Noise-robust recognition of objects by humans and deep neural networks

Hojin Jang*, Devin McCormack, and Frank Tong*

Department of Psychology and Vanderbilt Vision Research Center

Vanderbilt University

Corresponding authors: frank.tong@vanderbilt.edu and hojin.jang@vanderbilt.edu

ABSTRACT

Deep neural networks (DNNs) can accurately recognize objects in clear viewing conditions, leading to claims that they have attained or surpassed human-level performance. However, standard DNNs are severely impaired at recognizing objects in visual noise, whereas human vision remains robust. We developed a noise-training procedure, generating noisy images of objects with low signal-to-noise ratio, to investigate whether DNNs can acquire robustness that better matches human vision. After noise training, DNNs outperformed human observers while exhibiting more similar patterns of performance, and provided a better model for predicting human recognition thresholds on an image-by-image basis. Noise training also improved DNN recognition of vehicles in noisy weather. Layer-specific analyses revealed that the contaminating effects of noise were dampened, rather than amplified, across successive stages of the noise-trained network, with greater benefit at higher levels of the network. Our findings indicate that DNNs can learn noise-robust representations that better approximate human visual processing.

Introduction

Research in vision science has demonstrated that people excel at recognizing objects quickly, accurately, and across a diversity of viewing conditions [1, 2, 3, 4]. In most domains, human vision has constituted the gold standard for evaluating visual performance, whether in the context of everyday tasks such as searching for a friend in a crowd, navigating through a busy city street, or when performing specialized tasks that require considerable expertise, such as diagnostic radiology [2, 5, 6, 7].

However, recent advances in deep learning have led to a seismic shift in what might be considered the gold standard of visual performance. There is a growing body of evidence that many diverse tasks of visual recognition can be solved given a suitable and sufficiently deep neural architecture, extensive opportunities for supervised learning, and a large, heterogeneous set of data for training [8]. The original convolutional neural network (CNN) model was inspired by functional architecture of the primary visual cortex [9], and performs a series of operations akin to spatial filtering (or convolution) followed by local spatial pooling [10, 11, 12, 13]. Following the success of AlexNet [13], deeper and more sophisticated CNNs have been developed, leading to more powerful and accurate performance at recognizing images of real world objects from large data bases [14, 15, 16]. Recent comparisons between deep networks and humans have led to claims that state-of-the-art CNNs have achieved or even surpassed human-level performance at object recognition [8, 17, 18] and other challenging visual tasks [19, 20, 21].

In parallel, research in basic neuroscience has suggested that the visual object representations learned by these deep neural networks (DNNs) share a strong resemblance with those found in the primate visual system. Indeed, the response preferences of neurons in the inferotemporal cortex of monkeys can be well fitted by an appropriately weighted combination of units sampled from higher layers of deep neural networks that have been trained to recognize objects [22]. Likewise, the response patterns of DNNs to diverse objects have been successfully linked to the patterns of activity found in occipitotemporal cortex of humans [23, 24, 25]. An emerging view in cognitive neuroscience is that DNNs offer a compelling model of human visual function, and that the development of future DNNs will lead to more biologically realistic models of the visual system [18, 26, 27, 28].

However, a potential shortcoming of current DNNs is their tendency to become overspecialized within a narrow range of training conditions, such that they are unable to generalize to new stimuli that are noisy, variable or ambiguous. In particular, there is some evidence to suggest that DNNs for object recognition are unusually susceptible to visual noise and clutter, and that human observers may be better at recognizing objects in noisy viewing conditions [29, 30, 31]. If such a performance gap exists, this may bring into question existing claims that DNNs have attained human-level object recognition performance or that they process visual information in a manner that closely resembles human vision.

The goal of our study was to evaluate the performance of DNNs and human observers when tasked to recognize objects presented at the very limits of perceptual visibility. Object images were presented with varying levels of visual noise by manipulating the signal-to-signal-plus-noise ratio (SSNR), which is bounded between 0 (noise only) and 1.0 (signal only). This allowed us to quantify changes in performance accuracy as a function of SSNR level. We further compared recognition accuracy for objects in pixelated Gaussian noise (i.e., white noise) and spatially correlated Fourier phase-scrambled noise (akin to pink noise) to test for the possibility of qualitative differences in performance between DNNs and human observers (**Figure 1a**).

We found that humans outperform well-known DNNs by a considerable margin, and that standard DNNs are unusually susceptible to pixelated noise, whereas human vision is more severely disrupted by spatially structured noise. However, we go on to show that a noise-training protocol can allow DNNs to acquire considerable robustness to noise, such that they can outperform human observers. Of particular interest, these noise-trained DNNs exhibit more human-like patterns of performance than standard DNNs, in response to different types of noise and to individual images. Moreover, these noise-trained networks exhibit some ability to generalize to real world conditions, allowing for improved classification of vehicles in noisy weather conditions. A network-level analysis indicated that noise training led to widespread changes in the robustness of the network, especially in the middle and higher layers. Our findings suggest that noise-trained DNNs provide a viable model of the noise-robust nature of human visual processing.

Results

We evaluated the performance of 8 pre-trained DNNs (AlexNet, VGG-F, VGG-M, VGG-S, VGG-16, VGG-19, GoogLeNet, and ResNet-152) and 20 human observers at recognizing object images presented in either pixelated Gaussian noise or spatially correlated noise (**Figure 1**, see **Methods**). Performance was assessed using images from 16 object categories (8 animate, 8 inanimate) obtained from the validation data set of the ImageNet database [32]. These images were novel to participants and never used for DNN training.

Figure 2a shows the mean performance accuracy of DNNs and humans plotted as a function of SSNR level, with the performance of individual DNNs shown in **Figure 2b**. Although DNNs could match the performance of human observers under noise-free conditions, consistent with previous reports [17], DNN performance became severely impaired in the presence of moderate levels of noise. Most DNNs exhibited a precipitous drop in recognition accuracy as SSNR declined from 0.6 to 0.4, whereas human performance was much more robust across this range.

Of particular interest, the DNNs appeared to be impaired by noise in a manner that qualitatively differed from human performance. Spatially correlated noise proved more challenging to human observers, whereas the DNNs were more severely impaired by pixelated Gaussian noise (in 7 out of 8 cases). We fitted a logistic function to the performance accuracy data of each participant and each DNN to determine the threshold SSNR level at which performance reached 50% accuracy. This analysis confirmed that human observers outperformed DNNs by a highly significant margin at recognizing objects in pixelated noise ($t(26) = 15.94$, $p < 10^{-14}$) and in Fourier phase-scrambled noise ($t(26) = 12.29$, $p < 10^{-11}$). Moreover, humans showed significantly lower SSNR thresholds for objects in pixelated noise as compared to spatially correlated noise (0.255 vs. 0.315; $t(19) = 13.41$, $p < 10^{-10}$), whereas DNNs showed higher SSNR thresholds for objects in pixelated noise as compared to spatially correlated noise (0.535 vs. 0.446; $t(7) = 3.81$, $p = 0.0066$).

The fact that spatially independent noise proved more disruptive for DNNs was unexpected, given that a simple spatial filtering mechanism, such as averaging over a local spatial window, should allow a recognition system to reduce the impact of spatially independent noise while preserving relevant information about the object. Instead, these DNNs are unable to effectively pool information over larger spatial regions in the presence of pixelated Gaussian noise.

We performed additional analyses to compare the patterns of errors made by DNNs and human observers, plotting confusion matrices for each of four SSNR levels (**Supplementary Figure 1**).

Human performance remained quite robust even at SSNR levels as low as 0.2, as the majority of responses remained correct, falling along the main diagonal. Also, error responses were generally well distributed across the various categories, though there was some degree of clustering and greater confusability occurred among animate categories. In contrast, DNNs were severely impaired by pixelated noise when SSNR declined to 0.5 or lower, and showed a strong bias towards particular categories such as "hare", " cat" and "couch". For objects in spatially correlated noise, the DNNs exhibited a preponderance of errors at SSNR levels of 0.3 and below, with bias towards "hare", "owl" and "cat".

Development and evaluation of a noise-training protocol to improve DNN robustness
We devised a noise-training protocol to determine whether it would be possible to improve the robustness of DNNs to noisy viewing conditions. For these computational investigations, we primarily worked with the VGG-19 network, as this pre-trained network performed quite favorably in comparison to much deeper networks (e.g., GoogLeNet, ResNet-152), and could be trained and evaluated in an efficient manner to evaluate a variety of manipulations.

First, we investigated the effect of training VGG-19 on images from the 16 object categories presented at a single SSNR level with either type of noise. After training, the noise-trained network was tested on a novel set of object images presented with the corresponding noise type across a full range of SSNR levels. We observed that training the DNN at a progressively lower SSNR level led to a consistent leftward shift of the recognition accuracy by SSNR curve (**Figure 3a**). However, this improvement in performance for noisy images was accompanied by a loss of performance accuracy for noise-free images. The latter was evident from the prominent downward shift in the recognition accuracy by SSNR curve. Such loss of accuracy for noise-free images would be unacceptable for any practical applications of this noise-training procedure, and clearly deviated from human performance.

Next, we investigated whether robust performance across a wide range of SSNR levels might be attained by providing intermixed training with both noise-free and noisy images. **Figure 3b** indicates that such combined training was highly successful, with the strongest improvement observed for noisy images presented at challenging SSNR level of 0.2. When the training SSNR was reduced to levels as low as 0.1, the task became too difficult and the learning process suffered.

Given the excellent performance of VGG-19 after training with images at 0.2 and 1.0 SSNR, we sought to compare these noise-trained DNNs with human performance. **Figure 4a** shows that noise-trained VGG-19 performed far better than standard DNNs at recognizing objects in either type of noise. Moreover, the noise-trained networks now showed an advantage at recognizing objects in pixelated Gaussian noise as compared to spatially correlated noise, in a manner that better matched the qualitative performance of human observers. It was also striking that the noise-trained networks appeared to perform better than the human observers on average.

To analyze these performance differences in detail, we fitted a logistic function to identify the SSNR thresholds of each DNN and human observer, separately for each noise condition. A histogram of SSNR thresholds revealed that noise-trained VGG-19 outperformed all 20 human observers and all 8 original DNNs at recognizing objects in noise (**Figure 4b**). These results indicate that the noise-training protocol can greatly enhance the robustness of DNNs, such that they can match or surpass human performance when tasked to recognize objects in extreme levels of visual noise.

Although the benefits of noise training were specific to the noise type encountered during training (**Supplementary Figure 2**), we found that it was possible to train a single DNN to acquire robustness to both pixelated Gaussian noise and spatially structured noise concurrently (**Supplementary Figure 3**). Likewise, we confirmed that other networks (e.g., ResNet-152) showed similar improvements in robustness after noise training with these 16 object categories (**Supplementary Figure 4**). We also evaluated the impact of training VGG-19 on the full 1000-category image set from ImageNet with both types of noise, and found that the network was capable of recognizing objects in noise when discerning among a large number of possible categories (**Supplementary Figure 5**).

Using the 1000-category noise-trained VGG-19, we developed another test to determine whether noise training might improve the robustness of DNNs at recognizing novel real-world examples of objects in noisy weather conditions (e.g., snow, rain, fog). We evaluated classification performance for 8 types of vehicles in the ImageNet data set. A web-based search protocol was used to gather candidate images of vehicles in noisy weather conditions, and test images were selected based on the ratings of three independent observers. The final test set consisted of 102 noisy vehicle images and 102 noise-free vehicle images (see **Supplementary Figure 6** for examples). **Figure 5a** shows that both standard and noise-trained versions of VGG-19 performed equally well at recognizing noise-free images of vehicles. In contrast, noise-trained VGG-19 outperformed the standard DNN at recognizing vehicles in noisy weather conditions. Performance was further analyzed according to the human-rated noise level of individual vehicle images. This analysis indicated that noise-trained VGG-19 performed significantly better with images rated as containing moderate or strong noise (**Figure 5b**). Our findings indicate that DNNs trained on images with artificially generated noise can successfully generalize, to some degree, to real-world examples of noisy viewing conditions.

Characterizing network changes caused by noise-training

To identify the stages of DNN processing that are most affected by noise training, we devised a layer-specific noise sensitivity analysis. Specifically, we calculated the correlation strength between the layer-specific pattern of activity evoked by a noise-free image and the pattern of activity evoked by that same image when presented at varying SSNR levels (**Figure 6a**). Correlation strength will monotonically increase with increasing SSNR level (from an expected value of 0 to 1.0), and the threshold SSNR level needed to reach a correlation value of 0.5 can be identified. Here, a lower threshold SSNR indicates greater robustness, whereas a higher threshold SSNR indicates greater noise susceptibility. As can be seen in **Figure 6b**, the standard VGG-19 network exhibits a gradual increase in noise susceptibility in progressively higher layers, implying that the contaminating effects of visual noise tend to become amplified across successive stages of feedforward processing.

After the noise-training protocol, however, the network shows considerable improvement, especially in the middle and higher layers where the difference between noise-trained and standard networks most clearly diverges. For pixelated Gaussian noise, the responses of the noise-trained DNN actually become more robust across successive stages of processing. In effect, the convolutional processing that occurs across successive stages of the noise-trained network leads to a type of de-noising process. This finding is consistent with the notion that the disruptive impact of spatially independent noise can be curtailed if signals over progressively larger spatial regions are pooled together in an appropriate manner to dampen the impact of random, spatially independent noise. This can be contrasted with the results found for spatially correlated noise. Here, the threshold SSNR level appears quite stable across successive layers for the noise-trained network, and noise-training allows the network to avoid an increase in

noise susceptibility as information is passed from one layer to the next, in contrast with standard network.

As a complementary analysis, we measured classification-based SSNR thresholds by applying a multi-class support vector machine (SVM) classifier to the activity patterns of each layer of a given network. Each SVM was trained on activity patterns evoked by noise-free training images, and then tested on its ability to predict the object category of test stimuli presented at varying SSNR levels. The SSNR level at which classification accuracy reached 50% was identified as the classification-based SSNR threshold. For standard DNNs, we found that classification accuracy for noise-free test images gradually improved across successive layers of the network (**Supplementary Figure 7**), and this trend largely accounted the improvement (i.e., decrease) in the classification-based SSNR threshold in the middle and higher layers (**Figure 6c**). Of greater interest, the divergence between standard and noise-trained networks became more pronounced in the middle and higher layers due to the benefits of noise training. These results provide further evidence that noise robustness is largely achieved through the modification of learned representations in the middle and higher layers of the noise-trained network.

Comparison of human and DNN performance for individual object images

It has been suggested that DNNs represent visual object information in a manner that strongly resembles the human visual system [18, 23, 24, 25]. However, we found that standard DNNs process noisy object images in a qualitatively different manner than human observers (e.g., **Figure 2**). We devised a follow-up behavioral experiment to test whether noise-trained DNNs can provide a suitable model of human performance and predict people's recognition thresholds on an image-by-image basis.

Twenty observers were shown each of 800 object images (50 per category), which slowly emerged from pixelated Gaussian noise. The SSNR level gradually increased from an initial value of 0 in small steps of 0.025 every 400ms, until the observer pressed a key to pause the dynamic display in order to make a categorization decision. A reward-based payment scheme provided greater reward for correct responses made at lower SSNR levels. After the categorization response, participants used a mouse pointer to demarcate the portions of the image that they relied on for their recognition judgment.

The resulting data allowed us to compare the similarity of humans and DNNs in their SSNR thresholds, as well as the portions of each image that were diagnostic for recognition judgments. Mean performance accuracy was high (90.3%), and human SSNR thresholds for each image were calculated based on responses for correct trials only. Accordingly, SSNR thresholds were calculated for standard and noise-trained VGG-19 by requiring accuracy to reach 90%. As can be seen in the scatterplot in **Figure 7,** noise-trained DNNs provided a much better fit of human SSNR thresholds for individual images ($r = 0.55$, slope = 0.67) than standard DNNs ($r = 0.27$, slope = 0.32). That said, it should be noted that human-to-human similarity was greater still (mean $r = 0.94$, based on a split-half correlation analysis), indicating that further improvements can be made by future DNN models to account for human recognition performance.

To complement the diagnostic regions reported by human observers, we used layer-wise relevance propagation [33] to identify what portions of each image were important for the decisions of DNNs (see **Figure 8a**). We calculated the spatial correlation and amount of overlap between diagnostic regions reported by humans and those used by trained DNNs for their classification responses across a range of SSNR levels. Both standard and noise-trained DNNs performed reasonably well at predicting the diagnostic regions used by human observers under

noise-free conditions (**Figure 8b**). However, only the noise-trained DNN could reliably predict the diagnostic regions used by human observers in noisy viewing conditions.

Taken together, our findings indicate that noise-trained DNNs provide a superior model to account for people's ability to recognize objects in severely degraded viewing conditions. Moreover, their ability to generalize to real world conditions of visual noise suggests that noise-trained DNNs can successfully acquire a certain degree of the robustness that is exemplified by human vision.

Discussion

In this study, we evaluated whether DNNs can provide a viable model of the robust nature of human vision. This was done by tasking DNNs and humans to recognize objects in visual noise, often presented at the threshold of visibility. We found that standard DNNs lack robustness; moreover, their performance qualitatively differed from that of humans, as human observers find spatially correlated noise to be more disruptive than pixelated Gaussian noise. In comparison, our noise-trained DNNs provided a better qualitative match to human performance, and indeed, could outperform human observers by a modest but highly reliable margin. Moreover, noise training allowed DNNs to acquire better performance at recognizing novel images of vehicles in noisy weather. Thus, augmented training with artificial noise can support some generalization to real world examples of noisy viewing conditions.

The ability to predict human recognition performance at the individual image level remains a major challenge in neuroscience [27]. We showed that noise-trained DNNs provide a better model than standard DNNs at predicting SSNR thresholds for human recognition on an image-by-image basis. Likewise, the noise-trained DNNs relied on diagnostic regions that overlapped to a significant degree with those reported by human observers. Although there is certainly room for further improvement in the fitting of human performance data, we find that noise-trained DNNs provide a compelling model to account for human recognition of objects in noisy viewing conditions.

A network-level analysis revealed that noise training led to widespread changes in the robustness of the network, especially in the middle and higher layers. With respect to spatially uncorrelated noise, visual representations gradually became more noise-robust across successive stages of processing, akin to a hierarchical denoising process, with the greatest benefit observed at high levels of the network. These findings deviate from traditional notions of image processing, which typically rely on the modification of low-level visual filters to achieve noise filtering [34]. Our findings suggest that robustness to visual noise is acquired, at least in part, through learning and experience, with extensive modifications that take place at higher stages of visual processing. As a consequence, DNNs that are trained with challenging noisy images may acquire visual representations that better approximate human vision.

Previous studies have documented that the object recognition performance of DNNs can be disrupted by adding non-random adversarial noise [35, 36] and also by adding randomly generated pixelated noise to object images [29, 30, 31]. A few recent studies have reported some improvement in recognition performance by training DNNs on certain types of noisy object images but these studies did not establish strong links between DNN and human performance [29, 30, 31]. Concurrent with our own work [37], one study investigated the training of DNNs with various types of noise (i.e., uniform, salt-and-pepper) and image distortion (e.g., blur), and reported that the benefits of training were highly specific and failed to generalize to new conditions [31]. Indeed, the

researchers used the same set of object images across training and test, whereas in the current study, we used novel test images to ensure that any improvements in performance required a certain degree of generalization by the network. Although we found that training with either pixelated Gaussian noise or Fourier phase-scrambled noise led to noise-specific improvement, we also found the DNNs have the capacity to acquire robustness to both types of noise concurrently. An ability to generalize from artificial noise to real world conditions was further demonstrated by our evaluation of DNN performance with vehicles in noisy weather.

The main focus of this study was to determine whether DNNs, with suitable training, can provide a viable model of the noise-robust nature of human vision. Our findings suggest that this is indeed the case, thereby supporting our neuroscientific goals. In addition, our procedure for training deep neural networks to acquire robustness to noisy viewing conditions may have relevance for applications in computer vision and artificial intelligence, including the development of autonomous vehicles, visually guided robots, and the analysis of real world images with adverse viewing conditions.

In future work, it will be of considerable interest for researchers to investigate whether modifications of network architecture, such as the incorporation of lateral interactions, top-down feedback or recurrent connections [28, 38, 39], can provide further benefits to the robust performance of DNNs when tasked to recognize objects in noise. While it is conceivable that a DNN with modified architecture might prove somewhat more robust to noise after training exclusively with objects in clear viewing conditions, we suspect that some form of noise training will be necessary for DNNs to acquire a high degree of robustness, at least comparable to what we have shown here. Our noise-trained DNN outperformed each of 20 human observers in terms of SSNR threshold, indicating a marked benefit of DNN noise training. Although we cannot say whether humans or noise-trained DNNs are close to optimal at recognizing objects in noise, we do know that there must be an upper limit of performance as SSNR levels gradually decline towards a value of 0, at which point no object information will remain present in the image. Along these lines, our protocol for measuring SSNR thresholds for recognition performance may provide a useful metric for evaluating the performance of future DNNs. The SSNR threshold provides a measure of the limit at which a recognition system can still classify or identify an object, and avoids potential concerns of near-ceiling performance that can be more readily achieved for objects in clear viewing conditions. Moreover, performance for individual object images can be rigorously quantified, as a multitude of randomly generated noise images can be combined with any source image to evaluate performance accuracy, with confidence bounds, across a range of SSNR levels. Such methodology can support a rigorous evaluation of the correspondence between human and machine vision.

Methods

Participants

We recruited 20 participants for Experiment 1 (15 females, 5 males) and another group of 20 participants for Experiment 2 (12 females, 8 males); ages ranged from 19 to 33 years old. All participants reported having normal or corrected-to-normal visual acuity, and provided informed consent. The study was approved by the Institutional Review Board of Vanderbilt University. Participants were compensated monetarily or through a combination of course credit and monetary payment.

Visual stimuli

Object images were obtained from the ImageNet database [32], 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 1B**). Both humans and DNNs 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. DNNs were trained using images from the training set (1300 images per category), so the images used for testing were novel to both humans and DNNs.

In Experiment 1, objects were presented using two different types of visual noise: pixelated Gaussian noise and Fourier phase-scrambled noise (**Figure 1A**). To create each Gaussian noise image, the intensity of every pixel was randomly and independently drawn from a Gaussian distribution, 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.

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 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:

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

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. All 20 participants completed both sessions.

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.

Convolutional neural networks

We evaluated the performance of 8 pre-trained convolutional neural networks using the MatConvNet toolbox [40]: AlexNet, VGG-F, VGG-M, VGG-S, VGG-16, VGG-19, GoogLeNet, and ResNet-152 [13, 14, 15, 16]. The training of CNNs with noisy object images was primarily performed using MatConvNet, with ancillary analyses performed using PyTorch. 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. We initially evaluated pre-trained VGG-19 by training the network with noisy object images presented at a single SSNR level (**Figure 3A**), 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.

For 16-category training, the DNNs 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. 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 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.

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 across 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. The CNN's decision was determined based on the highest softmax value found the classification layer for the 16 categories that were tested. 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 sensitivity 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 DNN. 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.

References

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

2. DiCarlo JJ, Zoccolan D, Rust NC. How does the brain solve visual object recognition? *Neuron* **73**, 415-434 (2012).

3. Li FF, VanRullen R, Koch C, Perona P. Rapid natural scene categorization in the near absence of attention. *Proc Natl Acad Sci U S A* **99**, 9596-9601 (2002).

4. Potter MC. Meaning in visual search. *Science* **187**, 965-966 (1975).

5. Tong F, Nakayama K. Robust representations for faces: evidence from visual search. *J Exp Psychol Hum Percept Perform* **25**, 1016-1035. (1999).

6. Land MF. Eye movements and the control of actions in everyday life. *Prog Retin Eye Res* **25**, 296-324 (2006).

7. Krupinski EA. Current perspectives in medical image perception. *Atten Percept Psycho* **72**, 1205-1217 (2010).

8. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* **521**, 436-444 (2015).

9. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. . *Journal of Physiology* **160**, 106-154 (1962).

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

11. LeCun Y*, et al.* Backpropagation applied to handwritten zip code recognition. *Neural computation* **1**, 541-551 (1989).

12. Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. *Nat Neurosci* **2**, 1019-1025 (1999).

13. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems* (2012).

14. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:14091556*, (2014).

15. Szegedy C*, et al.* Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015).

16. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Computer Vision and Pattern Recognition*. IEEE (2016).

17. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *IEEE International Conference on Computer Vision* (2015).

18. Yamins DL, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci* **19**, 356-365 (2016).

19. Mnih V*, et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529-533 (2015).

20. Esteva A*, et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115-118 (2017).

21. Majkowska A*, et al.* Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-adjudicated Reference Standards and Population-adjusted Evaluation. *Radiology* **294**, 421-431 (2020).

22. Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci U S A* **111**, 8619-8624 (2014).

23. Khaligh-Razavi SM, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput Biol* **10**, e1003915 (2014).

24. Guclu U, van Gerven MA. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *J Neurosci* **35**, 10005-10014 (2015).

25. Horikawa T, Kamitani Y. Generic decoding of seen and imagined objects using hierarchical visual features. *Nat Commun* **8**, 15037 (2017).

26. Kriegeskorte N. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annu Rev Vis Sci* **1**, 417-446 (2015).

27. Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ. Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *J Neurosci* **38**, 7255-7269 (2018).

28. Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat Neurosci* **22**, 974-983 (2019).

29. Rodner E, Simon M, Fisher RB, Denzler J. Fine-grained recognition in the noisy wild: Sensitivity analysis of convolutional neural networks approaches. In: *British Machine Vision Conference* (2016).

30. Dodge S, Karam L. A study and comparison of human and deep learning recognition performance under visual distortions. In: *International Conference on Computer Communications and Networks* (2017).

31. Geirhos R, Medina Temme CR, Rauber J, Schutt HH, Bethge M, Wichmann FA. Generalisation in humans and deep neural networks. In: *Neural Information Processing Systems* (2018).

32. Russakovsky O*, et al.* Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* **115**, 211-252 (2015).

33. Bach S, Binder A, Montavon G, Klauschen F, Muller KR, Samek W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS One* **10**, e0130140 (2015).

34. Buades A, Coll B, Morel JM. A review of image denoising algorithms, with a new one. *Multiscale Modeling and Simulation* **4**, 490-530 (2005).

35. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. (2014).

36. Szegedy C*, et al.* Intriguing properties of neural networks (2014).

37. Jang H, Tong F. Can deep learning networks acquire the robustness of human recognition when faced with objects in visual noise? In: *Vision Sciences Society*. Journal of Vision (2018).

38. Kietzmann TC, Spoerer CJ, Sorensen LKA, Cichy RM, Hauk O, Kriegeskorte N. Recurrence is required to capture the representational dynamics of the human visual system. *Proc Natl Acad Sci U S A* **116**, 21854-21863 (2019).

39. Tang H*, et al.* Recurrent computations for visual pattern completion. *Proc Natl Acad Sci U S A* **115**, 8835-8840 (2018).

40. Vedaldi A, Lenc K. Matconvnet: Convolutional neural networks for matlab. In: *Proceedings of the 23rd ACM international conference on Multimedia*. ACM (2015).

Acknowledgements

The authors would like to thank Malerie McDowell, Echo Sun, Feyisayo Adegboye and Haley Frey for technical assistance. This research was supported by a grant from the National Eye Institute and a Vanderbilt Discovery Grant to FT, and a core grant from the National Eye Institute to the Vanderbilt Vision Research Center (Director David Calkins).

Author Contributions

FT devised the study, experimental design, protocol for noise training of DNNs, and layer-specific noise sensitivity analysis. DM coded the first behavioral experiment. HJ coded the DNN networks, analyzed both the human and DNN data, and devised the layer-specific classification analysis. FT and HJ wrote the paper.

Competing Interests

The authors (FT and HJ) have submitted a U.S. patent application with respect to the noise-training methods used in this study to train deep neural networks.

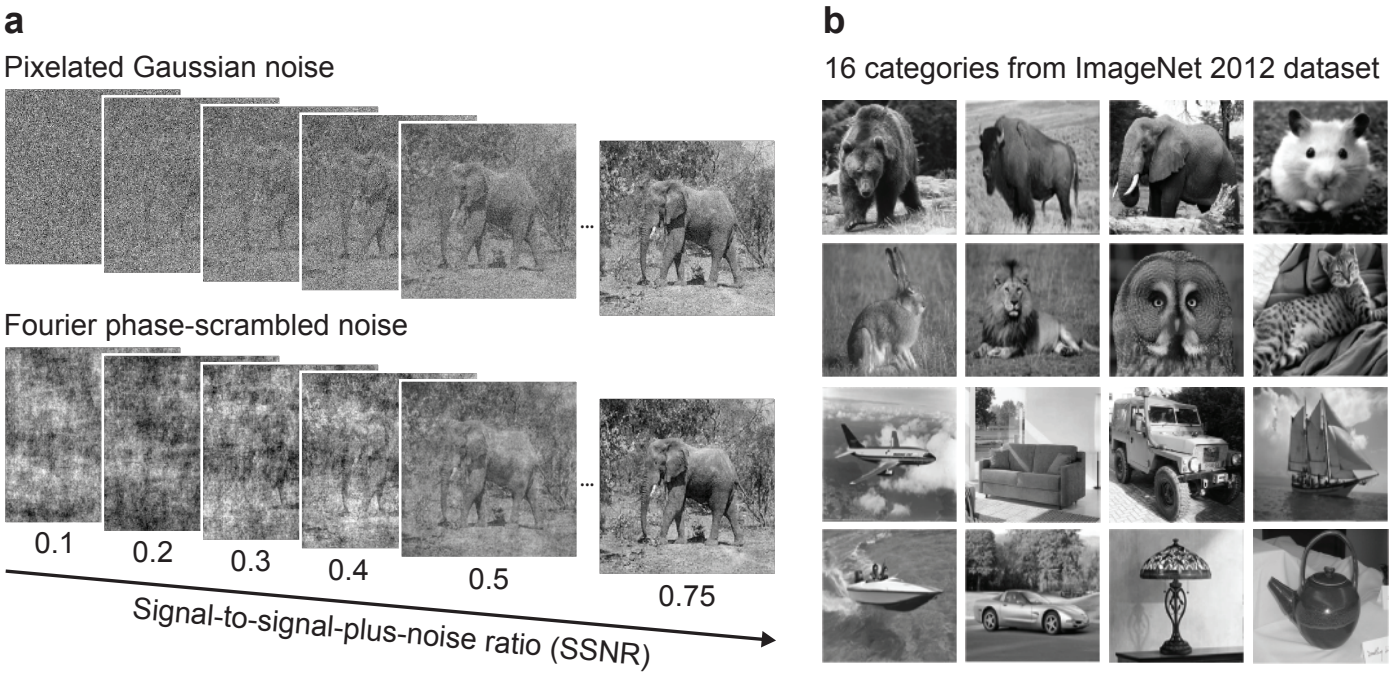

Figure 1. a Examples of an object image in pixelated Gaussian noise or Fourier phase-scrambled noise, shown at varying SSNR levels. **b** Example images from the 16 object categories used in this study: bear, bison, elephant, hamster, hare, lion, owl, tabby cat, airliner, couch, jeep, schooner, speedboat, sports car, table lamp, teapot.

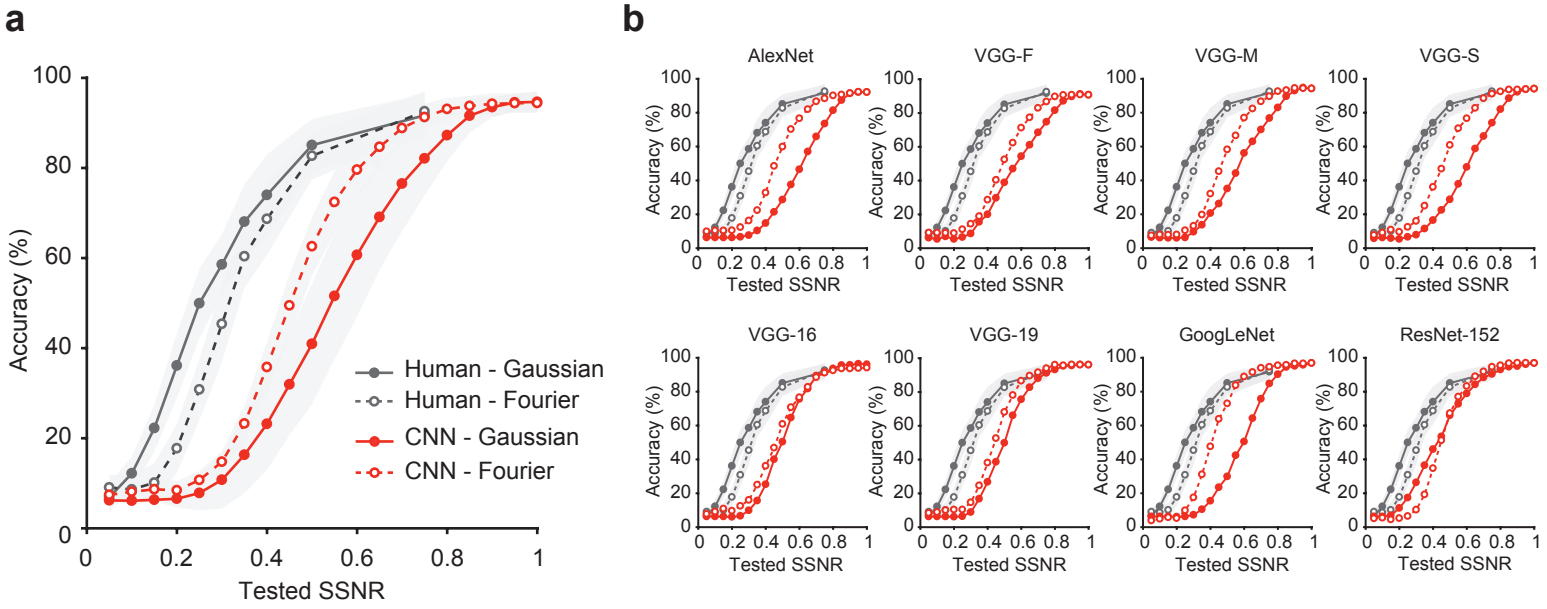

Figure 2. a Mean performance accuracy in a 16-alternative object classification task plotted as a function of SSNR level for human observers (black curves) and 8 standard pre-trained CNNs (red curves) with ± 1 standard deviation in performance indicated by the shaded area around each curve. Separate curves are plotted for pixelated Gaussian noise (solid lines with closed circles) and Fourier phase-scrambled noise (dashed lines with open circles). **b** Classification accuracy plotted as a function of SSNR level for individual pre-trained CNN models.

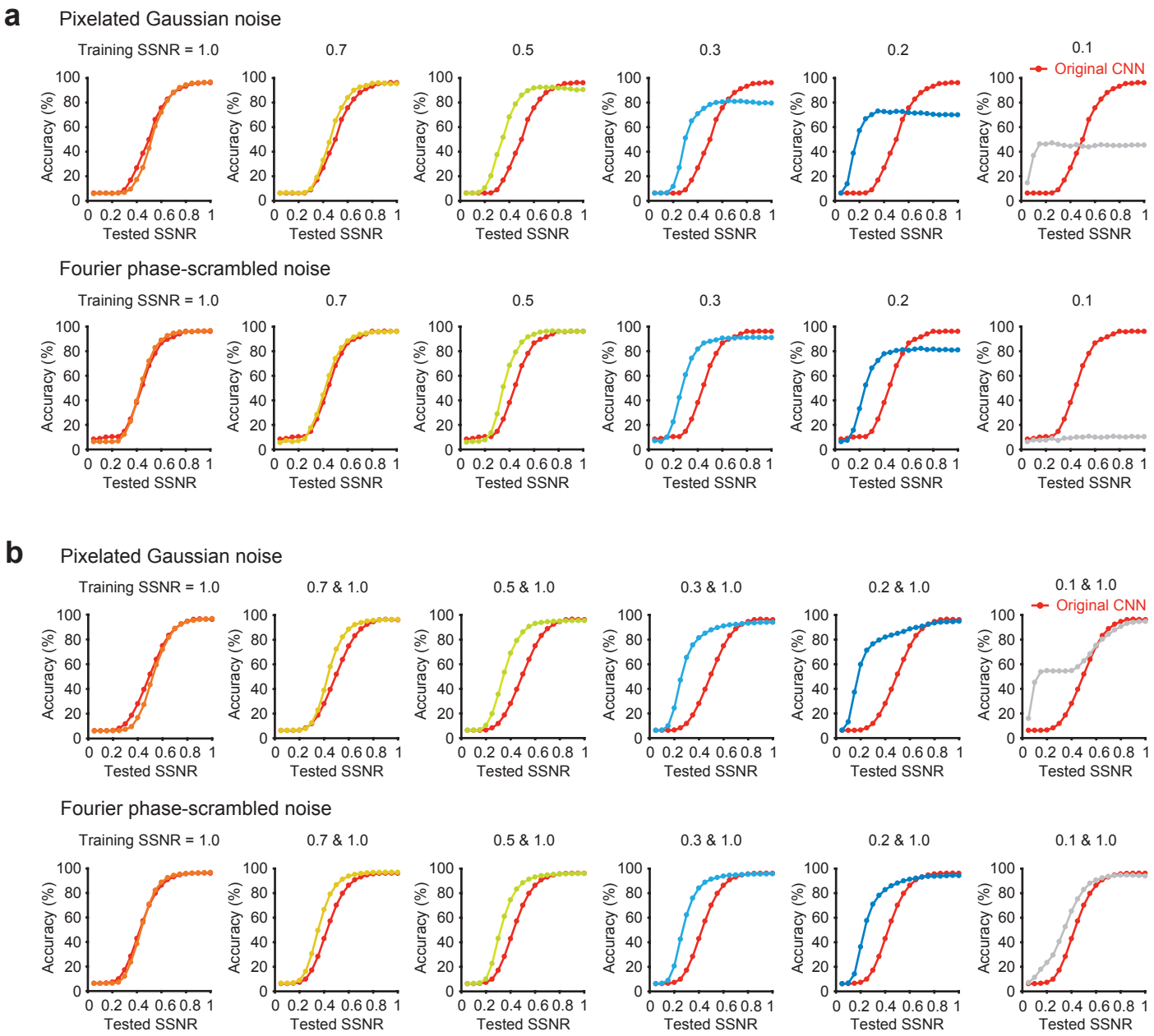

Figure 3. a Impact of training VGG-19 with object images presented at a single SSNR level (1.0, 0.7, 0.5, 0.3, 0.2, or 0.1) when evaluated with novel test images presented at multiple SSNR levels. Accuracy of standard VGG-19 (red curve) serves as a reference in each plot. **b** Impact of training VGG-19 with a combination of noise-free images (SSNR 1.0) and noisy images at a specified SSNR level.

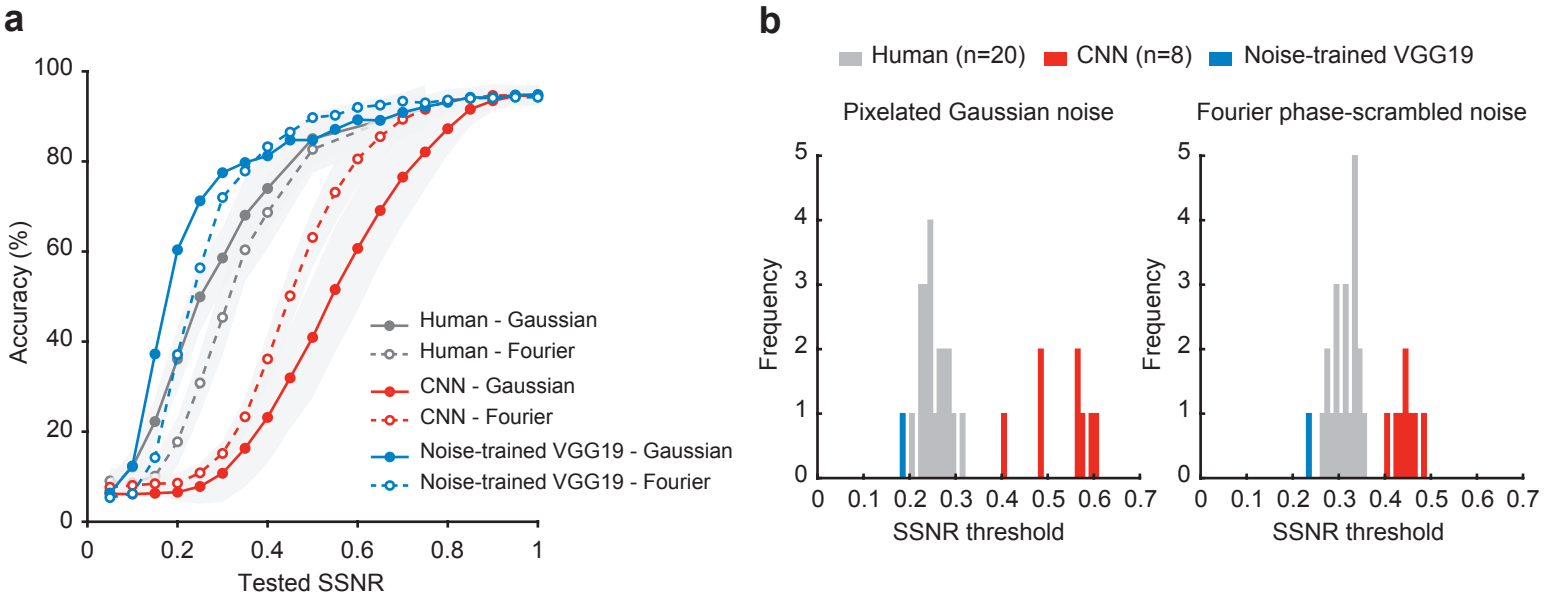

Figure 4. a Mean classification accuracy of noise-trained VGG-19 (blue), human observers (gray), and standard CNNs (red) for objects in pixelated Gaussian noise (solid lines, closed circles) and Fourier phase-scrambled noise (dashed lines, open circles). **b** Frequency histograms comparing the SSNR thresholds of noise-trained VGG-19 (blue), individual human observers (gray), and 8 standard pre-trained CNNs (red).

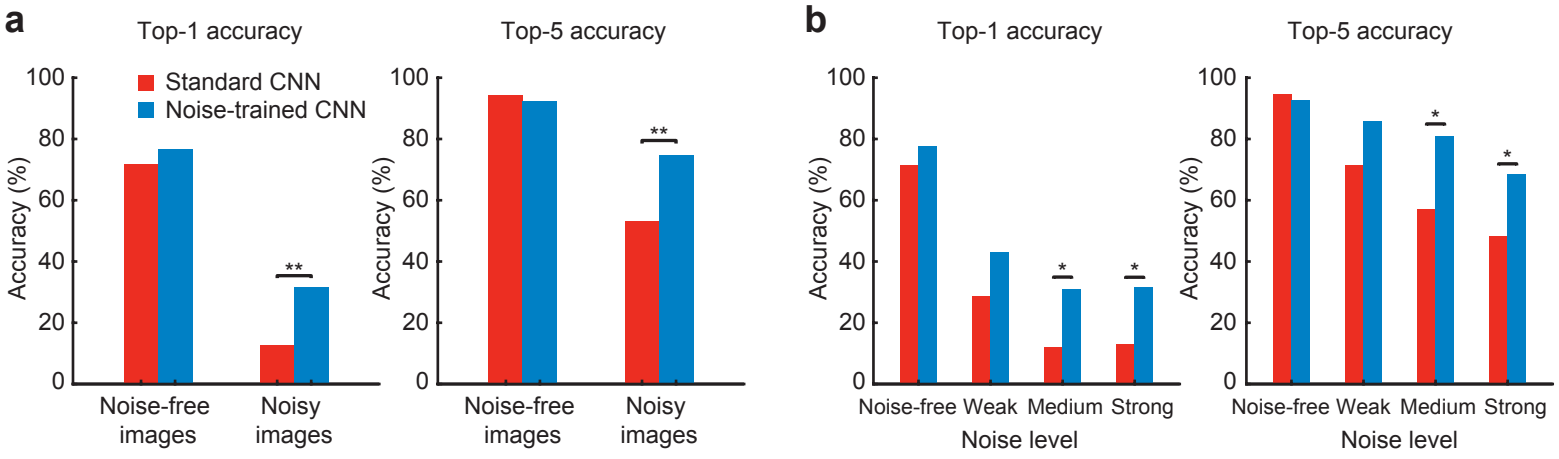

Figure 5. a Top1 and top5 accuracies of standard VGG-19 (red) and noise-trained VGG-19 (blue) at classifying vehicles in noise-free or noisy weather conditions. Noise-trained VGG-19 outperformed standard VGG-19 at recognizing noisy vehicle images (top1 accuracy, $\chi 2 = 10.29$, $p = .0013$; $\chi 2 = 10.26$, $p = .0014$). **b** 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 ($\chi 2 > 4.5$, $p < 0.05$ in all cases). Asterisks indicate * $p < .05$, ** $p < .01$.

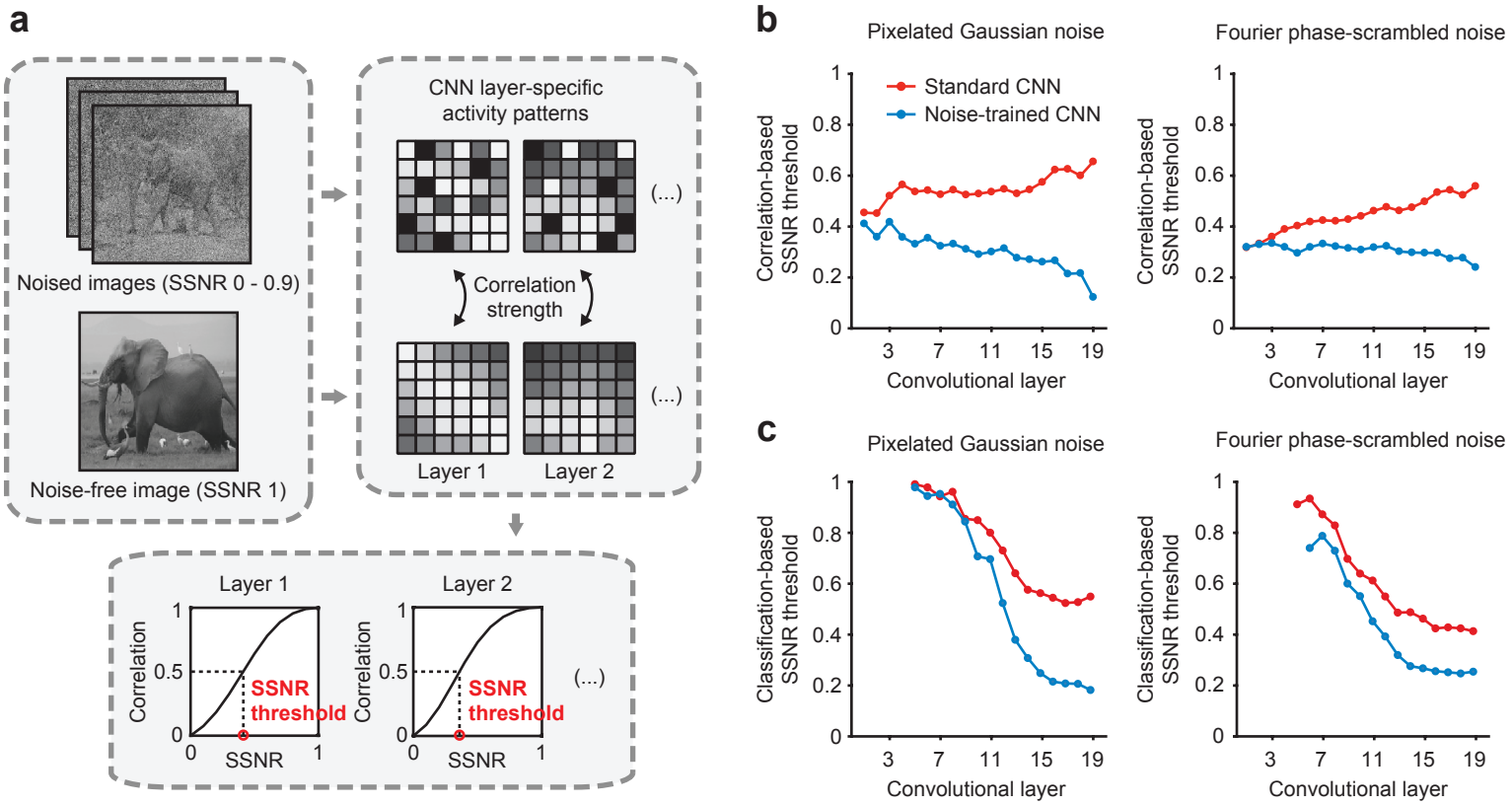

Figure 6. a Depiction of the method used for layer-specific noise sensitivity analysis. **b** Correlation-based SSNR thresholds for every convolutional layer (including fully-connected layers and the last classification layer) were measured for object images in pixelated Gaussian noise and Fourier phase-scrambled noise. The thresholds of standard (red) and noise-trained (blue) VGG-19 were compared. **c** Classification-based SSNR thresholds plotted by layer for standard and noise-trained networks. Multi-class support vector machines were used to predict object category from layer-specific activity patterns.

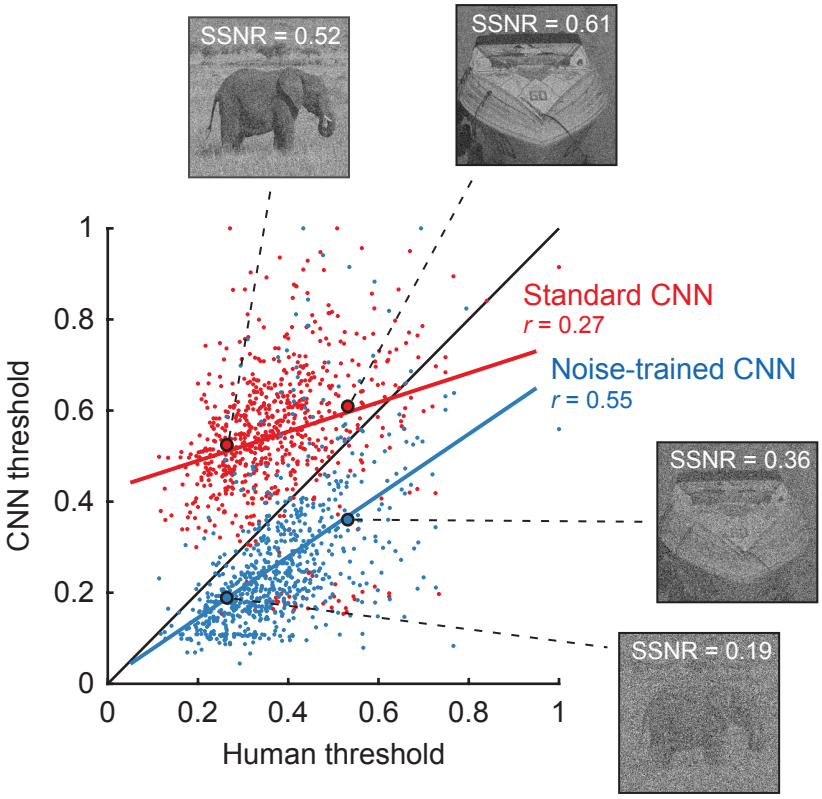

Figure 7. Scatter plot comparing SSNR thresholds of human observers with the thresholds of standard VGG-19 (red) and noise-trained VGG-19 (blue). Each data point depicts SSNR thresholds for an individual object image. Examples of two object images, shown at the SSNR threshold obtained from standard or noise-trained networks.

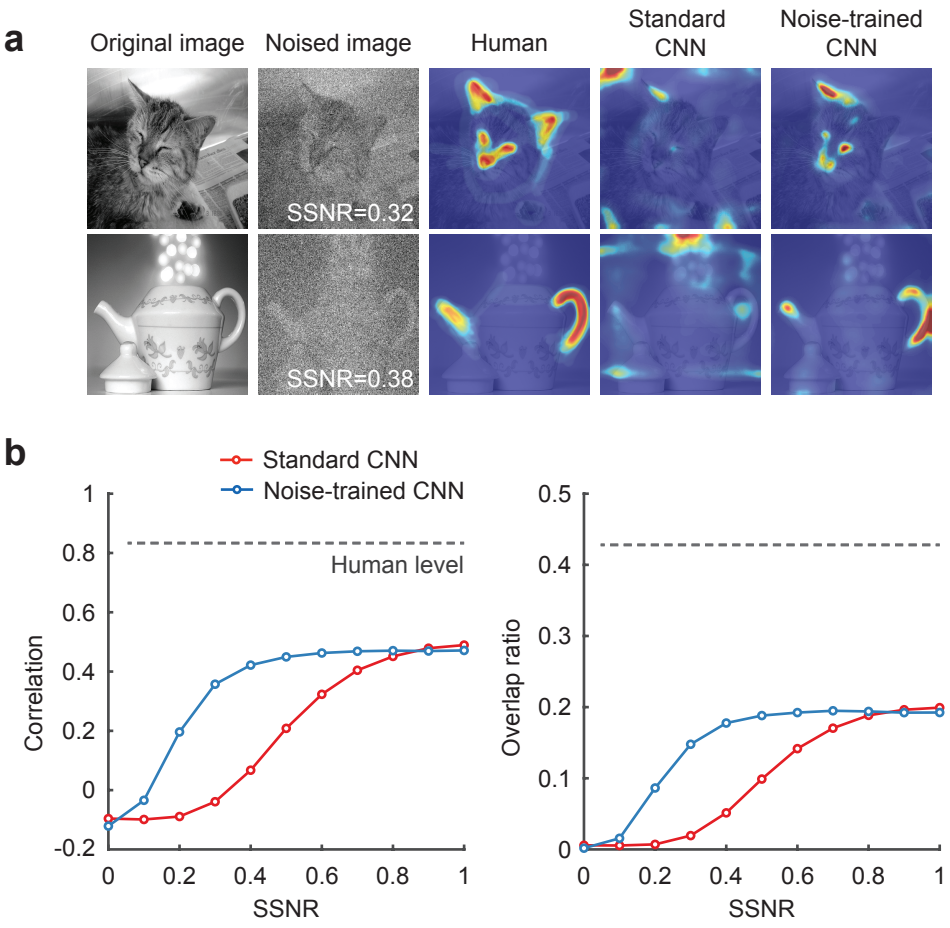

Figure 8. a Examples of diagnostic object features from human observers, standard VGG-19, and noise-trained VGG-19. The mean SSNR level at which human observers correctly recognized the objects is indicated. **b** Correlational similarity and overlap ratio of the spatial profile of diagnostic features reported by human observers and those measured in CNNs across a range of SSNR levels. Gray dashed lines indicate ceiling-level performance based on human-to-human correspondence.