Le Grand et al., 2001; Geldart et al., 2002). Other studies have also found that these patients demonstrate poor performance at illusory contour perception (Putzar et al., 2007) and at using pictorial depth cues (McKyton et al., 2015). These findings suggest that the early developmental sequence of blurry to clear visual inputs may be critical for developing integrative feature-binding processes, which may contribute to the robustness of human object recognition.

If this hypothesis is true, can this idea be applied to CNNs as a means to improve their object recognition skills to be more robust? Recent studies have shown that CNNs are exceptionally poor at recognizing blurry objects as compared to human observers (Dodge and Karam, 2017; Geirhos et al., 2018). The poor robustness of CNNs might be a consequence of the lack of early training period with blurry inputs, as suggested by the findings with congenital cataract patients.

That is, it is conceivable to think that a developmental sequence of blurry to clear image training may allow CNNs to develop object recognition skills based on more human-like and blur-robust features. A recent study has directly addressed this hypothesis and has offered some support for this view (Vogelsang et al., 2018). The authors initially trained a CNN with blurry face images, followed by clearer ones, and observed that the CNN showed greater robustness to a wide range of blur levels than a control CNN trained exclusively with clear face images. Furthermore, the CNN progressively trained with blurry to clear faces were shown to have larger receptive fields at the earliest layer than the control CNN, which supported the view that a sequence of blurry to clear image training promotes integrative feature processing of faces.

On the other hand, it is conceivable that these reported findings with faces may not be directly applicable to general object recognition. A long-standing view is that faces are a special category of objects; humans are experts at identifying different faces, despite the considerable challenges due to the highly similar and shared configuration across faces (Tong and Nakayama, 1999; Sinha, 2002). The high proficiency of face processing may naturally arise from its social significance as a developmental necessity to interact with caregivers and others. Many neurophysiological and neuroimaging studies have indicated that the face recognition system is supported by distinct neural regions that are separate from those that represent other object categories (Kanwisher et al., 1997; Farah et al., 2000; Grill-Spector et al., 2001; Tsao et al., 2006; Moeller et al., 2008). In particular, faces are known to engage holistic processing by which a face is perceived as a whole rather than a combination of its parts (Farah et al., 1998). Previous studies have suggested that the low spatial frequency component of faces plays an important role in characterizing face holistic processing (Goffaux et al., 2003; Goffaux and Rossion, 2006; Harel and Bentin, 2009). This collective body of literature raises the possibility that CNN training with a sequence of blurry to clear faces was successful because face recognition favors low spatial frequencies; as a consequence, this might not necessarily be the case for general object categories.

To evaluate these issues, we trained AlexNet on 1,000 ImageNet object categories using a sequence of blurry to clear images and evaluated its performance by comparing it to a control version of AlexNet trained with only clear images. Contrary to the findings reported by Vogelsang et al. (2018), we failed to observe any beneficial effect of training with blurry to clear object images. By probing each sequence of training, we observed that initial training with blurry objects led to good performance on blurry test objects, but this improvement quickly diminished after training with progressively clearer images. Furthermore, the distribution of spatial frequency preferences of the object-trained CNN rapidly shifted from low to high spatial frequencies, whereas the face-trained CNN maintained a preference for lower spatial frequencies across subsequent stages of training. We conducted two control analyses by matching the power spectrum of face and object training images and by training and testing CNNs on object categorization at subordinate as well as superordinate levels. Nevertheless, we still failed to find any significant benefit of sequential training with blurry to clear objects. We should note that our results do not necessarily contradict the idea that a developmental sequence of blurry to clear visual experience may be beneficial, since we observed a clear benefit in face processing. Rather, our findings suggest that object recognition favors the processing of fine-detailed information of object features, while faces are more processed in a holistic manner by which low spatial frequencies are sufficient for recognition.

4.2 Materials and methods

Visual stimuli

Face images were collected from the FaceScrub database (Ng & Winkler, 2014), which consisted of 100,000 face images sampled from 530 celebrities. The dataset only provided the URLs to the images, and if image URLs were invalid (as of October 13, 2019), those images were excluded. We also excluded any face identities with fewer than 100 examples. This resulted in a final face image dataset with 395 face identities to train CNNs on face recognition. Object images were obtained from the ILSVRC-2012 or ImageNet database (Russakovsky et al., 2015), which has 1,000 object categories with roughly 1.25 million training and validation images. All 1,000 object categories were used to train the object-trained CNNs. All stimuli were converted to grayscale and resized to 224 x 224 pixels to meet the image processing requirements for CNNs.

For our behavioral face recognition task, we chose 10 celebrities (5 females and 5 males) who we considered likely to be well known to the general public: Jennifer Aniston, Mila Kunis, Ellen Degeneres, Selena Gomez, Anne Hathaway, Jim Carrey, Matt Damon, Robert Downey Jr., Ryan Gosling, and Samuel L. Jackson. One of the authors reviewed and sorted out mislabeled or idiosyncratic photographs of faces. Furthermore, we excluded any images that shared a pixel-wise correlation exceeding 0.9 with any other image. The final face image set consisted of 80 images per celebrity or 800 images in total. Regarding image variability, the face images of a given celebrity could vary to a considerable degree due to variations in lighting, viewpoint (ranging from front to three-quarter view), facial expression, hairstyle, make-up, facial hair, age and/or accessories worn (e.g., glasses, hat). We applied a Gabor wavelet pyramid model with 5 spatial scales and 8 orientations to calculate the Pearson correlational similarity of simulated complex cell responses to the images. Normalization was first applied to the all responses at a given spatial scale to control for greater power at lower spatial frequencies. The pairwise correlational similarity of face images was somewhat greater for within-celebrity comparisons (mean r = 0.464, sd = 0.141) than between celebrities (mean r = 0.405, sd = 0.122).

For the behavioral object recognition task, 16 object categories were selected to compare human and CNN performance: bear, bison, elephant, hamster, hare, lion, owl, tabby cat, airliner, couch, jeep, schooner, speedboat, sports car, table lamp, and teapot. Half of the object stimuli were animate and the other half were inanimate. Fifty images per category from the ImageNet validation dataset were used, and thus we had 800 images in total. We performed the same Gabor wavelet pyramid model analysis to the object images. The correlational similarity of the object images was somewhat greater for within-category comparisons (mean r = 0.292, sd = 0.159) than between category (mean r = 0.255, sd = 0.148). As expected, the object images were more heterogeneous than the face images, and within-category (or within-identity) images shared somewhat greater low-level similarity than between-category images.

To generate the blurred images, we applied a Gaussian kernel to each image, adjusting the standard deviation (σ) of the Gaussian function to attain different levels of blur. All image processing was performed using MATLAB. For both behavioral experiments, all images were upsampled by a factor of 2 for presentation on a CRT monitor at a size of 19 × 19 degrees of visual angle.

Participants

We recruited 20 participants to take part in the behavioral object recognition study. A separate group of 20 participants were recruited to take part in the face recognition study. Each of the two studies required approximately 1 hour to complete. All participants reported having normal or corrected-to-normal visual acuity and provided informed written consent. The study was approved by the Institutional Review Board of Vanderbilt University. Participants were compensated monetarily or through course credit.

Behavioral experiments

We measured the abilities of human observers at recognizing faces and objects presented with varying degrees of blur (σ = 0, 1, 2, 4, 8, 12, 16, 20, 24, and 32). Here, σ = 0 indicated clear images without any blurring. Eight face images per celebrity were assigned to each blur level for the face recognition task, while five images per object category were assigned to each blur level for the object recognition task. Both experiments consisted of a total of 800 images, with 80 images presented at each blur level. Image assignment across blur levels was counterbalanced across participants and the order of image presentation was randomized.

Each visual stimulus was briefly presented for 200 ms on a gray background, subtending a visual angle of 19°. After stimulus presentation, participants were asked to report the face identity or object category by entering in a corresponding number code on a numerical pad. The correspondence between the number code and the stimulus identity remained on the screen throughout the study. The mapping between number codes and stimulus identity was counterbalanced across participants. The experiment required approximately 1 hour to complete, including informed consent, instructions, and debriefing. The experiment was implemented using MATLAB and the Psychophysics Toolbox (http://psychtoolbox.org/).

Training of convolutional neural networks

The majority of all CNN experiments and analyses were performed using AlexNet, which can achieve a high level of classification performance while still being quite fast to train from scratch (Krizhevsky et al., 2012). We performed supplementary analyses using VGG-19, which is a deeper CNN with greater learning capacity (Simonyan & Zisserman, 2014).

With the face dataset of 395 celebrities, we divided the images into separate training and validation sets using an approximately 90/10 split. On average, this led to 117 examples per identity for training and 13 examples per identity for validation. For object images obtained from ImageNet, we used their training images (~1.2 million) for training and their validation dataset (50k images) for testing the CNNs. For data augmentation, the training images were randomly rotated from -10º to +10º and about half were flipped about the vertical axis. Across all images within a training set, we calculated the mean and standard deviation of the pixel intensity values and used these values to normalize the pixel intensities of the images.

The models were trained using stochastic gradient descent with a fixed learning rate of 0.01, momentum of 0.9, and weight decay of 0.0001. To train the network initially with blurred images, we applied a Gaussian kernel with the standard deviation of σ = 8, and subsequently reduced the blur level to 4, 2, 1, and 0. The blur level was changed every 100 training epochs for the face recognition task and every 10 training epochs for the object recognition task. Given that the number of training examples per category of ILSVRC-2012 was approximately 10 times larger than that of FaceScrub, the networks in both tasks processed similar numbers of training images per category for each blur level. For comparison, a control CNN was trained with only clear images using the same number of training epochs. All training procedures were implemented in PyTorch on a workstation equipped with multiple GPUs.

Receptive field analysis

We fitted a 2D elliptical Gabor model to the first-layer receptive fields (11 × 11 pixels) of the trained CNNs. The function we used obtained the best fitting model after sampling from 100 different starting points using a gradient descent method. The filters whose R-squared value was less than 0.4 were excluded from analysis. After fitting, the average of standard deviation values of the 2D Gaussian envelope was determined as the size of the receptive field.

Peak spatial frequency in tuning curves

To estimate the peak spatial frequencies of the network across layers, we devised a method in which we presented grating patterns to CNNs and examined the responses of feature maps across layers. Specifically, the gratings were created by sinusoidal patterns using 15 orientations (0, 12, …, 168 in degree), 25 spatial frequencies (4.48, 8.96, ..., 112 cycles/stimulus), and 4 phases (0, 45, 90, 135 in degree). We measured the average responses to the gratings from individual feature maps across convolutional layers and plotted the tuning curves for spatial frequency by averaging across orientations and phases. Each tuning curve was normalized to a range from 0 to 1. The peak spatial frequency was determined from each tuning curve as it yielded the maximum. This peak spatial frequency indicated which spatial frequency was mostly preferred by each feature map of the network.

Spatial frequency control images

As a supplementary analysis, we manipulated the spatial frequency content of the training object images in two ways. First, we calculated the average amplitude spectrum of all training face images and replaced the amplitude spectrum of individual training object images with the average amplitude spectrum from the faces. This was done by performing the fast Fourier transform on each object image in MATLAB, adjusting the amplitude spectrum accordingly, and then performing the inverse fast Fourier transform to reconstruct the amplitude-matched object image. Our second approach relied on low-pass filtering applied to training object images using a cutoff frequency of either 32 or 16 cycles per image. This was done by zeroing out all amplitude values below the cutoff frequency in the Fourier domain, and then performing the inverse fast Fourier transform to reconstruct the image.

4.3 Results

We first compared the performance of human observers and CNNs in face and object recognition tasks across a range of blur. Faces were collected from 10 celebrities of the FaceScrub database (Ng & Winkler, 2014) and objects were collected from 16 categories of the ImageNet database (Russakovsky et al., 2015). To measure CNN performance, a separate AlexNet model was trained on either faces or objects with clear grayscale images and tested on a novel image set. Visual stimuli were blurred by a Gaussian kernel with its standard deviation varied across σ = 0 to 32. Consistent with the previous reports (Dodge and Karam, 2017; Geirhos et al., 2018), we found that human observers outperformed CNNs when images were blurred in both tasks (i.e., gray curves in **Figures 15a-b**). At modest levels of blur, CNN performance quickly dropped to almost chance level, whereas human observers showed more reliable performance across variations in blur. We also observed that both human observers and CNNs were better at recognizing blurry faces than blurry objects.

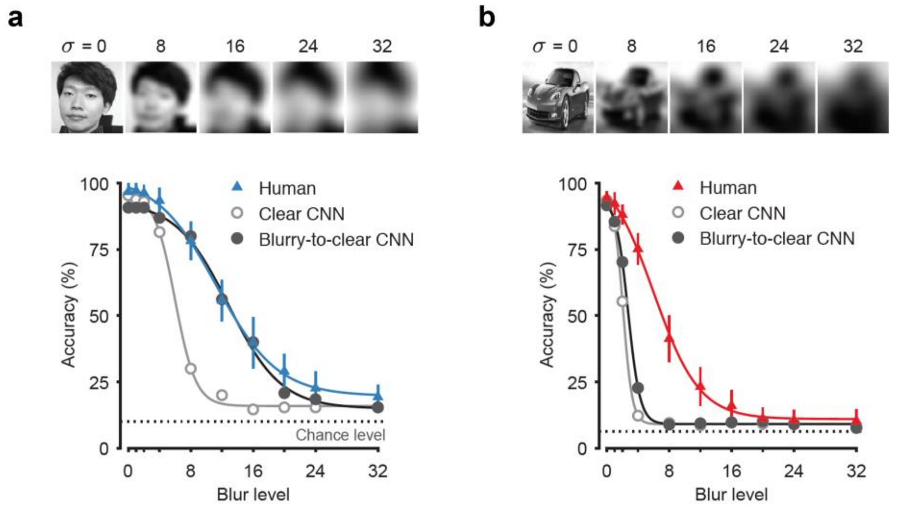

Figure 15. a Face recognition accuracy of human observers (blue triangles), AlexNet trained on clear face images (gray open circles), and AlexNet trained on a sequence of blurry to clear faces (black filled circles) when tested with a wide range of blur levels (chance level performance with a dashed line, 1/10 or 10%). The solid lines represent a logistic function fitted to the data. Images of one of the authors are shown (with permission) for illustrative purposes. **b** Object recognition accuracy of human observers (red triangles), AlexNet trained on clear objects (gray open circles), and AlexNet trained on a sequence of blurry to clear objects (black filled circles) when tested with a wide range of blur levels (chance level performance, 1/16 or 6.25%). Error bars indicate ±1 standard error of the mean.

Our primary interest was whether a sequence of blurry to clear image training would lead to improvement in robustness to blur in both face and object recognition tasks. To this end, two separate AlexNet models were trained on faces and objects initially with blurry images followed by progressively clearer ones across successive training stages (σ = 8, 4, 2, 1, and 0). We found a striking difference in the achievement of robustness between the face- and object-trained CNNs. Similar to the previous finding by Vogelsang et al. (2018), the face-trained CNN showed significant enhancement in robustness and achieved human-level performance in the accuracy by blur curve (**Figure 15a**). In stark contrast, the effect of training in the object-trained CNN was barely noticeable (**Figure 15b**).

We sought to determine the cause of the difference between face and object recognition tasks by probing the performance of these CNNs after each stage of training. **Figure 16a** shows the performance changes of the clear face-trained (gray curve) and blurry to clear face-trained (blue curve) CNNs across successive stages of training. Training with a blur level of 8 (leftmost plot) allowed the CNN to achieve highly robust performance at blurry conditions. This robustness was well preserved after successive clearer images were trained, and at the end of all training stages, the CNN showed stable performance across the full tested range of blur levels. By comparison, the CNN trained on a sequence of blurry to clear objects (red curve) exhibited a different pattern in performance changes across training stages (**Figure 16b**). Similar to the face-trained CNN, training with a blur level of 8 initially improved the performance at σ = 8; however, this enhanced performance soon disappeared with following clearer training images. Strikingly, one epoch was enough to lose the robustness to the previously trained blur level between the transition of two training stages. In the end, the blurry to clear object-trained CNN showed a negligible difference in performance from the clear object-trained CNN. We also observed the same pattern of results in top5 accuracy (bottom in **Figure 16b**), indicating that the failure of observing any training effects in the object-trained CNN was not due to the low sensitivity of top1 accuracy.

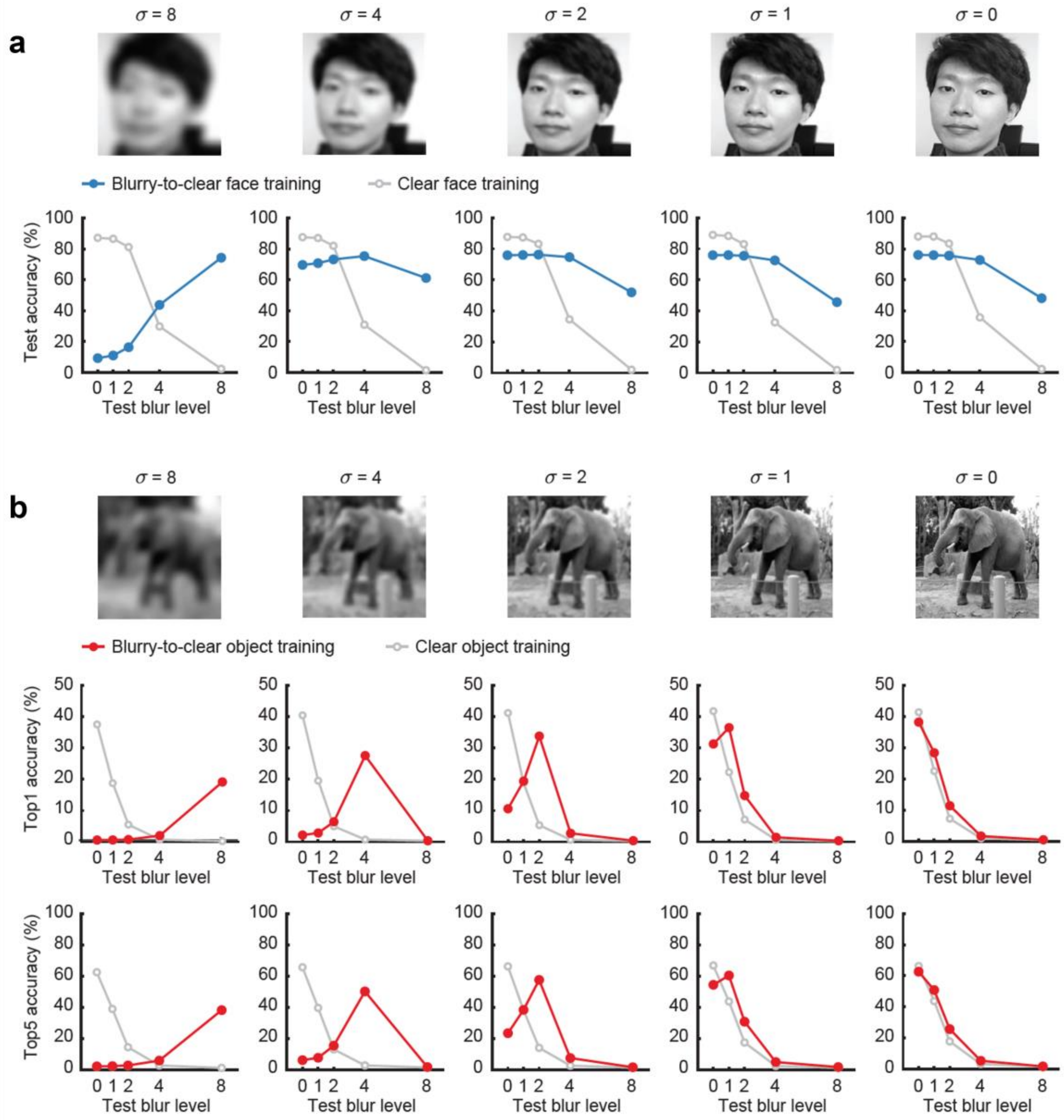

Figure 16. Performance accuracy of AlexNet after training with faces (**a**) or objects (**b**) presented at different blur levels (σ = 8, 4, 2, 1 or 0) over a series of training stages. Gray curves indicate the performance of control CNNs trained on clear images only. To illustrate the amount of blur applied at each training stage, images of one of the authors are shown (with permission) at each blur level.

By monitoring the performance accuracy of training images, the difference between face and object recognition tasks became further clear (**Figure 17**). The face-trained network reached nearly ceiling performance at the first stage of training (σ = 8) and was fine-tuned thereafter by a small margin with clearer face images. In contrast, the training accuracy of the object-trained network did not show early convergence and kept increasing as clearer object images were followed. These results imply that face recognition can be readily resolved by the low spatial frequency content of faces, while object recognition can not.

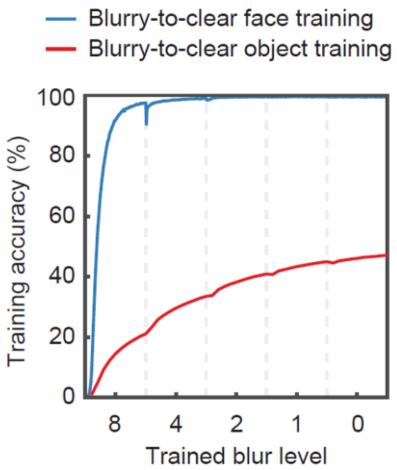

Figure 17. Performance accuracy on training images for CNNs trained on a progression of blurry to clear faces (blue) or objects (red). Vertical dashed lines indicate the transitions between blur levels.

Besides the performance accuracy, we examined how the internal representations of the blurry to clear face- and object-trained CNNs differed by visualizing the first layer receptive fields of the networks (**Figure 18a**). The face-trained CNN exhibited larger receptive fields than the object-trained CNN and they maintained stable sizes across successive stages of training. By comparison, while the object-trained CNN appeared to have large receptive fields at the first stage of training, their sizes became shrinking across training stages, indicating the progressive shift towards preferring sharper and more fine-grained features. This was further quantified by fitting a 2d Gabor model to each of the first layer receptive fields and estimating the average of the standard deviations in the 2d Gaussian profile (**Figure 18b**). The receptive field size of the blurry to clear face-trained CNNs was consistently larger than that of the clear face-trained CNNs across all training stages. In object training, the differences between the receptive field sizes of the clear object-trained and the blurry to clear object-trained CNNs decreased over training and eventually fell to nearly zero, consistent with the pattern of performance accuracy.

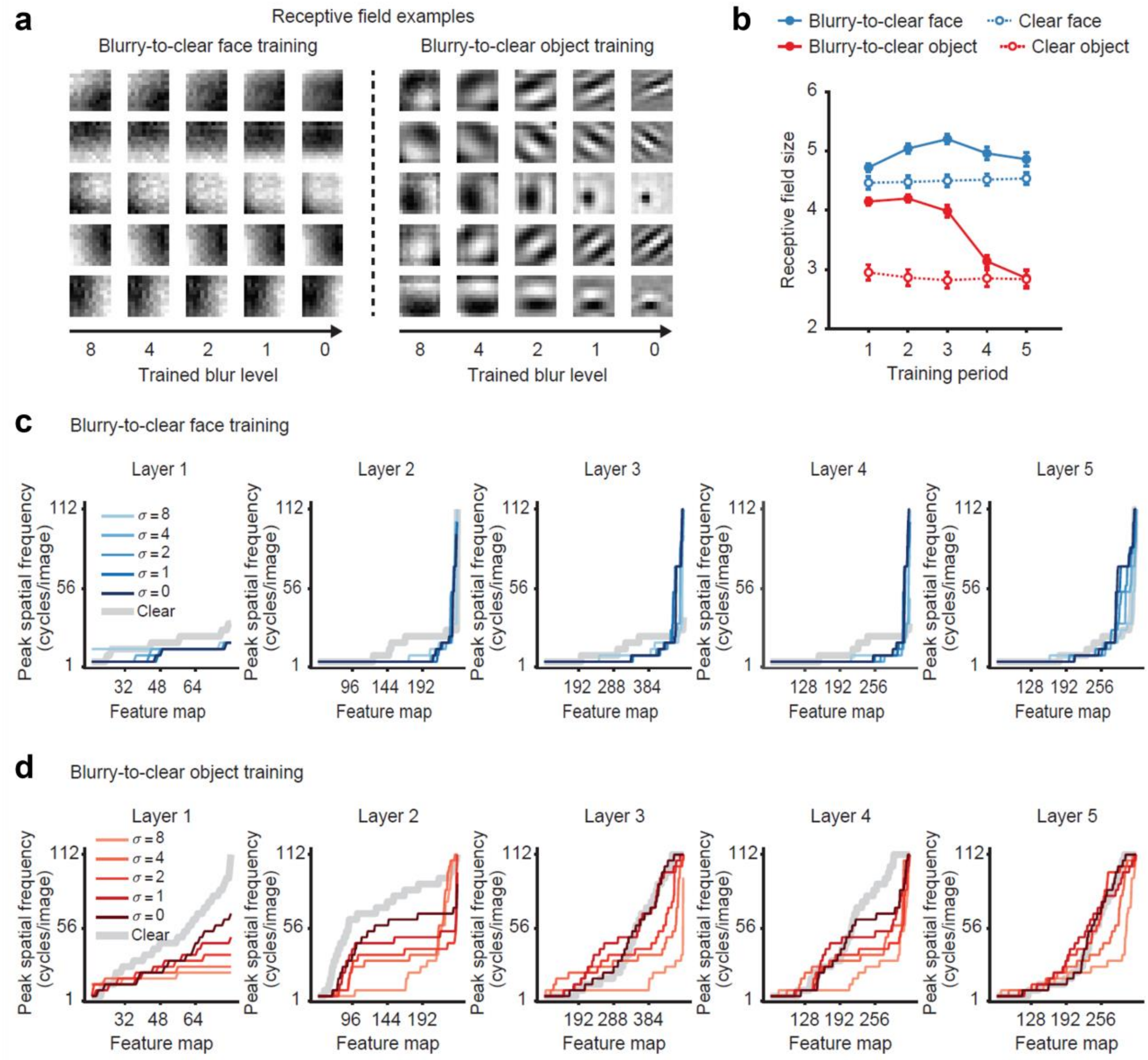

Figure 18. a Examples of the learned receptive fields obtained from a CNN trained on a sequence of blurry to clear faces (left) or blurry to clear objects (right). **b** Receptive field sizes were measured after each training period for both blurry to clear trained CNNs and CNNs trained on clear images only. **c** Peak spatial frequency preferences of face-trained CNNs across successive stages of training, with separate plots shown for each convolutional layer. For visualization purposes, the feature maps are sorted by their peak spatial frequency preference. The gray lines indicate the peak spatial frequency preferences of the clear face-trained network to serve as a reference. **d** Peak spatial frequency preferences of CNNs trained on object recognition, following the conventions of **c**.

To examine the effect of blurry to clear image training beyond the first layer of the networks, we estimated the spatial frequency tuning curves of individual convolutional units that responded to grating patterns across the first 5 layers, and then sorted the peaks of the tuning curves from low to high (**Figure 18c-d**). The distributions of the peak spatial frequencies at different training stages were depicted by different colors in each plot. As shown in **Figure 18c**, while the peak spatial frequencies of the blurry to clear face-trained CNN significantly changed between the first and second stages of training (σ = 8 to 4, $p < 0.01$ in all layers; Mann-Whitney U test), they remained stable thereafter. In particular, a marked difference between the clear face-trained and blurry to clear face-trained CNNs was observed in the range of 1-30 cycles/image, probably accounting for the robust performance of the blurry to clear face-trained CNNs to highly blurred faces. By contrast, the peak spatial frequencies of the blurry to clear object-trained CNN underwent dramatic changes across training stages, particularly in layer 2 ($p < 0.001$ in all stages; Mann-Whitney U test), demonstrating an upward trend in the curve that signified a shift in preference toward higher spatial frequencies (**Figure 18d**). Collectively, these findings clearly reveal the difference between face and object recognition in spatial frequency and are reminiscent of the hallmark feature of face processing that faces are processed in a holistic manner.

One could ask whether the difference in spatial frequency between face and object recognition tasks might be due to the fact that object images likely contain more high spatial frequencies than face images. This is not necessarily equivalent to the notion that object recognition needs more high spatial frequency than face recognition. That is, it is possible that the failure of a sequence of blurry to clear training in objects was simply attributed to the different image statistics in faces and objects. Indeed, we observed greater power at higher spatial frequencies in the training images of objects than faces (**Figure 19a**). To further differentiate between those two effects, we conducted a control analysis by adjusting the power spectrum of the training object images to match the average power spectrum of the training face images (examples are shown in **Figure 19b**; second column). Then, we trained those object images with a sequence of blurry to clear training. We found that the pattern of results was almost identical to those observed previously (**Figure 19c**), except that it caused a small loss of accuracy at clear test images after all training was completed. We further constructed two training image sets where object images were low-pass filtered by either a threshold of 32 or 16 cycles per image (third and fourth columns in **Figure 19b**). As the network trained with the more severe threshold of 16 cycles per image, it showed somewhat greater accuracies at modest levels of blur ($\sigma = 4$ or 8 in **Figure 19d**). However, this improvement was accompanied by a greater loss of accuracy at clear test objects. Taken together, these control analyses indicate that the different patterns in performance to acquire robustness from a sequence of blurry to clear training between face and object recognition cannot be explained simply by their image statistics.

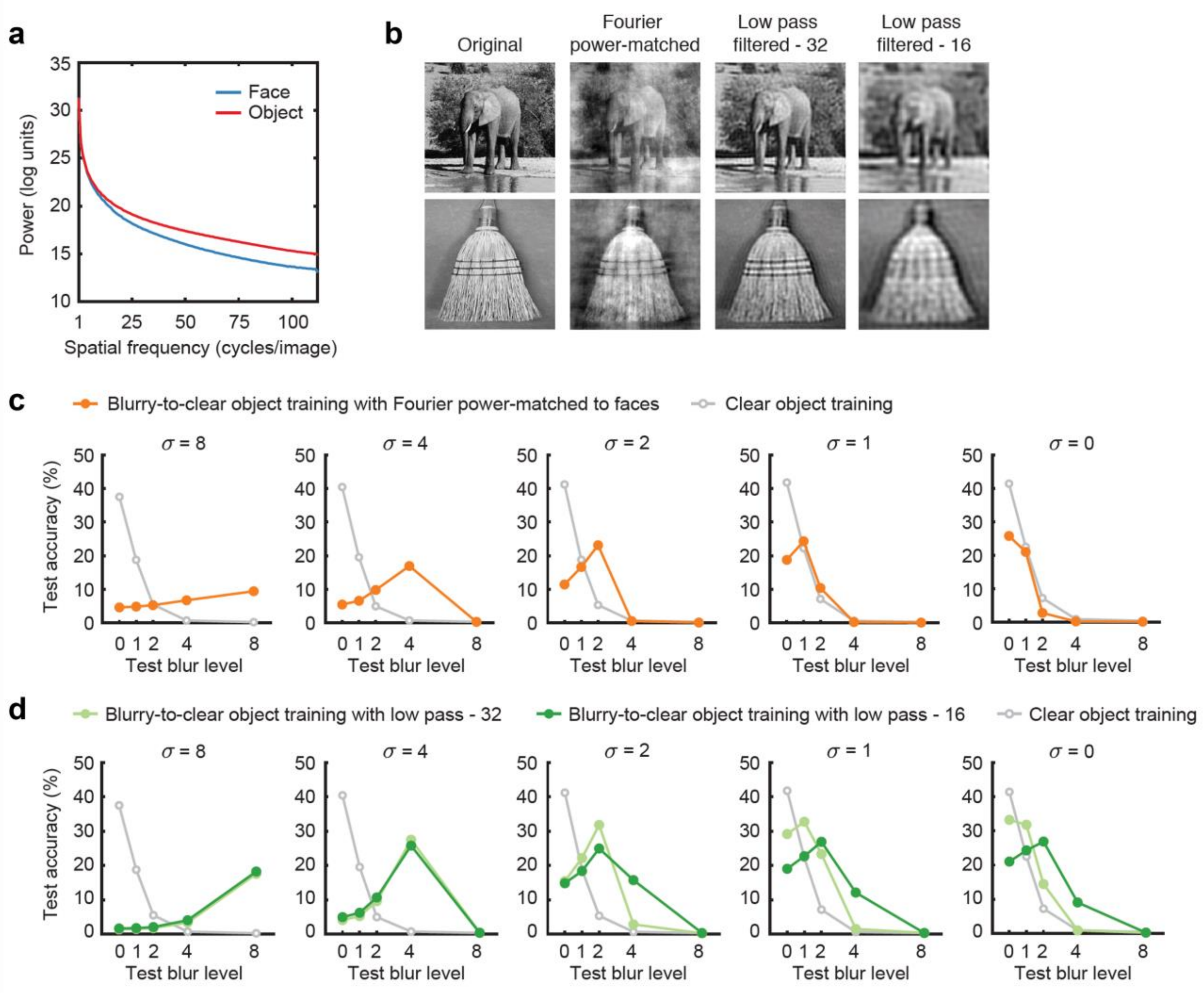

Figure 19. a Average power spectrum of face and object training images plotted on a log scale (ordinate) as a linear function of spatial frequency. **b** Examples of original object images (first column), object images with Fourier power spectrum matched to the average power spectrum of the face training dataset (second column), and low-pass filtered images with a cut-off frequency of either 32 cycle/images (third column) or 16 cycle/images (fourth column). **c** Recognition accuracy for a CNN trained on blurry to clear objects, after objects were first matched to the Fourier power spectrum of faces (orange curve). For comparison, performance of the CNN exclusively trained on the original clear objects is also shown (gray). **d** Object recognition accuracy of CNNs trained on blurry to clear objects, after the objects were low-pass filtered with a cut-off frequency of 32 cycles per stimulus (light green) or 16 cycles per stimulus (dark green). Again, the CNN originally trained on clear objects is shown in gray.

Additionally, one may wonder whether different categorization levels of objects would have different impacts on the degree of robustness achieved from blurry to clear training. For example, compared to faces that share the same configuration across different identities, ImageNet object categories greatly vary in their shapes and textures. This raises the possibility that the greater variation in objects might obscure the effect of blurry to clear image training. On the other hand, the developmental literature has demonstrated that infants formulate categorization abilities in a progressive manner, initially with superordinate-level categories followed by basic and subordinate levels (Mandler and McDonough, 1993; Quinn, 2004). Accordingly, it is also plausible to think that infants may rather benefit from the developmental sequence of blurry to clear inputs at the superordinate level of categorization.

To address these concerns, we first created two subsets of ImageNet categories including 116 dogs and 52 birds, and trained a separate AlexNet model for either dog breed or bird species categorization task with the sequence of blurry to clear training. Similar to the previous procedure, we compared these networks to the control networks trained by clear images only. As shown in **Figure 20a**, we still failed to observe a significant improvement in robustness in both tasks. There was a marginal improvement at σ of 2 in the dog breed categorization task, though this became almost invisible at higher blur levels. We also created another image set where most of ImageNet categories were sorted into two superordinate-level categories, i.e., 407 animate and 522 inanimate, by leveraging the WordNet hierarchy. Likewise, we compared two versions of the network trained for an animate/inanimate discrimination task, one trained with clear images and the other trained with blurry to clear images. Again, the performance of the blurry to clear object-trained network did not significantly differ from that of the clear object-trained network (**Figure 20b**). Therefore, we concluded that, regardless of the categorization levels of objects, fine-detailed features are necessary to achieve optimal performance for object recognition.

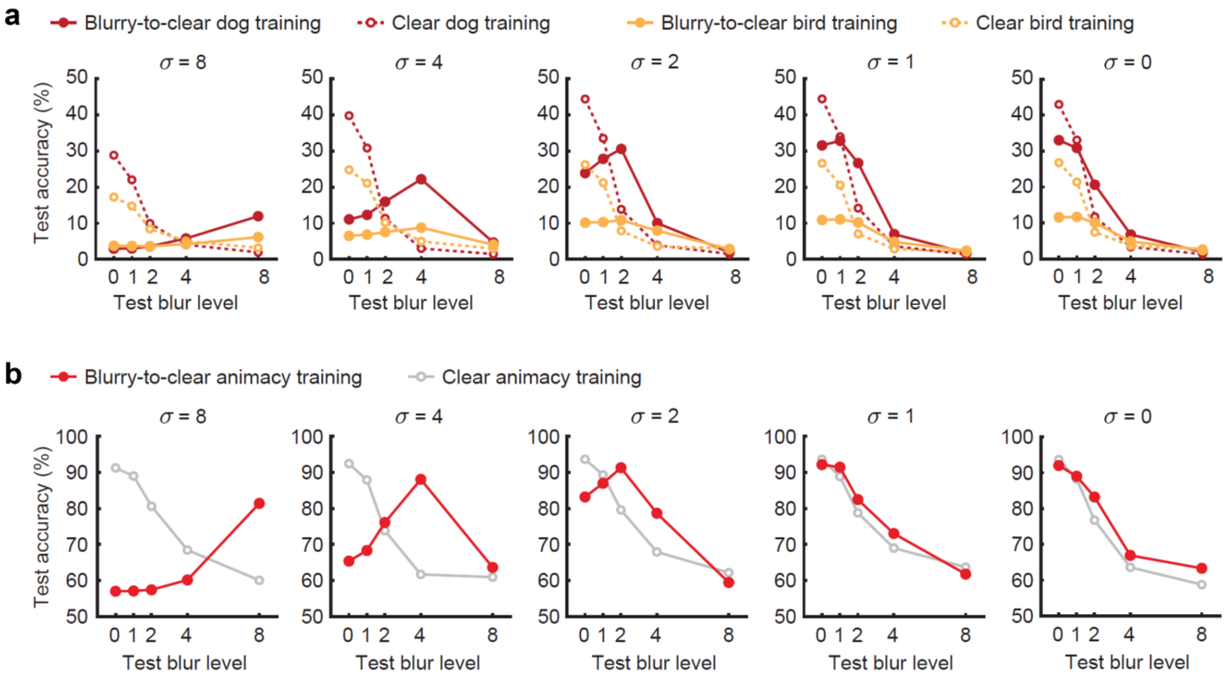

Figure 20. a Comparison of CNNs trained on blurry-to-clear (solid) or clear (dashed) images of different dog breeds (red) or bird species (orange). **b** Comparison of CNNs trained on blurry-to-clear (red) or clear (gray) images of ImageNet objects to perform an animate/inanimate discrimination task.

4.4 Discussion

CNNs are known to have difficulty recognizing blurred images of objects (Dodge and Karam, 2017; Geirhos et al., 2018). In the present study, we evaluated the hypothesis that a developmental sequence of blurry to clear image training might lead to more robust object recognition, inspired by clinical literature suggesting that such a sequence of training may be beneficial for developing integrative feature-binding processes (Le Grand et al., 2001; Geldart et al., 2002). We found that a CNN trained with a developmental sequence of blurry to clear faces showed stronger robustness to blur than a CNN trained with only clear faces, consistent with a previous report by Vogelsang et al. (2018). However, this benefit was not observed in object-trained CNNs. This result does not necessarily contradict the notion that initial experience with blurry vision may be critical for developing reliable face recognition systems in adults. Rather, our findings question the idea that initial blurry vision is ecologically important for developing blur-robust object recognition systems.

Although we failed to find evidence that early experiences with blur can lead to robust object recognition, the current findings illuminate fundamental differences between face and object processing with respect to their underlying preferences for spatial frequency. We observed that CNNs trained with blurry to clear objects demonstrated progressive changes in their spatial frequency preferences to favor higher spatial frequencies, whereas CNNs trained with blurry to clear faces showed stable preferences for lower spatial frequencies. Particularly interesting was the finding that the low spatial frequency content of faces was sufficient to achieve very high and stable performance for training stimuli across a range of blur levels (**Figure 17**). Such robustness was observed despite the inherent challenge of the face recognition task, which entailed far less between-identity variability than the object recognition task. Our findings provide a novel line of support for the notion that faces are processed in a holistic manner, as it has been suggested by many behavioral and neuroimaging studies (Farah et al., 1998; Goffaux and Rossion, 2006; Liu et al., 2010). In addition, our findings are concordant with other computational studies on face holistic processing. For example, Tan and Poggio (2016) proposed that three characteristic markers of face holistic processing, including the composite face effect, face inversion effect, and whole-part effect, can be accounted for by larger receptive field size in an HMAX model, consistent with our own findings (see **Figure 18a-b)**. A recent study has also suggested that both humans and CNNs tend to rely on lower spatial frequencies for face recognition tasks (Song et al., 2021). These studies including ours indicate the potential of CNNs to study face processing; that said, it is not yet clear whether CNNs trained on a face dataset process faces in a manner similar to human observers. Only in recent years has face recognition become tractable alongside the rise of CNNs (e.g., Taigman et al., 2015; Parkhi et al., 2015; Schroff et al., 2015; O'Toole et al., 2018), this should be further explored in future studies.

Our results suggest that initial visual experience with blurry inputs is not sufficient to account for the blur-robust nature of object recognition in humans. An alternative possibility is that repeated experiences with blurry visual inputs throughout life may be necessary to maintain robustness to blur. This may well be true if we consider that blurry objects are repeatedly encountered in peripheral vision, moreover, vision in the fovea can also be blurry if it is out of focus prior to accommodative compensation (Strasburger et al., 2011; Sprague et al., 2016). This repeated experience of blurry vision may explain some of the discrepancies between humans and CNNs that have been reported in recent years. This will be discussed later in detail in Chapter 5.

Finally, our computational approach could provide a promising opportunity to study the developmental trajectory of visual learning. This approach enables us to avoid a fundamental limitation of infant studies, as one cannot interfere with the developmental visual experiences of infants in an appropriate or ethical manner, whereas the training experiences of CNNs are readily manipulatable. Indeed, there are a few recent studies that have leveraged CNNs to examine issues of visual development, similar to the work done here (Voglesang et al., 2018; Bambach et al., 2018). However, a critical issue that remains to be established is that CNNs may use a different learning algorithm from humans and thereby do not follow the developmental trajectory of humans. For example, CNNs are known to suffer greatly from catastrophic forgetting that refers to the phenomenon that machine learning models abruptly forget what has been previously learned as new information comes in (Goodfellow et al., 2013; Li and Hoiem, 2017; Kirkpatrick et al., 2017). By comparison, humans do not simply forget what they have learned before, but instead, they effectively utilize prior information when learning new skills or knowledge. Such findings may potentially suggest that humans and CNNs rely on different principles for learning (e.g., supervised vs. unsupervised). On the other hand, it has been also suggested that deep learning models may capture some characteristics of critical period learning that are observed in humans and animals (Achille et al., 2018). In future studies, it will be important to be mindful of potential differences between humans and CNNs in their developmental trajectories and to clarify the validity of CNNs for developmental research.

Acknowledgments

This work has been recently accepted for publication in the Journal of Vision in July 2021. Hojin Jang and Frank Tong are co-authors on this project.

5. Repeated visual experiences of blurry objects may be beneficial to the development of robust object recognition systems

5.1 Introduction

People often think that blurry vision should be corrected. The main causes of blurry vision are refractive errors such as myopia, hyperopia, and astigmatism. They occur when the eye does not correctly focus light onto the retina and are often corrected with eyeglasses, contact lenses, or laser surgery. However, in addition to the refractive errors, we also experience blurry vision quite often more than we sense in our daily life. When we pay attention to a portion of the visual field, out-of-focus objects become blurry due to defocus aberration (Sprague et al., 2016). If they are somewhere at the far periphery, they will look even blurrier (Anstis, 1974; Strasburger et al., 2011). Despite the fact that many objects present in our visual field appear blurry, we do not have any problem recognizing them in our daily vision tasks.

Though high visual acuity is critical to performing our daily vision tasks, the notion that blurry vision needs to be corrected sometimes led us to overlook the remarkable ability of individuals to recognize blurry objects and the potential significance of blurry vision in object recognition. What if blurry vision is actually advantageous to our daily vision? What if blurry vision is ecologically important for maintaining our recognition system to be stable and robust?

Previous research has suggested that humans may leverage blurry vision to assist in our object recognition processes. For example, binocular disparity is useful for depth perception, but may not solely explain it. Instead, blur could provide a complimentary cue for more precise depth perception (Marshall et al., 1996; Held et al., 2012). Real-world objects usually occur within particular background scenes. The contextual scene information that often looks blurry in the visual field can assist our object recognition behavior (Torralba, 2003; Oliva and Torralba, 2007). A recent neurophysiological study has reported blur-selective neurons in V4, the responses of which were modulated by the degree of blur besides stimulus shape, size, contrast, and curvature (Oleskiw et al., 2018), suggesting that blur may be a fundamental feature in vision.

This may bring about a critical change in our view of convolutional neural networks (CNNs). The visual world that CNNs experience may be more different from ours than we ever thought. Typically, a dataset on which CNNs are trained primarily consists of high resolution and clear images, e.g., ImageNet (Russakovsky et al., 2015). Unless strong data augmentation is specified, CNNs may view the world as we do only using foveal vision. Thus, CNNs may lack the opportunity to experience multiple representations of objects from clear to blurry ones. This may create an unexpected bias in CNNs while developing their recognition system, for example, leading to an over-reliance on high spatial frequency components of objects for recognition. If this is truly an issue, would training CNNs with a mixture of clear and blurry images lead to provide a more predictable model for human object recognition behavior?

By training CNNs on both clear and blurry images, we made the following specific predictions. The CNNs would show a stronger shape bias in recognition than other CNNs typically trained on clear images only. Previous studies have reported that CNNs tend to make a shape-agnostic but strong texture-dependent decision when recognizing objects (Ballester and Araujo, 2016; Baker et al., 2019; Geirhos et al., 2019). This was demonstrated by artificially generated stimuli that were synthesized from the texture of an object and the shape of the other object (e.g., cat shape with elephant texture; Geirhos et al., 2019). We expected that blur training would mitigate this texture bias of CNNs by making them less resort to fine-grained features of objects in recognition.

In addition, blur training would make CNNs more robust to various types of degraded conditions. Recent evidence has indicated that CNNs struggle to identify objects when visual inputs are perturbed even minimally, whereas humans are generally robust to a variety of visual noise (Dodge and Karam, 2017; Geirhos et al., 2018; Jang and Tong, 2018). We have previously shown the opposite accuracy pattern of human observers and CNNs, that is, CNNs were easily impaired by pixelated Gaussian noise, but human observers were more disrupted by Fourier phase-scrambled noise. Specifically, we expected that blur training would improve the robustness of CNNs in favor of the response patterns of human observers, which means, there would be a greater improvement in pixelated Gaussian noise than Fourier phase-scrambled noise.

Finally, we expected that the internal representations of the CNNs trained with both clear and blurry objects would be better aligned to those obtained from humans or primates. Many studies have argued that CNNs do not fully account for neural representations of objects in humans and non-human primates (Kar et al., 2019; Bashivan et al., 2019; Xu and Vaziri-Pashkam, 2021), nor do CNNs predict image-by-image behavioral patterns of those (Rajalingham et al., 2018). If humans (or primates) develop their visual systems relying on blurry vision to some extent, this might be revealed by the measurements of brain-CNN correspondences.

In the present study, we hypothesized that CNNs experiencing both clear and blurry objects would provide a better predictive model for human vision. To evaluate this hypothesis, we compared two versions of CNNs, one trained on clear ImageNet objects and the other trained on a mixture of clear and blurry ImageNet objects. We found that the CNNs trained with a mixture of clear and blurry images showed not only enhanced robustness to blur but also higher correlations with cortical representations in humans under blurry viewing conditions, as compared to the control CNNs that were only trained with clear images. A unit-level analysis of spatial frequency tuning curves suggested that the early layers of the CNNs may be critical to conferring greater robustness to blur. These CNNs also demonstrated a stronger shape bias and stronger robustness to noise than the control CNNs. Finally, the cortical representations of the early visual areas were better predicted by the CNNs trained with blurry objects. Altogether, our findings may suggest that CNNs typically trained on ImageNet are likely biased toward overrepresenting high spatial frequency representations of objects and that, by comparison, humans may benefit from blurry vision for their robust object recognition.

5.2 Materials and methods

Training of convolutional neural networks

In this study, we sought to compare two versions of CNNs, one exclusively trained with clear images (referred to as clear-trained CNNs) and the other trained with a mixture of clear and blurry images (referred to as blur-trained CNNs). Six CNN models including AlexNet, VGG16, VGG19, GoogLeNet, ResNet50, and Inception-v3 (Krizhevsky et al., 2012; Simonyan et al., 2014; Szegedy et al., 2015; He et al., 2016; Szegedy et al., 2016), were first trained on clear images from the ImageNet 1,000 categories. All input images were resized to 224 × 224 pixels and grayscaled. Images were randomly flipped horizontally and rotated within ±10 degrees. Images were then normalized using the mean and standard deviation of the ImageNet training samples. The networks were trained for 70 epochs using a stochastic gradient descent optimizer with a fixed learning rate of 0.001, momentum of 0.9, and weight decay of 0.0001.

Correspondingly, the same networks were trained on a mixture of clear and blurry images using the same training protocol. Two methods to generate blurry images are described below.

First, we assumed that clear vision at the fovea and blurry vision in the periphery were evenly contributing to shaping the object recognition system. To obtain a realizable approximation, it was assumed that ImageNet examples (224 × 224 pixels) were photographed by a 35 mm camera with a 54º field of view and foveal vision had a normal visual acuity of 20/20 which corresponds to approximately 30 cycles per degree. Then, an observer would resolve 30 cycles per degree / 224 pixels × 54 degrees on the stimulus for foveal vision. According to the Anstis' measurement at retinal eccentricities of up to around 60º, the decline in visual acuity follows $E_2/(E_2+E)$ where E is eccentricity and E_2 is set to a constant value of 2 (Anstis, 1974; Strasburger et al., 2011). With the formula, an observer would resolve (30 × 2/62) cycles per degree / 224 pixels × 54 degrees on the stimulus at 60º eccentricity. By setting the spatial frequency at the fovea to be the Nyquist frequency, we could compute the cut-off frequency f_c at 60º eccentricity. The standard deviation of a Gaussian blur filter (σ) in the pixel domain was then determined as approximately 9.8676 by:

$$f_c = \frac{1}{2\pi\sigma}.$$

Based on this, we ended up choosing 5 blur levels, σ = 0, 1, 2, 4, and 8. During the training procedure, an individual image was randomly blurred by one of the 5 levels (σ = 0 corresponds to original images without blur), with each of the levels having the same probability of being selected.

In addition to the method above, we also considered a more conservative approach to approximate the real-life blurry visual experiences of humans. Here, we assumed that an object that is placed at the center of the visual field would greatly contribute to the recognition process, where the object could appear either clear or blurry depending on an observer's point of focus. The degree of out-of-focus blur could then provide the amount of blur needed to be presented in the training samples of CNNs.

Out of focus blur can be quantified by the diameter of a blur disk computed by:

$$b = p \cdot D,$$

where p is pupil diameter in mm and D is the absolute difference in diopters between the focal distance and the distance at which an object appears. We relied on the data by Sprague et al. (2016) to estimate the degree of out-of-focus blur in the visual field (b). In the study, an observer performed four daily living tasks, i.e., walking outside, walking inside a building, ordering coffee, and making a sandwich, while wearing a head-mounted camera with an eye-tracking device. These measures allowed the authors to estimate the distances from the observer to fixation and to a scene point and obtain the relative distance, D, on every video frame. The average measurement of pupil diameter in the experiment, 5.8 mm, was used for the calculation of blur. The estimated blur disk diameter was converted to radius for convenience. To convert the measuring unit of blur disk radii in degrees of visual angle to pixels, we again assumed that ImageNet examples were viewed with a 54º field of visual angle. After the conversion, the frequencies of all adjacent bins (0-0.5, 0.5-1, …, 4.5-5; the maximum was 4.86 pixels) were counted for each task. Finally, to better account for daily visual experiences, a different weight was multiplied by the frequencies of each task when the four tasks were combined, i.e., 0.16 for outside walk, 0.10 for inside walk, 0.53 for order coffee, and 0.21 for make sandwich, following the method by Sprague et al. (2016). The final distribution of blur disk radius was fitted to an exponential decay function. The frequencies of radii from 0 to 5 were converted to the probabilities of blur levels in training samples: 69.39% for the radius of 0, 21.28% for the radius of 1, 6.53% for the radius of 2, 2% for the radius of 3, 0.61% for the radius of 4, and 0.19% for the radius of 5. The radius of 0 indicated a clear image without any blur. One thing to note is that we ended up using a Gaussian blur kernel rather than a circular blur disk, partly because a circular blur disk cannot account for longitudinal chromatic aberration (Cholewiak et al., 2018) and also because a Gaussian blur kernel is more commonly used in the machine learning literature and thereby more convenient for evaluation.

To further differentiate the two types of blur-trained CNNs, the CNN models trained by the first and second methods were referred to as strong-blur-trained CNNs and weak-blur-trained CNNs, respectively.

Evaluation of CNNs on blurry scenes

To evaluate whether blur-trained CNNs better accounted for human recognition of blurry objects than clear-trained CNNs, we leveraged the behavioral and neuroimaging data that were previously acquired. The behavioral data were collected by 20 human observers where their recognizing abilities were measured using a total of 800 images from 16 ImageNet categories across 10 different blur levels (σ = 0, 1, 2, 4, 8, 12, 16, 20, 24, and 32, as σ of the standard deviation in a Gaussian blur kernel; refer to Chapter 4 for details). Both clear- and blur-trained CNNs were tested on the same images using the same blur levels. Note that these images were never used during training. To measure 16-way classification performance, the softmax outputs of the 16 categories of the networks were compared. In addition to recognition performance, the confusion matrices of the CNNs across 16 categories were obtained and compared to those of human observers.

Neuroimaging data collected by Abdelhack and Kamitani (2018) were obtained for 5 human observers who viewed both clear and blurred stimuli in a 3T fMRI scanner. We only analyzed test image runs that contained stimuli degraded by different blur levels (0%, 6%, 12%, and 25%). A total of 40 unique images were presented with those 4 blur levels. The test image runs consisted of two conditions, prior and no-prior. In the prior condition, observers were provided with semantic information about the categories of test images (airplane, bird, car, cat, and dog) before the experiment. Each category contained 4 different examples. In the no-prior condition, observers did not receive any prior and viewed 20 object images (one example per category). Seven regions of interest (ROIs) were created from a separate retinotopy experiment, including V1-V4 and LOC/FFA/PPA.

To compare the representations of clear- and blur-trained CNNs to the brain responses, we performed the representation similarity analysis (RSA) across four conditions: 1) blur levels of 0% and 6%, 2) blur levels of 0% and 12%, 3) blur levels of 0% and 25%, and 4) all blur levels of 0%, 6%, 12%, and 25%. This RSA matrix was obtained from each layer of the CNNs and each brain region of the human participants. The similarity of the RSA matrices between humans and CNNs was measured by Pearson correlation. For each brain region, the highest correlation was chosen across all layers of individual CNNs, as done by other studies (Schrimpf et al., 2018; Zhuang et al., 2021). The brain and CNN responses were z-normalized before calculating RSA matrices. The diagonals of the RSA matrix were excluded for analysis.

Orientation and spatial frequency tuning curves of convolutional units

To better understand the effect of training with blurry visual inputs, we estimated the responses of individual convolutional units to sinusoidal grating patterns generated by 15 orientations (0, 12, …, 168 in degree), 25 spatial frequencies (4.48, 8.96, ..., 112 cycles/stimulus), and 4 phases (0, 45, 90, 135 in degree), across layers. The responses of convolutional units were averaged across spatial positions. Both orientation and spatial frequency tuning curves were normalized to have a range of 0 to 1. To estimate the peak and bandwidth of the orientation tuning curve, we fitted a von Mises distribution to individual curves as follows (Swindale, 1998):

$$f(\theta) = A \exp\{\kappa[\cos^2(\theta - \mu) - 1]\},$$

where A is the value of the function at $\theta = \mu$, μ is the preferred orientation, and κ is the precision parameter determining the width of the distribution. The bandwidth of the tuning curve was defined as the full width at half maximum calculated below:

$$\mathrm{FWHM} = \cos^{-1}[(\ln 0.5 + \kappa)/\kappa], \text{ where } \kappa > -0.5 \ln 0.5.$$

To obtain the peak and bandwidth of the spatial frequency tuning curve, a Gaussian function was fitted to the curve on a logarithmic scale. Similarly, the full width at half maximum of a Gaussian distribution served as the bandwidth of the spatial frequency tuning curve.

Texture and shape biases

Geirhos et al. (2019) have suggested that ImageNet-trained CNNs tend to recognize objects relying on texture cues, whereas, in stark contrast, humans show a strong bias to shape cues. This was demonstrated by texture-shape cue conflict stimuli generated by style transfer (Gatys et al., 2016), such as the shape of a cat rendered using the texture of elephant skin. The texture-shape cue conflict stimuli consisted of 1,280 images from 16 ImageNet categories that included airplane, bear, bicycle, bird, boat, bottle, car, cat, chair, clock, dog, elephant, keyboard, knife, oven, and truck (available at https://github.com/rgeirhos/texture-vs-shape). The degree of shape bias was determined by calculating the proportion of shape decisions within a total number of texture and shape decisions. Using the texture-shape cue conflict stimuli, the shape biases of clear- and blur-trained CNNs were estimated. Additionally, the degree of shape bias obtained from 10 participants in the original study was also reported as a reference.

Evaluation of CNNs on noisy scenes

The recognition abilities of clear- and blur-trained CNNs under noisy viewing conditions were also evaluated. We first leveraged a benchmark dataset that has garnered increasing interest with respect to the robustness of CNNs, ImageNet-C (Hendrycks and Dietterich, 2019; available at https://github.com/hendrycks/robustness). The dataset contains ImageNet 50,000 validation images, each of which is degraded by 19 types of corruptions, including 4 cases of blur (defocus blur, glass blur, motion blur, and zoom blur), 4 digital (contrast, elastic transform, JPEG compression, and pixelate), 3 noise (Gaussian noise, impulse noise, shot noise), 4 weather (brightness, fog, frost, and snow), and 4 extra categories (Gaussian blur, saturate, spatter, and speckle noise). Each type of corruption has 5 levels of severity. The details of the image corruption methods are described in their paper (Hendrycks and Dietterich, 2019).

We also evaluated the clear- and blur-trained CNNs on noisy images by comparing their performance with human behavioral and fMRI data previously collected in Chapter 2. For the behavioral data, we measured the recognition performance of 20 participants under degraded conditions with pixelated Gaussian noise and Fourier phase-scrambled noise. We also recorded the brain activity of 8 participants while they viewed the two noise types as well as clear images. Following the method by Chapter 2, we compared the internal representations of the clear- and blur-trained CNNs to those of the ventral visual stream using an RSA approach.

Evaluation of CNNs on semantic representations and neural predictivity

To further determine whether blur-trained CNNs provide a more suitable model for object recognition in humans than clear-trained CNNs, two additional datasets were tested. The first is a large-scale fMRI dataset that collected the brain responses to 1870 natural images from 2 observers (Kay et al., 2008). The original goal of the study was to investigate whether complex natural stimuli could be decoded by brain activity. To do so, a Gabor wavelet pyramid model was estimated using 1750 images and tested by 120 images for image identification. For our purposes, we combined the two image sets to one and calculated the 1870 × 1870 RSA matrix. Due to its large scale, the matrix could provide insights into the neural representations of semantic relations across object categories. Using the same stimuli, we obtained the RSA matrices from clear- and blur-trained CNNs and measured their correspondences to those by the 2 human observers.

The other dataset, brainscore, is a public benchmark that has been designed to assess the neural predictivity of computational vision models (Schrimpf et al., 2018). The responses of single neurons are predicted by the internal representations of computational models using a partial least square regression and the quality of prediction is estimated by a median correlation between the actual and predicted responses across multiple cross-validation runs. In our study, we compared the brainscores of clear- and blur-trained CNNs across 4 brain regions including V1, V2, V4, and IT (Freeman et al., 2013; Majaj et al., 2015).

5.3 Results

Six standard CNN models were employed in the study: AlexNet, VGG-16, VGG-19, GoogLeNet, ResNet-50, and Inception-v3 (Krizhevsky et al., 2012; Simonyan et al., 2014; Szegedy et al., 2015; He et al., 2016; Szegedy et al., 2016). All inputs were converted to grayscale and resized to 224 × 224 pixels. Each CNN model was trained on ImageNet from scratch in two different versions, one trained by clear original images and the other trained by a combination of clear and blurry images. Blur was generated by applying a 2D Gaussian filter. In order to mimic the nature of blurry visual inputs in humans, two different methods were evaluated, one trained with higher blur levels than the other (see **Materials and methods**).

We first assessed the recognition abilities of 20 human observers, clear-trained CNNs, and two versions of blur-trained CNNs under blurry viewing conditions. As previously reported in Chapter 4, clear-trained CNNs displayed poorer performance than human observers when inputs were degraded by blur (red versus yellow curves in **Figure 21a**). At a blur level of 8, those CNNs showed almost chance level performance, whereas the recognition performance of human observers was still fairly high. By comparison, both versions of blur-trained CNNs showed enhanced robustness to blur. To be specific, weak-blur-trained CNNs closely matched the humans' performance, while strong-blur-trained CNNs exceeded human-level performance at blur levels of 8 and 12. Interesting is the fact that including a small number of blurry images in training, e.g., σ = 4 and 5 only accounting for 0.8% of the total training images in the weak-blur-trained CNNs, allowed them to achieve approximately 90% of accuracy performance tested at σ = 4 and 40% at σ = 8. The efficacy of weak blur training implies that rare events of blurry visual experiences could have a significant impact on recognition systems. We also examined the similarity of confusion matrices between human observers and each of the CNNs. Both versions of blur-trained CNNs showed greater similarity than clear-trained CNNs at blur levels of 4 and 8 with a rightward shift of the correlation by blur level curve (**Figure 21b**). This was primarily because clear-trained CNNs were easily biased towards certain categories even with minimal blur levels, though both versions of blur-trained CNNs also demonstrated such bias above σ =

8. At σ = 32, strong-blur-trained CNNs showed higher correlation than clear-trained CNNs, likely better capturing the pattern of bias made by human observers (**Figure 21c**).

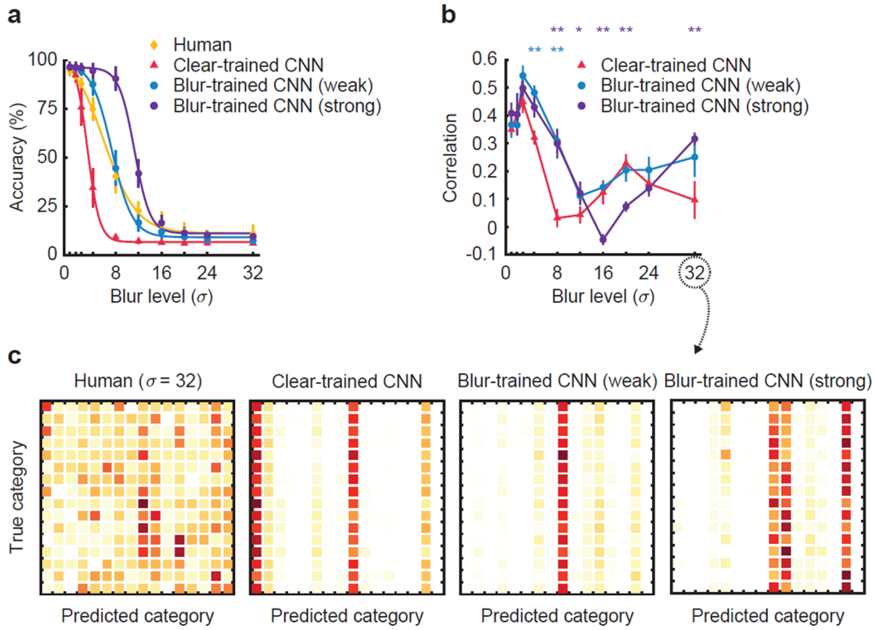

Figure 21. a Accuracy performance on 16 categories tested across a range of blur from σ of 0 to 32 by human observers (yellow), clear-trained CNNs (red), weak-blur-trained CNNs (blue), and strong-blur-trained CNNs (purple). **b** Similarity of confusion matrices between human observers and CNNs. **c** Examples of confusion matrices at σ of 32 from human observers, clear-trained CNNs, weak-blur-trained CNNs, and strong-blur-trained CNNs.

We also examined whether blur-trained CNNs provided more similar response patterns of blurry objects than clear-trained CNNs when compared to the cortical responses of humans. This was evaluated by using previously collected neuroimaging data (Abdelhack and Kamitani, 2018). Five participants were scanned using fMRI while viewing 40 clear images as well as degraded ones by different blur levels (6%, 12%, and 25%). We computed RSA matrices from each brain region of the ventral stream and from each layer of the CNNs using a total of 80 images (i.e., 40 clear images plus 40 blurry ones from a selected blur level). **Figure 22** shows the similarity of the RSA matrices between human observers and CNNs determined by the maximum correlation across the layers of the CNNs for each brain region. When the degree of blur was weak (leftmost plot in **Figure 22**), both versions of blur-trained CNNs exhibited significantly higher correlations with humans than the clear-trained CNNs across all brain regions. The difference between the clear- and blur-trained CNNs became reduced when the degree of blur was strong (i.e., a blur level of 25%), although the blur-trained CNNs still showed higher correlations in the early visual areas, V1-V4. The strong-blur-trained CNNs always showed higher correlations than the weak-blur-trained CNNs, suggesting that training with stronger blur levels led to a closer correspondence to the cortical representations of humans in blurry viewing conditions.

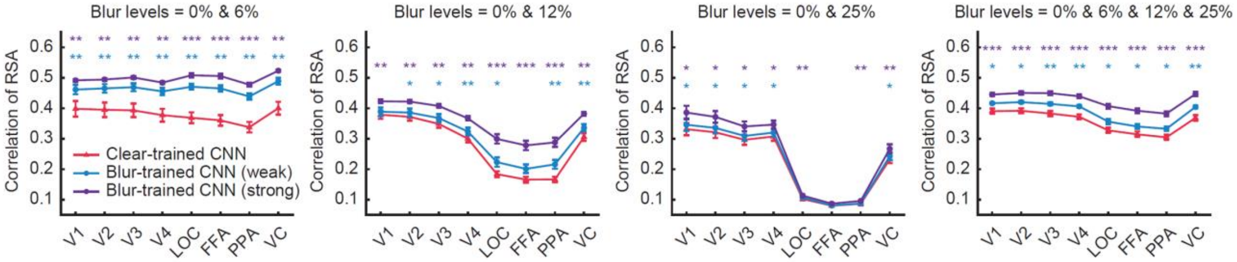

Figure 22. Correlation of RSAs between human observers and CNNs across brain regions. Clear-trained CNNs (red), weak-blur-trained CNNs (blue), and strong-blur-trained CNNs (purple) were evaluated. The diagonals of RSA matrices were excluded for analysis.

We further sought to determine how blur-trained CNNs achieved more robust performance to blurry objects than clear-trained CNNs by probing the orientation and spatial frequency tuning curves of individual convolutional units across the layers of AlexNet. **Figure 23** shows the histograms of the peak and bandwidth of orientation tuning curves and the peak and bandwidth of spatial frequency tuning curves for all CNNs. A significant difference between the clear-trained CNN and weak-blur-trained CNN was observed in their spatial frequencies peaks in layers 1 and 2 (**Figure 23a**). The weak-blur-trained CNN had more convolutional units preferring low spatial frequencies than the clear-trained CNN in the early layers, possibly accounting for its greater robustness to blur, as revealed in the accuracy patterns. In addition, the weak-blur-trained CNN demonstrated broader orientation tuning and spatial frequency tuning curves in layer 1. Interestingly, these patterns recurred in layer 5, the last convolutional layer. Differences between the clear-trained and blurred-trained CNNs were even more pronounced with strong blur training (**Figure 23b**). Across all 5 layers, the strong-blur-trained CNN showed greater preferences for lower spatial frequencies than the clear-trained CNN. The wider orientation and spatial frequency tuning curves were also observed in the strong-blur-trained CNN across most of the layers.

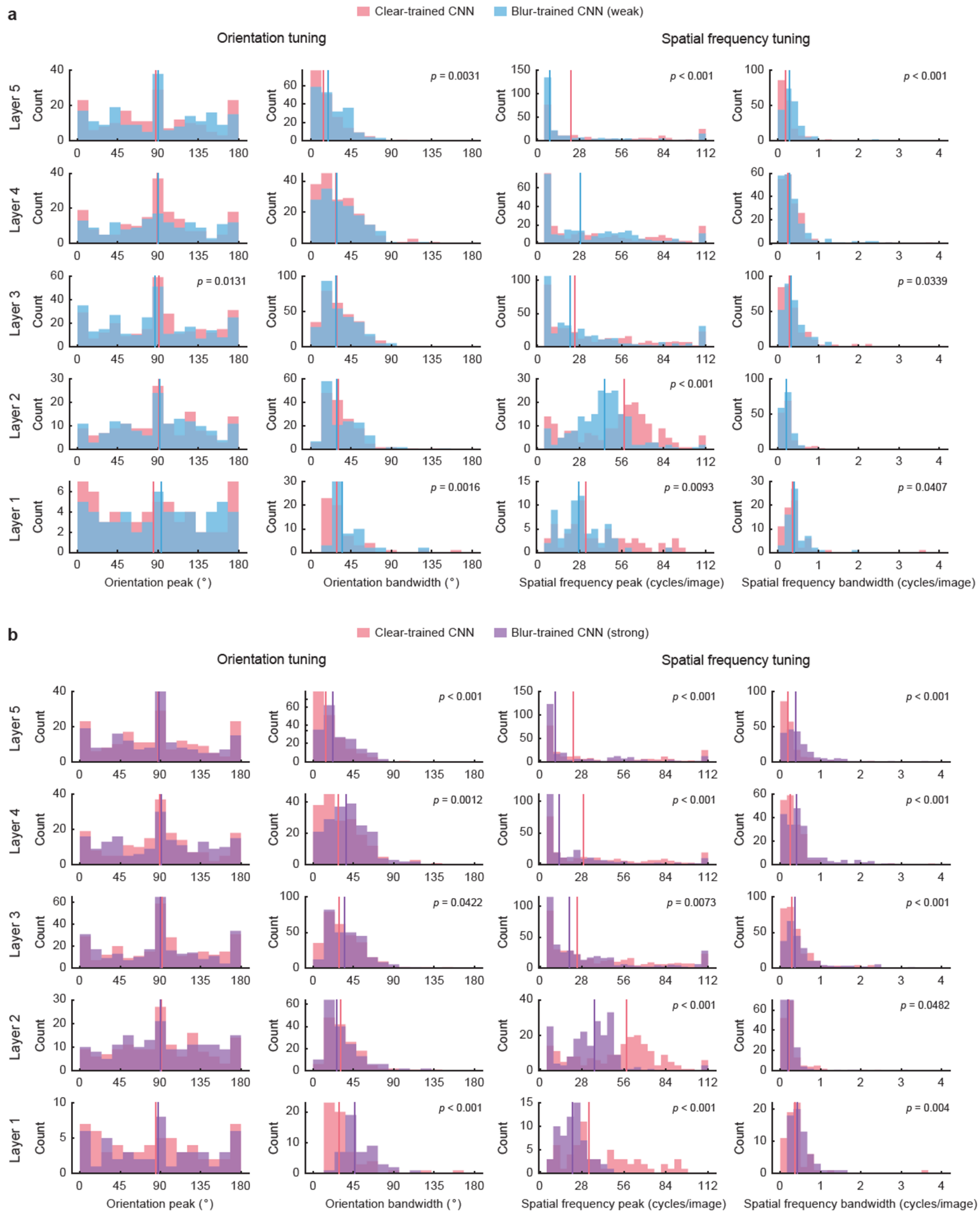

Figure 23. Histograms of the peak and bandwidth of orientation tuning curve and the peak and bandwidth of spatial frequency tuning curve (from left to right) examined by layerwise individual convolutional units of clear-trained CNNs, weak-blur-trained CNNs (**a**), and strong-blur-trained CNNs (**b**). The horizontal line indicates the median of the histogram.

If humans truly benefit from blurry visual experiences, we speculated that blur-trained CNNs would better match other human recognition behaviors than clear-trained CNNs. Previously, Geirhos et al. (2019) have reported a striking difference between humans and CNNs that humans tend to recognize objects based on their shape information while CNNs do so based on texture cues. We found that the weak-blur-trained CNNs exhibited a stronger shape bias across categories in general (**Figure 24a**), though only 3 categories revealed statistical significance. By comparison, the strong-blur-trained CNNs clearly showed higher shape bias for most categories, demonstrating that blur training can significantly mitigate the high texture bias of CNNs. In **Figure 24b**, the layerwise relevance propagation visualization technique demonstrates how the clear-trained CNN was biased to texture cues (e.g., dog textures over keyboard shape) and the blur-trained CNN was not (e.g., keyboard shape over dog textures).

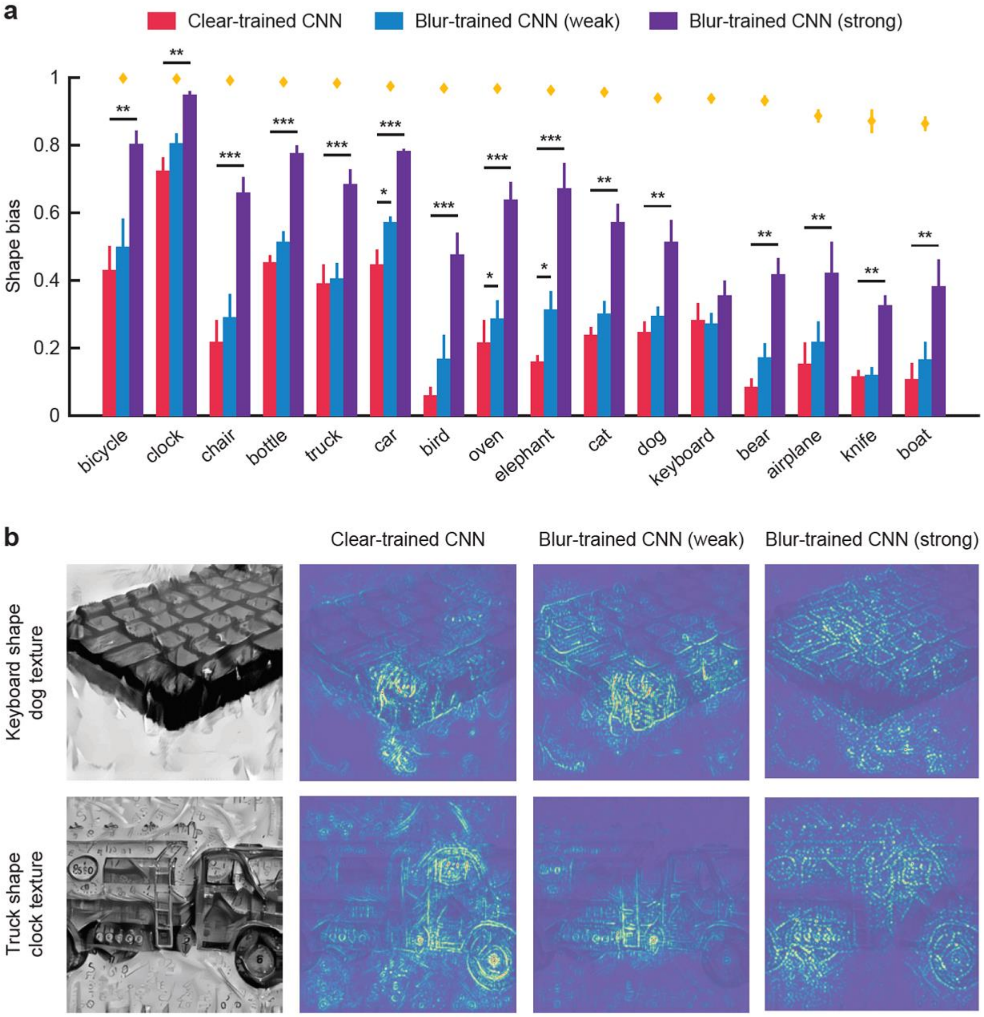

Figure 24. a Shape bias of clear-trained CNNs (red), weak-blur-trained CNNs (blue), and strong-blur-trained CNNs (purple) on shape-texture cue conflict stimuli by Geirhos et al. (2019). Human performance is displayed as yellow diamonds. **b** Heatmap examples of shape-texture cue conflict stimuli obtained by each of the CNNs using layerwise relevance propagation.

Many recent studies have demonstrated that CNNs show extremely poor performance at recognizing objects when inputs are corrupted by noise (Dodge and Karam, 2017; Geirhos et al., 2018; Jang and Tong, 2018). A recent benchmark dataset, ImageNet-C, introduced 19 different noise types under 5 categories to evaluate the robustness of CNN models (Hendrycks and Dietterich, 2019; **Figure 25b**). We observed that both versions of blur-trained CNNs outperformed the clear-trained CNNs on every noise type (**Figures 25a**). In particular, by comparing the two versions of blur-trained CNNs, the strong-blur-trained CNNs usually provided better performance than the weak-blur-trained CNNs, suggesting that blurry visual experiences would help increase the robustness to various types of visual noise in general.

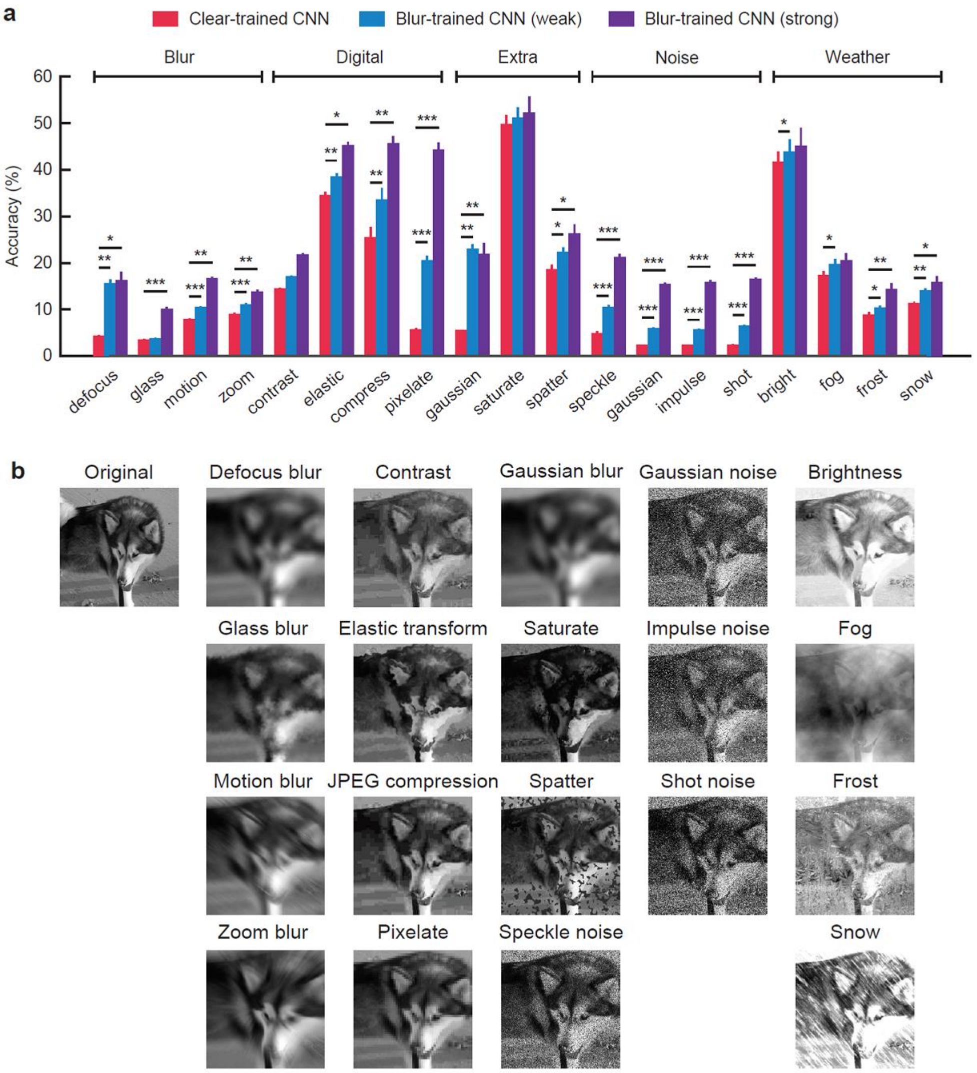

Figure 25. a Accuracy performance of clear-trained CNNs (red), weak-blur-trained CNNs (blue), and strong-blur-trained CNNs (purple) on ImageNet-C. **b** Examples of a dog image in ImageNet-C.

We further tested whether blur training allowed CNNs to better capture the response patterns of human behavioral and neural data to pixelated Gaussian noise and Fourier phase-scrambled noise, previously collected in Chapter 2. We found that both versions of blur-trained CNNs increased the robustness to pixelated Gaussian noise, while they did not show any improvement in their performance with respect to Fourier phase-scrambled noise (**Figure 26a**). When the neural representation similarities between humans and each of the CNNs were compared, the weak-blur-trained CNNs showed higher correlations than the clear-trained CNNs in higher visual cortical regions including V4, LOC, and PPA with pixelated Gaussian noise and in early visual areas from V1-V4 with Fourier phase-scrambled noise (**Figure 26b**). We would like to note that it was previously reported in Chapter 2 that Fourier phase-scrambled noise was particularly detrimental to humans, and this might explain why training with blurry objects did not improve the robustness to Fourier phase-scrambled noise. However, blur training helped decrease the gap of neural representations between humans and CNNs in Fourier phase-scrambled noise. The strong-blur-trained CNNs showed even higher correlations with humans. The correlations of the RSAs where all conditions were combined between humans and the strong-blur-trained CNNs were significantly higher than those between humans and the clear-trained CNNs across all brain regions. Interestingly, the strong-blur-trained CNNs were better correlated with humans than the clear-trained CNNs even at clear viewing conditions in the earliest visual area, V1.

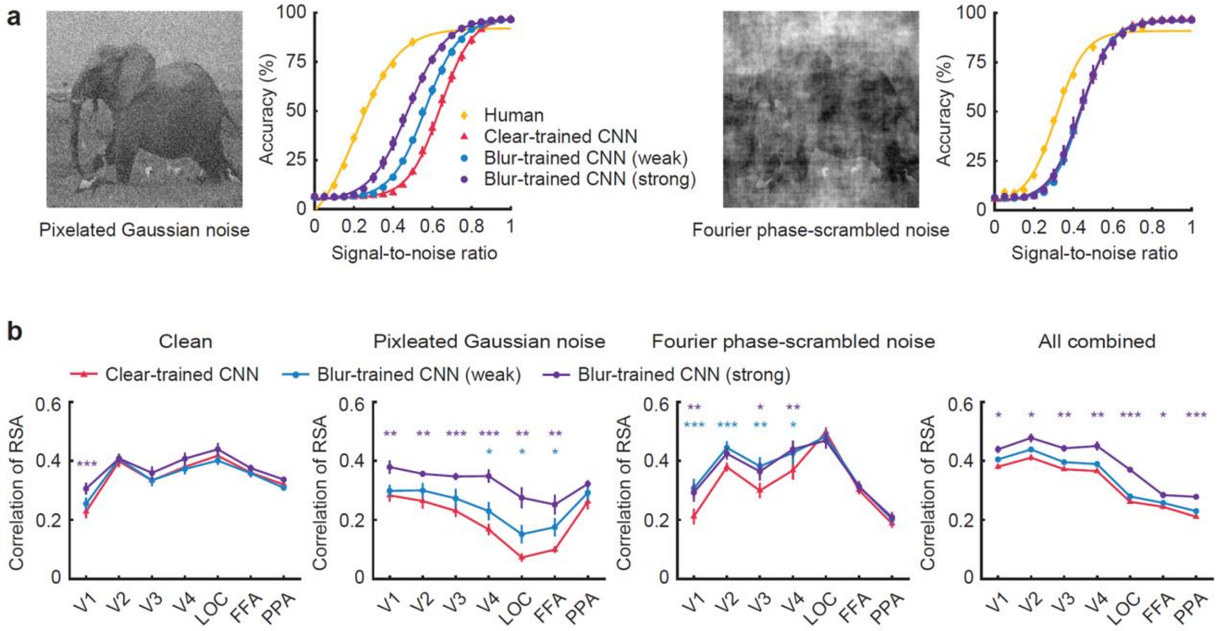

Figure 26. a Behavioral accuracy performance of human observers (yellow), clear-trained CNNs (red), weak-blur-trained CNNs (blue), and strong-blur-trained CNNs (purple) on pixelated Gaussian noise and Fourier phase-scrambled noise. **b** Correlation of RSAs between human observers and CNNs across brain regions under clean and noisy conditions. Clear-trained CNNs (red), weak-blur-trained CNNs (blue), and strong-blur-trained CNNs (purple) were evaluated. The diagonals of RSA matrices were excluded for analysis.

We further speculated that blur-trained CNNs might not just show an advantage at accounting for human cortical responses under challenging viewing conditions but might also perform better for clear viewing conditions. We evaluated this question by leveraging a large fMRI dataset previously collected by Kay et al. (2008) in which the brain responses of 2 observers were collected from 1870 natural images. In a similar manner, the RSA matrices between each brain region of the two human observers and each layer of the clear- and blur-trained CNNs were compared via Pearson correlation. As a result, however, we failed to find a significant difference between the clear-trained CNNs and both versions of blur-trained CNNs (**Figure 27a**). The blur-trained CNNs might show higher correlations than the clear-trained CNNs in the early visual areas, V1-V2, but those did not reach a significant level.

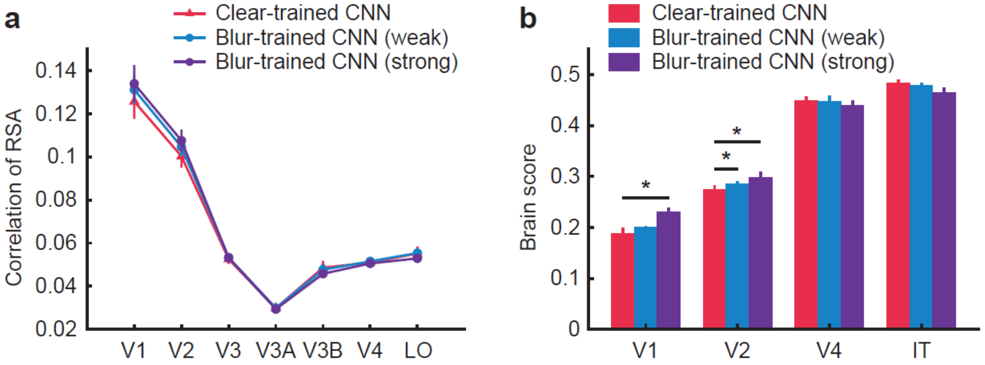

Figure 27. a Correlation of RSAs between human observers and CNNs across brain regions, evaluated by the fMRI dataset from Kay et al. (2008). Clear-trained CNNs (red), weak-blur-trained CNNs (blue), and strong-blur-trained CNNs (purple) were tested. **b** Brainscore of Clear-trained CNNs (red), weak-blur-trained CNNs (blue), and strong-blur-trained CNNs (purple).

Perhaps, the difference between the clear- and blur-trained CNNs at clear viewing conditions may not be sufficiently large to be detected at the macroscopical level by fMRI, but could be detected at the neural level. We further compared the clear- and blur-trained CNNs using another benchmark dataset, brainscore, in their neural predictivities of V1, V2, V4, and IT (Schrimpf et al., 2018). We observed that the weak-blur-trained CNNs better predicted the neural responses of V2 than the clear-trained CNNs (**Figure 27b**). Similarly, the strong-blur-trained CNNs better predicted the neural responses of V1 and V2 than the clear-trained CNNs. Based on these results, we conclude that modern CNNs typically trained on ImageNet are likely optimized in a biased manner; that is, the networks tend to hone in on fine-detailed features of objects. Blur training can help mitigate this bias and thereby better explain human recognition behavior.

5.4 Discussion

In the present study, we investigated the possibility that blurry vision may benefit humans performing robust object recognition under various viewing conditions, as demonstrated by the comparison of two CNNs, one trained by a mixture of blurry to clear images and the other trained by clear images only. Previous studies have reported several discrepancies between humans and CNNs (Geirhos et al., 2018; Jang and Tong, 2018; Geirhos et al., 2019). Here, we showed that simply adding a small fraction of blurry images to training not only mitigated previously reported issues with CNNs, such as the strong texture bias and poor robustness to noise, but also increased their neural correspondence to biological vision.

The fact that ImageNet-trained CNNs are biased to emphasize fine-grained features while recognizing objects has been observed by other studies as well. Geirhos et al. (2019) have nicely demonstrated the texture bias of ImageNe-trained CNNs by utilizing artificial stimuli conflicting shape and texture cue information. To mitigate the strong texture bias, the authors proposed training CNNs with so-called stylized images, in which ImageNet images were transformed by a style transfer algorithm (Huang & Belongie, 2017), thereby reducing the dependence on texture cues. However, this approach seems rather technical and like an engineering perspective but lacks biological plausibility as natural images do not appear to be “stylized” in the wild. More importantly, the authors observed that stylized-ImageNet-trained CNNs performed worse than ImageNet-trained CNNs on low-pass filtered or blurred images (Geirhos et al., 2019), which implies that our blur training approach fundamentally differs from theirs. A recent study has incorporated multiple Gabor filters at the front end of ResNet50 and observed enhanced robustness as well as increased neural predictivity (Dapello et al., 2020). Particularly, the low spatial frequency filters were critical for the increase in robustness to noise and blur. Another study by Kong et al. (2021) has examined the eigenspectrum of CNNs and demonstrated that ImageNet-trained CNNs exhibited a stronger preference towards high spatial frequencies than macaque V1. These results are concordant with our neural predictivity analysis that blur-trained CNNs better account for V1 and V2 responses than clear-trained CNNs.

Although several recent CNN studies suggest that blurry vision may be beneficial for acquiring robust object recognition, it remains difficult to confirm whether humans truly benefit from blurry vision. According to the coarse-to-fine hypothesis (Watt, 1987; Schyns and Oliva, 1994; Bullier, 2001), low spatial frequency information is first processed by the visual system and subsequently guides visual object processing of fine-detailed features. Bar et al. (2006) claimed that low spatial frequency components of images reached earlier the prefrontal cortex and are projected back to the ventral stream via top-down feedback processing. Low spatial frequency information in the periphery, such as contextual cues from scenes, may also influence object recognition (Oliva and Torralba, 2007; Roux-Sibilon et al., 2019). Altogether, it will be of future interest to determine the role of blurry vision in object recognition made by a direct link between machine and human vision.

Finally, although our approach was rather computationally straightforward and simply required including both clear and blurry images in CNN training, it should be noted that this may not be the best approach for approximating the nature of blurry vision in humans. For example, Deza and Konkle (2020) implemented a foveated vision model by applying different standard deviations of a Gaussian blur kernel at different eccentricities, as more similar to the approach by Freeman and Simoncelli (2011). Han et al. (2020) instead implemented an eccentricity-dependent CNN where multi-scaled versions of center cropped images were provided as an input of the CNN. Future studies may need to further clarify the nature of blur in our perception (Sprague et al., 2016; Cholewiak et al., 2018) and how this can be implemented in machine vision.

6. General discussion

In this thesis, I sought to examine the robust nature of human object recognition under a variety of challenging viewing conditions by directly comparing human performance to CNNs. It has been claimed that CNNs have achieved human-level recognition performance (He et al., 2015) and that they share strong similarities with the neural representations of objects in the ventral visual pathway of both humans and non-human primates (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015; Yamins and DiCarlo, 2016; Kubilius et al., 2016; Cadena et al., 2019). However, these studies have mostly focused on clear and typical viewing conditions, whereas very little attention has been paid to whether CNNs perform similarly to humans under challenging viewing conditions. More importantly, it was not known whether the CNN representations of degraded objects resemble those of the human visual system. Through this thesis research, I sought to address this gap in knowledge.

Summary of findings

We have shown that CNNs are generally unstable when a modest amount of perturbation is applied to the visual inputs, leading to substantial impairment in their recognition ability. This has been observed across studies (Dodge and Karam, 2017; Geirhos et al., 2018; Jang and Tong, 2018), implying that CNNs do not have a built-in mechanism for dealing with visual noise. In particular, the observation that humans and CNNs differ in their susceptibility to the different types of noise (Jang and Tong, 2018), together with other previous reports (Eckstein et al., 2017; Geirhos et al., 2019), suggests that CNNs process objects in a fundamentally different manner than humans do. As a means to stabilize the recognition performance of CNNs to visual noise, we showed that training CNNs with noisy examples is highly effective in enhancing robustness to noise. More importantly, we found that noise-trained CNNs were better at predicting the patterns of human behavioral and neural performance under degraded viewing conditions. These findings provide empirical support that noise-trained CNNs can be used for modeling the noise-robust human visual system. Further analyses revealed that the effect of noise training operates differently across layers of CNNs. Specifically, the early to middle layers of the CNNs appeared to benefit from noise training by dampening the effect of external noise, whereas the middle to higher layers of the CNNs appeared to act as a signal amplifier specifically for relevant trained categories. This implies that the robust nature of the human recognition system may involve multi-stage processing (Pratte et al., 2013).

We further sought to investigate whether initial blurry visual experiences would be of importance for developing a robust recognition system. While this question is difficult to address in humans, as one cannot ethically manipulate the prolonged visual experience of infants, it can readily be addressed by modifying the experiences of CNNs. We found that a developmental sequence of training from blurry to clear images was effective in a face recognition task, leading to human-level robustness to variations in image blur. By contrast, blurry to clear training with object images led to negligible improvement in dealing with blur, with performance far below human levels. Our findings are consistent with the notion that face recognition may be special in certain ways; specifically, it may favor low spatial frequencies to process the configural information of faces (Farah et al., 1998; Goffaux and Rossion, 2006). In contrast to face recognition, repeated experiences of blurry visual inputs in everyday life may be necessary for acquiring and maintaining human-level robustness in object recognition. We found that CNNs trained on a mixture of clear and blurry objects were better at capturing the patterns of human behavioral and neural performance under blurry conditions. Furthermore, this improved robustness even generalized to other degraded viewing conditions such as visual noise. Our findings suggest that the human object recognition system may naturally learn to utilize a wide range of spatial frequencies to recognize objects, resulting in a more stable recognition system that is robust to multiple visual challenges.

Taken all together, the findings above demonstrate that, despite many studies suggesting that CNNs may be the best computational model for human vision when evaluated on clear conditions, they do not reliably predict human behavioral patterns under challenging viewing conditions, nor do they share similar representations of degraded objects when compared to humans. That said, training CNNs with degraded viewing conditions appears to be effective in making them more robust, by shifting their internal representations to become more aligned with those of humans. These findings raise some important questions. Can training CNNs with degraded conditions, particularly noise and blur, prove sufficient to make them human-like? Are there other types of image degradation that occur in the natural world that humans benefit from? Are there any other factors that might contribute to the robust object recognition of humans besides their diversity of visual experiences?

Despite the effectiveness of training CNNs under degraded conditions, this approach may not be the only solution, nor does it lead to perfect alignment with human performance. For example, when human noise thresholds for individual images were predicted by CNNs in Chapter 2, the noise-trained CNN showed a significantly higher correlation with the human data (r = 0.55) than the control CNN trained without noise (r = 0.27). However, this correlation value was still far below our estimates of human-to-human similarity (r = 0.94 based on split-half analysis). In addition, the blur-trained CNN was able to mitigate its texture bias but still failed to reach human-level shape bias. These results raise the possibility that humans may benefit from other types of image degradation, that humans may rely on different learning principles, or that the human visual system may take advantage of other built-in mechanisms that CNNs lack. These issues are further discussed below.

Ecological relevance of training objects with degraded viewing conditions

Before discussing other types of degradation, it would be worthwhile to consider whether the degraded conditions used in our studies, namely visual noise and blur, are ecologically suitable for capturing aspects of natural vision. For instance, one might ask whether the benefit of noise training found in Chapters 2-3 is directly relevant to understanding human vision, given that our visual world does not usually appear noisy. Admittedly, real-world objects typically obscure each other via occlusion. However, there are some cases where visual noise does not perfectly obscure but impacts a background object in an additive manner, for example, raindrops or flakes of snow that are partially transparent. Pixelated Gaussian noise may resemble to some extent this type of visual noise. On the other hand, we sometimes experience foggy vision due to fog or low clouds; such clouds can obscure entire scenes while having their own visual structure. Fourier phase-scrambled noise may somewhat resemble this type of visual noise. Also, we considered that applying noise in an additive manner could be beneficial because it offers a direct method for systematic control of signal-to-noise ratios as we introduced SSNRs in the thesis.

Training CNNs with a mixture of clear and blurry images may sound more appropriate in terms of ecological relevance, given the fact that a large portion of the retinal image appears blurry. However, it is possible that our approach in Chapters 3-4 did not provide a sufficiently realistic model of the blurry visual experiences of humans. Although the CNNs were trained across samples on varying spatial blur, an individual training image was blurred by a single blur kernel. In human vision, however, the degree of blur in the visual field is not homogeneous, but rather increases systematically with eccentricity (Strasburger et al., 2011). The lack of heterogeneity in the degree of blur across the input spatial map of the CNNs would potentially lead them to lose the opportunity to learn higher-order relationships across a range of spatial frequencies. For instance, humans may take advantage of contextual scene cues in the periphery for object recognition where one has to rely on lower spatial frequency information (Torralba, 2003).

What other types of challenging viewing conditions are there that humans might benefit from? The visual world is inherently cluttered and objects are often occluded by other objects. Because of that, humans may have to develop a robust mechanism to recognize objects where only limited features are visible due to partial occlusion. In future studies, it will be of interest to investigate whether training on this occlusion type of degradation can help CNNs achieve more robust object recognition and thereby better mimic the human visual system. In addition, regarding the problem of viewpoint invariance, humans may achieve a substantial degree of viewpoint invariance by encountering a subset of discrete viewpoints on multiple occasions (Jolicoeur, 1985; Tarr and Pinker, 1989). With respect to these learning effects found in humans, CNNs could provide a useful model to investigate how familiarity impacts the attainment of viewpoint invariance by carefully manipulating the number of training examples for individual viewpoints and assessing recognition performance across a full range of viewpoints.

Another important consideration is whether training CNNs with artificially designed noise might allow them to generalize to real-world degraded conditions. Our findings have provided preliminary evidence of successful generalization, as we saw that training CNNs on both pixelated Gaussian noise and Fourier phase-scrambled noise led to improve performance at recognizing vehicles in noisy weather conditions. Moreover, we have observed that training on one type of noise can improve the robustness to other types of untrained noise, although the extent to which such generalization occurs is still subject to some debate (Geirhos et al., 2019; Rusak et al., 2020), Therefore, such findings should be further validated in future studies.

Adversarial noise

Although this thesis has primarily focused on degraded viewing conditions such as those involving the addition of random noise patterns, it would be remiss to avoid discussion of adversarial noise. Adversarial noise is a purposefully designed perturbation that can be imperceptible to humans but exceedingly harmful to deep learning models (Goodfellow et al., 2014; Szegedy et al., 2014). Adversarial noise can be added to real-world objects (e.g., a printout of a small adversarial patch added to a traffic sign) and even effectively attack real-life applications (Brown et al., 2017). Adversarial noise can also generalize fairly well across CNN architectures trained on ImageNet (Moosavi-Dezfooli et al., 2017). As it can be critical for many deep learning applications, multiple defense strategies against adversarial noise have been proposed. For instance, Madry et al. (2017) addressed this problem by directly training models on a particular set of adversarial examples that was purposefully designed to fool the models. This approach has been successful but it demands significant training time. In a subsequent study, Shafahi et al. (2019) proposed a method to reduce training time by simultaneously updating network parameters and adversarial noise in a single backward pass. Despite many efforts and much progress, the relationship between adversarial noise and various types of image corruptions such as random Gaussian noise is still not clear. However, a few recent studies have provided some empirical evidence that the two are perhaps related; for example, training models with adversarial noise led to improve robustness to Gaussian noise and vice versa (Ford et al., 2019; Rusak et al., 2020). Similarly, another study has reported that introducing stochasticity into each unit of the first layer of a CNN led to an increase in robustness to adversarial attacks (Dapello et al., 2020).

On the other hand, some research groups have focused more on how humans perceive adversarial images, asking whether humans might be affected by adversarial noise and why humans are much more robust than CNNs. Elsayed et al. (2018) demonstrated that human observers were negatively influenced by adversarial noise when the exposure time was limited to a fraction of a second (~70ms). Zhou and Firestone (2019) performed a series of experiments in unlimited exposure duration that was particularly designed to accentuate the subtle effect of adversarial examples on humans. To be specific, in one of the experiments, the authors added a small amount of adversarial noise to random Gaussian noise patterns to mislead CNNs to choose a particular target category and found that most of the human observers well predicted the category even though the noise was hardly recognizable. The two studies above suggest that adversarial noise may impact human recognition behavior in a similar manner as it does CNNs (reviewed by Buckner, 2020). That said, the question of why the human visual system is more robust to adversarial noise than CNNs is not yet answered. Humans are not trained by adversarial examples, unlike the typical defense method used for training deep learning models. Instead, humans may benefit from real-life experiences under degraded conditions. Although Ford et al. (2019) primarily focused on the relationship between additive Gaussian noise and adversarial noise, visual experiences with various types of suboptimal viewing conditions may be advantageous for enhancing the robustness of a visual system. In addition, the observation by Elsayed et al. (2018) that limited exposure time can increase people's susceptibility to adversarial noise raises the possibility that recurrent processing may be critical for the robustness of the human visual system to adversarial noise. Future studies will have to address adversarial noise comprehensively by linking it to both human vision and different types of image degradation.

Potential differences between CNNs and the human visual system in training samples or algorithms

As we have discussed, simply training CNNs with degraded viewing conditions may not be sufficient for attaining the robustness of the human visual system. Instead, CNNs may need more ecological training samples or it could be the case that the learning principles of CNNs fundamentally differ from those of humans. Although ImageNet provides the opportunity of training with a large set of natural images (Russakovsky et al., 2015), the dataset may not adequately reflect how we view and perceive the world. A recent study developed a new dataset, which the authors argue to be more ecologically relevant as it consists of 565 basic-level categories, including human categories, selected based on their frequencies in spoken language ("Ecoset"; Mehrer et al., 2021). The authors demonstrated that the Ecoset-trained CNNs exhibited more similar representations to visual representations in human ventral temporal cortex when compared to ImageNet-trained CNNs, particularly for animate objects. Nevertheless, it would be still expected that the Ecoset-trained CNNs would not differ from ImageNet-trained CNNs in terms of robustness, because the Ecoset dataset consists primarily of clear images of objects.

In addition, as posed in Chapter 4, our visual experience changes dramatically over the early stages of infancy but CNNs do not usually address this specific aspect. Early visual experience may have a critical role in shaping the visual system. For instance, about 100 days of postnatal experience is sufficient to induce a visual preference toward own-race faces and toward a caregiver's gender (Quinn et al., 2019). One of the characteristics of early visual experience is that faces largely dominate the infant visual experience; however, the frequency of face viewing declines with age, whereas the frequency of object viewing increases (Jayaraman et al., 2015; Fausey et al., 2016; Jayaraman et al., 2017). This notion of "early faces, later objects" suggests that the human visual system may be shaped primarily by faces first and it is later fine-tuned by non-face objects based upon the early face-tuned representations (Smith and Slone, 2017). Because CNNs are typically trained on object images from scratch, this may make CNNs lose opportunities to start from face-tuned representations, which may be potentially more robust under degraded viewing conditions.

Another potential limitation of CNNs is that they are typically optimized for object classification through supervised learning. However, supervised learning with gradient descent may not coincide with how our visual system naturally develops. An alternative is the Hebbian learning rule (Hebb, 1949), as often stated “cells that fire together wire together”. According to the rule, neurons do not need any supervision but will learn statistical patterns of inputs by themselves. In parallel to advances in supervised learning, many unsupervised learning methods have been developed and used in many research problems (Hinton and Salakhutdinov, 2006; Hinton et al., 2006; Vincent et al., 2008; Lee et al., 2009; Vincent et al., 2010). One form of unsupervised learning that has proven successful for training CNNs to learn useful visual representations is contrastive learning, by which a model learns visual representations of input data by clustering example images into similar and dissimilar pairs (He et al., 2019; Chen et al., 2020). These models are more often referred to as self-supervised models but here we consider them as unsupervised for simplicity, technically because they do not require any labeled data. Zhuang et al. (2021) recently showed that such unsupervised models can achieve comparable performance in predicting the response patterns of neurons in V1, V4, and IT, as standard CNNs trained in a supervised manner. Similarly, Konkle and Alvarez (2020) have observed that contrastive unsupervised models showed comparable or even higher correspondences than the counterpart supervised models to brain representations obtained by fMRI. However, another recent study has shown that there is little difference between supervised and unsupervised models when evaluated by psychophysical measures including noise robustness, texture/shape biases, and error patterns (Geirhos et al., 2020), suggesting that more work will be needed to clarify the importance or validity of unsupervised learning in terms of its biological plausibility and how well it can account for human behavioral and neural data.

Another piece of evidence implying that CNNs may leverage a biologically implausible learning rule may be catastrophic forgetting, which refers to the fact that neural networks are prone to forget old representations after new information is learned. Many engineering methods have been proposed to mitigate this issue, often, by using regularization-based approaches (Goodfellow et al., 2013; Li and Hoiem, 2017; Kirkpatrick et al., 2017; Lee et al., 2017; Zenke et al., 2017). Other approaches based on biological insights, such as functional modularity or hippocampal replay, have been also suggested (Ellefsen et al., 2015; van de Ven et al., 2020). This catastrophic forgetting problem could be particularly critical if one considers CNNs as a model to examine the developmental trajectory of visual learning, because the effect of early training periods will be obscured by catastrophic forgetting. Thus, this should be taken into consideration for future studies that attempt to make connections between CNNs and the developmental literature.

Importance of recurrent visual processing

Multiple visual areas in the brain are densely interconnected by feedforward and feedback pathways (Felleman and Van Essen, 1991). The human visual system does not simply rely on the feedforward pathway, but instead, neurons in the higher areas also project back to lower areas, thereby modulating activity in the lower areas in a top-down or recurrent manner (Lamme and Roelfsema, 2000). Although a feedforward computational model performs sufficiently well on simple object recognition tasks (Serre et al., 2007), it may not be enough to capture the complexity of many real-life object recognition tasks (Wyatte et al., 2012; O’Reilly et al., 2013). For instance, Wyatte et al. (2012) showed that human observers performed worse in a categorization task when an occluded stimulus was followed by a patterned mask. The authors additionally showed that this masking effect was also observed in a hierarchical recurrent model but not a feedforward model, demonstrating that a computational model with recurrent processing better captured the patterns of behavioral responses. Spoerer et al. (2017) demonstrated that CNN-based recurrent models showed superior performance on a digit classification task where digits were largely occluded, as compared to pure feedforward CNN models, suggesting that recurrent processing would be particularly important for robust processing. In recent years, several recurrent CNN models have been proposed (Wen et al., 2018; Tang et al., 2018; Nayebi et al., 2018; Kietzmann et al., 2019; Spoerer et al., 2019; Kar et al., 2019; Huang et al., 2020), and some of the models have been suggested to better account for brain responses (Spoerer et al., 2019; Kar et al., 2019; Huang et al., 2020). It will be of great interest to investigate how these models perform object recognition under various degraded viewing conditions, as compared to feedforward recognition models. Although recurrent CNN models have increased in popularity in recent years, their detailed implementations have varied across studies. For instance, a neural network called CORnet implemented locally recurrent computations within individual layers (Kar et al., 2019), while deep predictive coding networks incorporate the dynamics of both bottom-up and top-down processing (Wen et al., 2018). Future studies will need to address how different implementations of recurrent processing may contribute to robust object recognition.

Conclusions

The current thesis has provided an exciting glimpse into the intersection of human and computer vision, focusing on the robust nature of object recognition. Many findings of this thesis have highlighted that human object recognition is remarkably efficient and simultaneously robust, which leads to a fundamental question as to how humans achieve both in a balanced manner. By contrast, most CNNs have been found to lack robustness; they may be highly efficient in recognizing objects remain extremely vulnerable to variations in viewing conditions. Future studies will have to comprehensively consider various factors as discussed above to determine what accounts for robust object recognition. The current thesis has also demonstrated how an interdisciplinary approach can benefit both research fields, neuroscience and artificial intelligence, to better understand the nature of object recognition. Computer science and neuroscience often have very different views, since the former focuses more on practical applications and the latter on more fundamental research questions. Instead of focusing on one aspect, interdisciplinary research can bring unique insights that would not be offered by traditional views and help provide a broader perspective. Through this thesis research, my aim was to contribute to and enrich the intersection of these two fields.

Reference

Abdelhack, M., & Kamitani, Y. (2018). Sharpening of hierarchical visual feature representations of blurred images. Eneuro, 5(3).

Achille, A., Rovere, M., & Soatto, S. (2018, September). Critical learning periods in deep networks. In International Conference on Learning Representations.

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. Josa a, 2(2), 284-299.

An, G. (1996). The effects of adding noise during backpropagation training on a generalization performance. Neural computation, 8(3), 643-674.

Anstis S. M. (1974). Letter: A chart demonstrating variations in acuity with retinal position. Vision research, 14(7), 589–592. https://doi.org/10.1016/0042-6989(74)90049-2

Audhkhasi, K., Osoba, O., & Kosko, B. (2016). Noise-enhanced convolutional neural networks. Neural Networks, 78, 15-23.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7), e0130140.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. PLoS computational biology, 14(12), e1006613.

Ballester, P., & Araujo, R. M. (2016, February). On the performance of GoogLeNet and AlexNet applied to sketches. In Thirtieth AAAI Conference on Artificial Intelligence.

Bambach, S., Crandall, D. J., Smith, L. B., & Yu, C. (2018, December). Toddler-inspired visual object learning. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (pp. 1209-1218).

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., ... & Halgren, E. (2006). Top-down facilitation of visual recognition. Proceedings of the national academy of sciences, 103(2), 449-454.

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

Benson NC, Winawer J. Bayesian analysis of retinotopic maps. elife. 2018;7:e40224.

Biederman, I. (1985). Human image understanding: Recent research and a theory. Computer vision, graphics, and image processing, 32(1), 29-73.

Bishop, C. M. (1995). Training with noise is equivalent to Tikhonov regularization. Neural computation, 7(1), 108-116.

Brincat, S. L., & Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. Nature neuroscience, 7(8), 880-886.

Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch. arXiv preprint arXiv:1712.09665.

Buckner, C. (2020). Understanding adversarial examples requires a theory of artefacts for deep learning. Nature Machine Intelligence, 2(12), 731-736.

Bullier, J. (2001). Integrated model of visual processing. Brain research reviews, 36(2-3), 96-107.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proceedings of the National Academy of Sciences, 89(1), 60-64.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS computational biology, 15(4), e1006897.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., ... & Rust, N. C. (2005). Do we know what the early visual system does?. Journal of Neuroscience, 25(46), 10577-10597.

Carrasco, M., Ling, S., & Read, S. (2004). Attention alters appearance. Nature neuroscience, 7(3), 308-313.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. arXiv preprint arXiv:1405.3531.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

Cholewiak, S. A., Love, G. D., & Banks, M. S. (2018). Creating correct blur and its effect on accommodation. Journal of Vision, 18(9), 1-1.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D. D., & DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. BioRxiv.

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. Journal of Neuroscience, 4(8), 2051-2062.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. Annual review of neuroscience, 18(1), 193-222.

Deza, A., & Konkle, T. (2020). Emergent properties of foveated perceptual systems. arXiv preprint arXiv:2006.07991.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. Trends in cognitive sciences, 11(8), 333-341.

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

Dobson, V., & Teller, D. Y. (1978). Visual acuity in human infants: a review and comparison of behavioral and electrophysiological studies. Vision research, 18(11), 1469-1483.

Dodge, S., & Karam, L. (2017, July). A study and comparison of human and deep learning recognition performance under visual distortions. In 2017 26th international conference on computer communication and networks (ICCCN) (pp. 1-7). IEEE.

Dosher, B. A., & Lu, Z. L. (1998). External noise distinguishes attention mechanisms. Vision research, 38(9), 1183-1198.

Dosher, B. A., & Lu, Z. L. (2005). Perceptual learning in clear displays optimizes perceptual expertise: Learning the limiting process. Proceedings of the National Academy of Sciences, 102(14), 5286-5290.

Dosher, B. A., Jeter, P., Liu, J., & Lu, Z. L. (2013). An integrated reweighting theory of perceptual learning. Proceedings of the National Academy of Sciences, 110(33), 13678-13683.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. Science, 293(5539), 2470-2473.

Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but not deep neural networks, often miss giant targets in scenes. Current Biology, 27(18), 2827-2832.

Ellefsen, K. O., Mouret, J. B., & Clune, J. (2015). Neural modularity helps organisms evolve to learn new skills without forgetting old skills. PLoS computational biology, 11(4).

Elsayed, G. F., 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. arXiv preprint arXiv:1802.08195.

Engel, S. A., Glover, G. H., & Wandell, B. A. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cerebral cortex (New York, NY: 1991), 7(2), 181-192.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature, 392(6676), 598-601.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. nature, 542(7639), 115-118.

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. Nature reviews neuroscience, 9(4), 292-303.

Farah, M. J., Rabinowitz, C., Quinn, G. E., & Liu, G. T. (2000). Early commitment of neural substrates for face recognition. Cognitive neuropsychology, 17(1-3), 117-123.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is" special" about face perception?. Psychological review, 105(3), 482.

Fausey, C. M., Jayaraman, S., & Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. Cognition, 152, 101-107.

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.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. Nature neuroscience, 14(9), 1195-1201.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. Nature neuroscience, 16(7), 974-981.

Fukushima, K. (1980). Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biological Cybernetics, 36, 193--202.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2414-2423).

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), 665-673.

Geirhos, R., Narayanappa, K., Mitzkus, B., Bethge, M., Wichmann, F. A., & Brendel, W. (2020). On the surprising similarities between supervised and self-supervised models. arXiv preprint arXiv:2010.08377.

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

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. Advances in Neural Information Processing Systems, 31, 7538-7550.

Geldart, S., Mondloch, C. J., Maurer, D., De Schonen, S., & Brent, H. P. (2002). The effect of early visual deprivation on the development of face processing. Developmental Science, 5(4), 490-501.

Gilmer, J., Ford, N., Carlini, N., & Cubuk, E. (2019, May). Adversarial examples are a natural consequence of test error in noise. In International Conference on Machine Learning (pp. 2280-2289). PMLR.

Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

Glorot, X., Bordes, A., & Bengio, Y. (2011, June). Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 315-323). JMLR Workshop and Conference Proceedings.

Goffaux, V., & Rossion, B. (2006). Faces are" spatial"--holistic face perception is supported by low spatial frequencies. Journal of Experimental Psychology: Human Perception and Performance, 32(4), 1023.

Goffaux, V., Gauthier, I., & Rossion, B. (2003). Spatial scale contribution to early visual differences between face and object processing. Cognitive Brain Research, 16(3), 416-424.

Gold, J. M., Sekuler, A. B., & Bennett, P. J. (2004). Characterizing perceptual learning with external noise. Cognitive Science, 28(2), 167-207.

Gold, J., Bennett, P. J., & Sekuler, A. B. (1999). Signal but not noise changes with perceptual learning. Nature, 402(6758), 176-178.

Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., & Bengio, Y. (2013). An empirical investigation of catastrophic forgetting in gradient-based neural networks. arXiv preprint arXiv:1312.6211.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. Vision research, 41(17), 2261-2271.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. Neuroimage, 48(1), 63-72.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. Vision research, 41(10-11), 1409-1422.

Gross, C. G., Rocha-Miranda, C. D., & Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. Journal of neurophysiology, 35(1), 96-111.

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.

Han, Y., Roig, G., Geiger, G., & Poggio, T. (2020). Scale and translation-invariance for novel objects in human vision. Scientific reports, 10(1), 1-13.

Harel, A., & Bentin, S. (2009). Stimulus type, level of categorization, and spatial-frequencies utilization: implications for perceptual categorization hierarchies. Journal of Experimental Psychology: Human Perception and Performance, 35(4), 1264.

Harmon, L. D., & Julesz, B. (1973). Masking in visual recognition: Effects of two-dimensional filtered noise. Science, 180(4091), 1194-1197.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9729-9738).

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026-1034).

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).

Hebb, D. O. (1949). The organisation of behaviour: a neuropsychological theory. New York: Science Editions.

Held, R. T., Cooper, E. A., & Banks, M. S. (2012). Blur and disparity are complementary cues to depth. Current biology, 22(5), 426-431.

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

Hinton, G. E. (2009). Deep belief networks. Scholarpedia, 4(5), 5947.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. science, 313(5786), 504-507.

Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1501-1510).

Huang, Y., Gornet, J., Dai, S., Yu, Z., Nguyen, T., Tsao, D. Y., & Anandkumar, A. (2020). Neural networks with recurrent generative feedback. arXiv preprint arXiv:2007.09200.

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.

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.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. The Journal of physiology, 195(1), 215-243.

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.

Jang, H., & Tong, F. (2018). Can deep learning networks acquire the robustness of human recognition when faced with objects in visual noise?. Journal of Vision, 18(10), 903-903.

Jayaraman, S., Fausey, C. M., & Smith, L. B. (2015). The faces in infant-perspective scenes change over the first year of life. PloS one, 10(5), e0123780.

Jayaraman, S., Fausey, C. M., & Smith, L. B. (2017). Why are faces denser in the visual experiences of younger than older infants?. Developmental psychology, 53(1), 38.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage, 17(2), 825-841.

Jolicoeur, P. (1985). The time to name disoriented natural objects. Memory & cognition, 13(4), 289-303.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. Nature neuroscience, 8(5), 679-685.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. Journal of neuroscience, 17(11), 4302-4311.

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.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. Nature, 452(7185), 352-355.

Kell, A. J., & McDermott, J. H. (2019). Invariance to background noise as a signature of non-primary auditory cortex. Nature communications, 10(1), 1-11.

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

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

Kietzmann, T. C., Swisher, J. D., König, P., & Tong, F. (2012). Prevalence of selectivity for mirror-symmetric views of faces in the ventral and dorsal visual pathways. Journal of Neuroscience, 32(34), 11763-11772.

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

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13), 3521-3526.

Kong, N. C., Margalit, E., Gardner, J. L., & Norcia, A. M. (2021). Increasing neural network robustness improves match to macaque V1 eigenspectrum, spatial frequency preference and predictivity. bioRxiv.

Konkle, T., & Alvarez, G. A. (2020). Instance-level contrastive learning yields human brain-like representation without category-supervision. bioRxiv.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. PLoS computational biology, 12(4), e1004896.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. Trends in neurosciences, 23(11), 571-579.

Le Grand, R., Mondloch, C. J., Maurer, D., & Brent, H. P. (2001). Early visual experience and face processing. Nature, 410(6831), 890-890.

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

Lecun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In D. Touretzky (Ed.), Advances in Neural Information Processing Systems (NIPS 1989), Denver, CO (Vol. 2). Morgan Kaufmann.

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

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009, June). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the 26th annual international conference on machine learning (pp. 609-616).

Lee, S. W., Kim, J. H., Jun, J., Ha, J. W., & Zhang, B. T. (2017). Overcoming catastrophic forgetting by incremental moment matching. In Advances in neural information processing systems (pp. 4652-4662).

Li, Z., & Hoiem, D. (2017). Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence, 40(12), 2935-2947.

Liao, S. H. (2005). Expert system methodologies and applications—a decade review from 1995 to 2004. Expert systems with applications, 28(1), 93-103.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88.

Liu, J., Harris, A., & Kanwisher, N. (2010). Perception of face parts and face configurations: an fMRI study. Journal of cognitive neuroscience, 22(1), 203-211.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. Current biology, 5(5), 552-563.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

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

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.

Mandler, J. M., & McDonough, L. (1993). Concept formation in infancy. Cognitive development, 8(3), 291-318.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. Proceedings of the Royal Society of London. Series B. Biological Sciences, 200(1140), 269-294.

Marshall, J. A., Burbeck, C. A., Ariely, D., Rolland, J. P., & Martin, K. E. (1996). Occlusion edge blur: a cue to relative visual depth. JOSA A, 13(4), 681-688.

McKyton, A., Ben-Zion, I., Doron, R., & Zohary, E. (2015). The limits of shape recognition following late emergence from blindness. Current Biology, 25(18), 2373-2378.

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

Moeller, S., Freiwald, W. A., & Tsao, D. Y. (2008). Patches with links: a unified system for processing faces in the macaque temporal lobe. Science, 320(5881), 1355-1359.

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. (2019). Layer-Wise Relevance Propagation: An Overview. Explainable AI.

Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal adversarial perturbations. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1765-1773).

Morrone, M. C., Burr, D. C., & Ross, J. (1983). Added noise restores recognizability of coarse quantized images. Nature, 305(5931), 226-228.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., ... & Yamins, D. L. (2018). Task-driven convolutional recurrent models of the visual system. arXiv preprint arXiv:1807.00053.

Ng, H. W., & Winkler, S. (2014, October). A data-driven approach to cleaning large face datasets. In 2014 IEEE international conference on image processing (ICIP) (pp. 343-347). IEEE.

Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE international conference on computer vision (pp. 1520-1528).

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face space representations in deep convolutional neural networks. Trends in cognitive sciences, 22(9), 794-809.

Oleskiw, T. D., Nowack, A., & Pasupathy, A. (2018). Joint coding of shape and blur in area V4. Nature communications, 9(1), 1-13.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. Trends in cognitive sciences, 11(12), 520-527.

O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., & Jilk, D. J. (2013). Recurrent processing during object recognition. Frontiers in psychology, 4, 124.

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. Nature Reviews Neuroscience, 5(4), 291-303.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition.

Pasupathy, A., & Connor, C. E. (2002). Population coding of shape in area V4. Nature neuroscience, 5(12), 1332-1338.

Peissig, J. J., & Tarr, M. J. (2007). Visual object recognition: Do we know more now than we did 20 years ago?. Annu. Rev. Psychol., 58, 75-96.

Perrett, D. I., Mistlin, A. J., & Chitty, A. J. (1987). Visual neurones responsive to faces. Trends in Neurosciences, 10(9), 358-364.

Perrett, D. I., Rolls, E. T., & Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. Experimental brain research, 47(3), 329-342.

Poggio, T., & Edelman, S. (1990). A network that learns to recognize three-dimensional objects. Nature, 343(6255), 263-266.

Pospisil, D. A., Pasupathy, A., & Bair, W. (2018). 'Artiphysiology'reveals V4-like shape tuning in a deep network trained for image classification. Elife, 7, e38242.

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. Nature Reviews Neuroscience, 1(2), 125-132.

Pratte, M. S., Ling, S., Swisher, J. D., & Tong, F. (2013). How attention extracts objects from noise. Journal of Neurophysiology, 110(6), 1346-1356.

Putzar, L., Goerendt, I., Lange, K., Rösler, F., & Röder, B. (2007). Early visual deprivation impairs multisensory interactions in humans. Nature neuroscience, 10(10), 1243-1245.

Quinn, P. C. (2004). Development of subordinate-level categorization in 3-to 7-month-old infants. Child Development, 75(3), 886-899.

Quinn, P. C., Lee, K., & Pascalis, O. (2019). Face processing in infancy and beyond: The case of social categories. Annual review of psychology, 70, 165-189.

Rainer, G., & Miller, E. K. (2000). Effects of visual experience on the representation of objects in the prefrontal cortex. Neuron, 27(1), 179-189.

Rainer, G., Lee, H., Logothetis, N. K., & Desimone, R. (2004). The effect of learning on the function of monkey extrastriate visual cortex. PLoS biology, 2(2), e44.

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.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 91-99.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. Neuron, 61(2), 168-185.

Ricci, M., Kim, J., & Serre, T. (2018). Same-different problems strain convolutional neural networks. arXiv preprint arXiv:1802.03390.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nature neuroscience, 2(11), 1019-1025.

Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

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

Roux-Sibilon, A., Trouilloud, A., Kauffmann, L., Guyader, N., Mermillod, M., & Peyrin, C. (2019). Influence of peripheral vision on object categorization in central vision. Journal of vision, 19(14), 7-7.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.

Rusak, E., Schott, L., Zimmermann, R. S., Bitterwolf, J., Bringmann, O., Bethge, M., & Brendel, W. (2020, August). A simple way to make neural networks robust against diverse image corruptions. In European Conference on Computer Vision (pp. 53-69). Springer, Cham.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... & DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like?. BioRxiv, 407007.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time-and spatial-scale-dependent scene recognition. Psychological science, 5(4), 195-200.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. Proceedings of the national academy of sciences, 104(15), 6424-6429.

Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., ... & Goldstein, T. (2019). Adversarial training for free!. arXiv preprint arXiv:1904.12843.

Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. Annual review of biomedical engineering, 19, 221-248.

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

Sinha, P. (2002). Recognizing complex patterns. nature neuroscience, 5(11), 1093-1097.

Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning?. Frontiers in psychology, 8, 2124.

Solomon, J. A., & Pelli, D. G. (1994). The visual filter mediating letter identification. Nature, 369(6479), 395-397.

Song, Y., Qu, Y., Xu, S., & Liu, J. (2020). Implementation-independent representation for deep convolutional neural networks and humans in processing faces. Frontiers in computational neuroscience, 14.

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent networks can recycle neural resources to flexibly trade speed for accuracy in visual recognition. BioRxiv, 677237.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: a better model of biological object recognition. Frontiers in psychology, 8, 1551.

Sprague, W. W., Cooper, E. A., Reissier, S., Yellapragada, B., & Banks, M. S. (2016). The natural statistics of blur. Journal of vision, 16(10), 23-23.

Strasburger, H., Rentschler, I., & Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. Journal of vision, 11(5), 13-13.

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., & Harris, K. D. (2019). High-dimensional geometry of population responses in visual cortex. Nature, 571(7765), 361-365.

Swindale, N. V. (1998). Orientation tuning curves: empirical description and estimation of parameters. Biological cybernetics, 78(1), 45-56.

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

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).

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1701-1708).

Tan, C., & Poggio, T. (2016). Neural tuning size in a model of primate visual processing accounts for three key markers of holistic face processing. PloS one, 11(3), e0150980.

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., ... & Kreiman, G. (2018). Recurrent computations for visual pattern completion. Proceedings of the National Academy of Sciences, 115(35), 8835-8840.

Tanner Jr, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. Psychological review, 61(6), 401.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. Cognitive psychology, 21(2), 233-282.

Tarr, M. J., Williams, P., Hayward, W. G., & Gauthier, I. (1998). Three-dimensional object recognition is viewpoint dependent. Nature neuroscience, 1(4), 275-277.

Thorpe, S. J., Gegenfurtner, K. R., Fabre-Thorpe, M., & BuÈlthoff, H. H. (2001). Detection of animals in natural images using far peripheral vision. European Journal of Neuroscience, 14(5), 869-876.

Tong, F. (2018). Foundations of Vision. Leading chapter for Volume 2 of Sensation, Perception & Attention, The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Editors John T.

Tong, F., & Nakayama, K. (1999). Robust representations for faces: evidence from visual search. Journal of Experimental Psychology: Human Perception and Performance, 25(4), 1016.

Tong, F., Nakayama, K., Moscovitch, M., Weinrib, O., & Kanwisher, N. (2000). Response properties of the human fusiform face area. Cognitive neuropsychology, 17(1-3), 257-280.

Tong, F., & Jang, H. (2021). Noise-robust neural networks and methods thereof. (U.S. Patent No. 11,030,487). U.S. Patent and Trademark Office.

Torralba, A. (2003). Contextual priming for object detection. International journal of computer vision, 53(2), 169-191.

Treue, S., & Trujillo, J. C. M. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. Nature, 399(6736), 575-579.

Tsao, D. Y., & Livingstone, M. S. (2008). Mechanisms of face perception. Annu. Rev. Neurosci., 31, 411-437.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. Science, 311(5761), 670-674.

Tsunoda, K., Yamane, Y., Nishizaki, M., & Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. Nature neuroscience, 4(8), 832-838.

van de Ven, G. M., Siegelmann, H. T., & Tolias, A. S. (2020). Brain-inspired replay for continual learning with artificial neural networks. Nature communications, 11(1), 1-14.

Vasiljevic, I., Chakrabarti, A., & Shakhnarovich, G. (2016). Examining the impact of blur on recognition by convolutional networks. arXiv preprint arXiv:1611.05760.

Vedaldi, A., & Lenc, K. (2015, October). Matconvnet: Convolutional neural networks for matlab. In Proceedings of the 23rd ACM international conference on Multimedia (pp. 689-692).

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning (pp. 1096-1103).

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P. A., & Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research, 11(12).

Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., Yonas, A., Diamond, S., Held, R., & Sinha, P. (2018). Potential downside of high initial visual acuity. Proceedings of the National Academy of Sciences, 115(44), 11333-11338.

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

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. Perception & psychophysics, 33(2), 113-120.

Watt, R. J. (1987). Scanning from coarse to fine spatial scales in the human visual system after the onset of a stimulus. JOSA A, 4(10), 2006-2021.

Wen, H., Han, K., Shi, J., Zhang, Y., Culurciello, E., & Liu, Z. (2018, July). Deep predictive coding network for object recognition. In International Conference on Machine Learning (pp. 5266-5275). PMLR.

Wichmann, F. A., Braun, D. I., & Gegenfurtner, K. R. (2006). Phase noise and the classification of natural images. Vision research, 46(8-9), 1520-1529.

Wyatte, D., Curran, T., & O'Reilly, R. (2012). The limits of feedforward vision: Recurrent processing promotes robust object recognition when objects are degraded. Journal of Cognitive Neuroscience, 24(11), 2248-2261.

Xu, Y., & Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. Nature communications, 12(1), 1-16.

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

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.

Zenke, F., Poole, B., & Ganguli, S. (2017, August). Continual learning through synaptic intelligence. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 3987-3995). JMLR. org.

Zheng, S., Song, Y., Leung, T., & Goodfellow, I. (2016). Improving the robustness of deep neural networks via stability training. In Proceedings of the ieee conference on computer vision and pattern recognition (pp. 4480-4488).

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. Nature communications, 10(1), 1-9.

Zhu, S. C., & Mumford, D. (2007). A stochastic grammar of images. Now Publishers Inc.

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).