frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 05 August 2021
doi: 10.3389/fninf.2021.677925

# Dissociable Neural Representations of Adversarially Perturbed Images in Convolutional Neural Networks and the Human Brain

*Chi Zhang[1], Xiao-Han Duan[1], Lin-Yuan Wang[1], Yong-Li Li[2], Bin Yan[1], Guo-En Hu[1], Ru-Yuan Zhang[3,4]*† and Li Tong[1]*†*

*[1] Henan Key Laboratory of Imaging and Intelligent Processing, PLA Strategic Support Force Information Engineering University, Zhengzhou, China, [2] People's Hospital of Henan Province, Zhengzhou, China, [3] Institute of Psychology and Behavioral Science, Shanghai Jiao Tong University, Shanghai, China, [4] Shanghai Key Laboratory of Psychotic Disorders, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China*

Despite the remarkable similarities between convolutional neural networks (CNN) and the human brain, CNNs still fall behind humans in many visual tasks, indicating that there still exist considerable differences between the two systems. Here, we leverage adversarial noise (AN) and adversarial interference (AI) images to quantify the consistency between neural representations and perceptual outcomes in the two systems. Humans can successfully recognize AI images as the same categories as their corresponding regular images but perceive AN images as meaningless noise. In contrast, CNNs can recognize AN images similar as corresponding regular images but classify AI images into wrong categories with surprisingly high confidence. We use functional magnetic resonance imaging to measure brain activity evoked by regular and adversarial images in the human brain, and compare it to the activity of artificial neurons in a prototypical CNN—AlexNet. In the human brain, we find that the representational similarity between regular and adversarial images largely echoes their perceptual similarity in all early visual areas. In AlexNet, however, the neural representations of adversarial images are inconsistent with network outputs in all intermediate processing layers, providing no neural foundations for the similarities at the perceptual level. Furthermore, we show that voxel-encoding models trained on regular images can successfully generalize to the neural responses to AI images but not AN images. These remarkable differences between the human brain and AlexNet in representation-perception association suggest that future CNNs should emulate both behavior and the internal neural presentations of the human brain.

**Keywords: adversarial images, convolutional neural network, human visual cortex, functional magnetic resonance imaging, representational similarity analysis, forward encoding model**

OPEN ACCESS

***Edited by:***
*Tolga Cukur,*
*Bilkent University, Turkey*

***Reviewed by:***
*Shinji Nishimoto,*
*Osaka University, Japan*
*Mo Shahdloo,*
*University of Oxford, United Kingdom*

***Correspondence:**
*Ru-Yuan Zhang*
*ruyuanzhang@sjtu.edu.cn*
*Li Tong*
*tttocean_tl@hotmail.com*
*† These authors share senior authorship*

***Received:*** *08 March 2021*
***Accepted:*** *28 June 2021*
***Published:*** *05 August 2021*

***Citation:***
*Zhang C, Duan X-H, Wang L-Y, Li Y-L, Yan B, Hu G-E, Zhang R-Y and Tong L (2021) Dissociable Neural Representations of Adversarially Perturbed Images in Convolutional Neural Networks and the Human Brain. Front. Neuroinform. 15:677925. doi: 10.3389/fninf.2021.677925*

## INTRODUCTION

The recent success of convolutional neural networks (CNNs) in many computer vision tasks inspire neuroscientists to consider them as a ubiquitous computational framework to understand biological vision (Jozwik et al., 2016; Yamins and DiCarlo, 2016). Indeed, a bulk of recent studies have demonstrated that visual features in CNNs can accurately predict many spatiotemporal

characteristics of brain activity (Agrawal et al., 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015, 2017; Cichy et al., 2016; Hong et al., 2016; Horikawa and Kamitani, 2017; Khaligh-Razavi et al., 2017). These findings reinforce the view that modern CNNs and the human brain share many key structural and functional substrates (LeCun et al., 2015).

Despite the tremendous progress, current CNNs still fall short in several visual tasks. These disadvantages suggest that critical limitations still exist in modern CNNs (Grill-Spector and Malach, 2004). One potent example is adversarially perturbed images, a class of images that can successfully "fool" even the most state-of-the-art CNNs (Szegedy et al., 2013; Nguyen et al., 2015). Adversarial noise (AN) images (**Figure 1B**) look like meaningless noise to humans but can be classified by CNNs into familiar object categories with surprisingly high confidence (Nguyen et al., 2015). Adversarial interference (AI) images are generated by adding a small amount of special noise to regular images (**Figure 1C**). The special noise looks minimal to humans but severely impairs CNNs' recognition performance (Szegedy et al., 2013). *Perception* here can be operationally defined as the output labels of a CNN and object categories reported by humans. Therefore, adversarial images present a compelling example of double-dissociation between CNNs and the human brain, because artificially created images can selectively alter perception in one system without significantly impacting the other one.

It remains unclear the neural mechanisms underlying the drastically different visual behavior between CNNs and the human brain with respect to adversarial images. In particular, why do the two systems receive similar stimulus inputs but generate distinct perceptual outcomes? In the human brain, it has been known that the neural representations in low-level visual areas mostly reflect stimulus attributes whereas the neural representations in high-level visual areas mostly reflect perceptual outcomes (Grill-Spector and Malach, 2004; Wandell et al., 2007). For example, the neural representational similarity in human inferior temporal cortex is highly consistent with perceived object semantic similarity (Kriegeskorte et al., 2008). In other words, there exists a well-established representation-perception association in the human brain.

This processing hierarchy is also a key feature of modern CNNs. If the representational architecture in CNNs truly resembles the human brain, we should expect similar neural substrates supporting CNNs' "perception." For CNNs, AI images and regular images are more similar at the pixel level but yield different perceptual outcomes. By contrast, AN images and regular images are more similar at the "perceptual" level. We would expect that AI and regular images have more similar neural representations in low-level layers while AN and regular images have similar neural representations in high-level layers. In other words, there must exist at least one high-level representational layer that supports the same categorical perception of AN and regular images, similar to the representation-perception association in the human brain. However, as we will show later in this paper, we find no representational pattern that supports RE-AN perceptual similarity in all intermediate representation layers except the output layer.

The majority of prior studies focused on revealing similarities between CNNs and the human brain. In this paper, we instead leverage adversarial images to examine the differences between the two systems. We particularly emphasize that delineating the differences here does not mean to object CNNs as a useful computational framework for human vision. On the contrary, we acknowledge the promising utilities of CNNs in modeling biological vision but we believe it is more valuable to understand differences rather than similarities such that we are in a better position to eliminate these discrepancies and construct truly brain-like machines. In this study, we use a well-established CNN—AlexNet and investigate the activity of artificial neurons toward adversarial images and their corresponding regular images. We also use functional magnetic resonance imaging (fMRI) to measure the cortical responses evoked by RE and adversarial images in humans. Representational similarity analysis (RSA) and forward encoding modeling allow us to directly contrast representational geometries within and across systems to understand the capacity and limit of both systems.

# MATERIALS AND METHODS

## Ethics Statement

All experimental protocols were approved by the Ethics Committee of the Henan Provincial People's Hospital. All research was performed in accordance with relevant guidelines and regulations. Informed written consent was obtained from all participants.

## Subjects

Three healthy volunteers (one female and two males, aged 22∼28 years) participated in the study. The subject S3 was the author C.Z. The other two subjects were naïve to the purpose of this study. All subjects were monolingual native-Chinese speakers and right-handed. All subjects had a normal or corrected-to-normal vision and considerable experience of fMRI experiments.

## Convolutional Neural Network

We chose AlexNet and implemented it using the Caffe deep learning framework (Deng et al., 2009; Krizhevsky et al., 2012). AlexNet consists of five convolutional layers and three fully-connected layers (**Figure 1D**). The five convolutional layers each have 96, 256, 384, 384, and 256 linear convolutional kernels. The three fully-connected layers each have 4096, 4096, and 1000 artificial neurons. All convolutional layers perform linear convolution and rectified linear unit (ReLU) gating. Spatial max pooling is used only in layers 1, 2, and 5 to promote the spatial invariance of sensory inputs. In layers 1 and 2, local response normalization implements the inhibitory interactions across channels in a convolutional layer. In other words, the strong activity of a neuron in the normalization pool suppresses the activities of other neurons. Lateral inhibition of neurons is a well-established phenomenon in visual neuroscience and has proven to be critical to many forms of visual processing (Blakemore et al., 1970). The ReLU activation function and

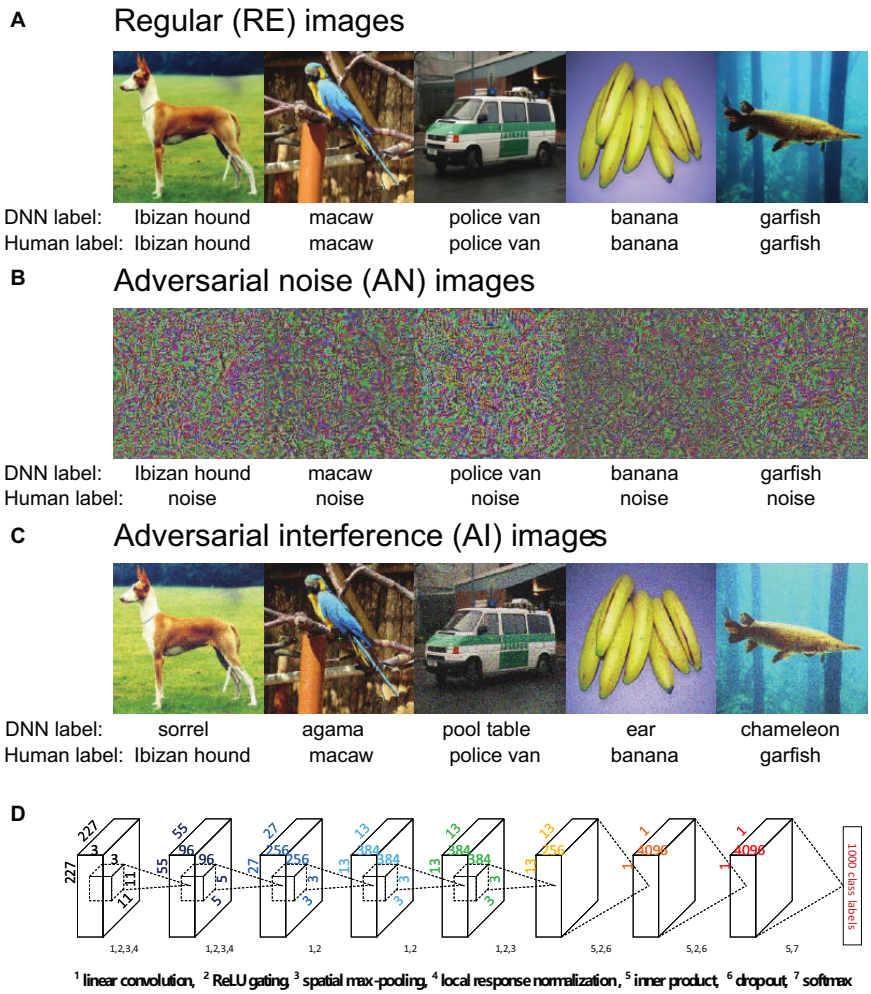


**FIGURE 1 | (A–C)** Example regular (RE, panel **A**), adversarial noise (AN, panel **B**) images and adversarial interference (AI, panel **C**) images. The five AN and five AI images one-by-one correspond to the five RE images. The labels provided by AlexNet and humans are listed under the images. The AI images contain a small amount of special image noise but overall look similar to the corresponding RE images. Humans can easily recognize the AI images as corresponding categories but the AN images as noise. AlexNet can classify the AN images into corresponding categories with over 99% confidence, but recognize the AI images as wrong categories. **(D)** The architecture of AlexNet. Details have been documented in Krizhevsky et al. (2012). Each layer uses some or all the following operations: linear convolution, ReLU gating, spatial max-pooling, local response normalization, inner product, dropout and softmax.

dropout are used in fully-connected layers 6 and 7. Layer 8 uses the softmax function to output the probabilities for 1000 target categories. In our study, all images were resized to 227 × 227 pixels in all three RGB color channels.

## Image Stimuli

### Regular Images

Regular (RE) images (**Figure 1A**) in our study were sampled from the ImageNet database (Deng et al., 2009). ImageNet is currently the most advanced benchmark database on which almost all state-of-the-art CNNs are trained for image classification. We selected one image (width and height > 227 pixels and aspect ratio > 2/3 and < 1.5) from each of 40 representative object categories. AlexNet can classify all images into their corresponding categories with probabilities greater than 0.99. The 40 images can be evenly divided into 5 classes: dogs, birds, cars, fruits, and aquatic animals (see **Supplementary Table 1** for details).

### Adversarial Images

Adversarial images include adversarial noise (AN) (**Figure 1B**) and adversarial interference (AI) images (**Figure 1C**). A pair of AN and AI images were generated for each RE image. As such, a total of 120 images (40 RE + 40 AN + 40 AI) were used in the entire experiment.

The method to generate AN images has been documented in Nguyen A et al. (Nguyen et al., 2015). We briefly summarize the method here. We first used the averaged image of all images in ImageNet as the initial AN image. Note that the category label of the corresponding RE image was known, and AlexNet had been fully trained. As such, we first fed the initial AN image to

AlexNet and forwardly computed the probability for the correct category. This probability was expected to be initially low. We then used the backpropagation method to transduce error signals from the top layer to image pixel space. Pixel values in the initial AN image were then adjusted accordingly to enhance the classification probability. This process of forwarding calculation and backpropagation was iterated many times until the pixel values of AN image converged.

We also included an additional regularization item to control the overall intensity of the image. Formally, let $P_c(I)$ be the probability of class $c$ (RE image label) given an image $I$. We would like to find an $L_2$-regularized image $I^*$, such that it maximizes the following objective:

$$I^* = \arg\max_I \; P_c(I) - \lambda \, ||I\text{-}I_{mean}||_2^2 \,, \qquad (1)$$

where, $\lambda$ is the regularization parameter and $I_{mean}$ is the grand average of all images in ImageNet. Finally, all the probabilities of generated AN images used in our experiment being classified into RE images were greater than 0.99. Note that the internal structure (i.e., all connection weights) of AlexNet was fixed throughout the entire training process, and we only adjusted pixel values in input AN images.

The AI images were generated by adding noise to the RE images. For an RE image (e.g., dog), a wrong class label (e.g., bird) was pre-selected (see **Supplementary Table 1** for details). We then added random noise (uniform distribution $-5\sim5$) to every pixel in the RE image. The resulted image was kept if the probability of this image being classified into the wrong class (i.e., bird) increased, and was discarded otherwise. This procedure was repeated many times until the probability for the wrong class exceeded 0.5 (i.e., wrong class label as the top1 label). We deliberately choose 0.5 because under this criteria the resulted images were still visually comparable to the RE images. A higher stopping criteria (e.g., 0.99) may overly load noises and substantially reduce image visibility. We further used the similar approach as AN images (change the $I_{mean}$ in Eq. 1 to $I_{RE}$.) to generate another set of AI images (with a probability of over 0.99 to be classified into the "wrong" class) and confirmed that the results in AlexNet RSA analyses did not substantially change under this regime (see **Supplementary Figure 4**). We adopted the former not the latter approach in our fMRI experiment because the differences between the AI and the RE images were so small that the human eye can hardly see it in the experiment. This is meaningless for an fMRI experiment as the AI and the RE images look "exactly" the same, which is equivalent to present the identical images twice.

## Apparatus

All computer-controlled stimuli were programmed in Eprime 2.0 and presented using a Sinorad LCD projector (resolution 1920 × 1080 at 120 Hz; size 89 cm × 50 cm; viewing distance 168 cm). Stimuli were projected onto a rear-projection monitor located over the head. Subjects viewed the monitor via a mirror mounted on the head coil. Behavioral responses were recorded by a button box.

## fMRI Experiments

### Main Experiment

Each subject underwent two scanning sessions in the main experiment. In each session, half of all images (20 images x 3 RE/AN/AI = 60 images) were presented. Each session consisted of five scanning runs, and each run contained 129 trials (2 trials per image and 9 blank trials). The image presentation order was randomized within a run. In a trial, a blank lasted 2 s and was followed by an image (12° × 12°) of 2 s. A 20 s blank period was included to the beginning and the end of each run to establish a good baseline and compensate for the initial insatiability of the magnetic field. A fixation point (0.2° × 0.2°) was shown at center-of-gaze, and participants were instructed to maintain steady fixation throughout a run. Participants pressed buttons to perform an animal judgment task—whether an image belongs to animals. The task aimed to engage subjects' attention onto the stimuli.

### Retinotopic Mapping and Functional Localizer Experiments

A retinotopic mapping experiment was also performed to define early visual areas, as well as two functional localizer experiments to define lateral occipital (LO) lobe and human middle temporal lobe (hMT+).

The retinotopic experiment used standard phase-encoding methods (Engel et al., 1994). Rotating wedges and expanding rings were filled by textures of objects, faces, and words, and were presented on top of achromatic pink-noise backgrounds (http://kendrickkay.net/analyzePRF/). Early visual areas (V1–V4) were defined on the spherical cortical surfaces of individual subjects.

The two localizer experiments were used to create a more precise LO mask (see region-of-interest definition section below). Each localizer experiment contained two runs. In the LO localizer experiment, each run consisted of 16 stimulus blocks and 5 blank blocks. Each run began with a blank block, and a blank block appeared after every 4 stimulus blocks. Each block lasted 16 s. Intact images and their corresponding scrambled images were alternately presented in a stimulus block. Each stimulus block contained 40 images (i.e., 20 intact + 20 scramble images). Each image (12° × 12°) lasted 0.3 s and was followed by a 0.5 s blank.

In the hMT+ localizer experiment, each run contained 10 stimulus blocks, and each block lasted 32 s. In a block, a static dot stimulus (24 s) and a moving-dot stimulus (8 s) were alternately presented. All motion stimuli subtended a 12° × 12° square area on a black background. An 8 s blank was added to the beginning and the end of each run. Note that hMT+ here is only used to remove motion-selective vertices from the LO mask (see Region-Of-Interest definitions). We did not analyze motion signals in hMT+ as all our images were static.

## MRI Data Acquisition

All MRI data were collected using a 3.0-Tesla Siemens MAGNETOM Prisma scanner and a 32-channel head coil at the Department of Radiology at the People's Hospital of Henan Province.

An interleaved T2*-weighted, single-shot, gradient-echo echo-planar imaging (EPI) sequence was used to acquire

functional data (60 slices, slice thickness 2 mm, slice gap 0 mm, field of view 192 × 192 $mm^2$, phase-encode direction anterior-posterior, matrix size 96 × 96, *TR/TE* 2000/29 ms, flip angle 76°, nominal spatial resolution 2 × 2 × 2 $mm^3$). Three B0 fieldmaps were acquired to aid post-hoc correction for EPI spatial distortion in each session (resolution 2 × 2 × 2 $mm^3$, $TE_1$ 4.92 ms, $TE_2$ 7.38 ms, *TA* 2.2 min). In addition, high-resolution T1-weighted anatomical images were also acquired using a 3D-MPRAGE sequence (*TR* 2300 ms, *TE* 2.26 ms, *TI* 900 ms, flip angle 8°, field of view 256 × 256 $mm^2$, voxel size 1. × 1. × 1. $mm^3$).

## MRI Data Preprocessing

The pial and the white surfaces of subjects were constructed from T1 volume using FreeSurfer software (http://surfer.nmr.mgh.harvard.edu). An intermediate gray matter surface between the pial and the white surfaces was also created for each subject.

Our approach for dealing with EPI distortion followed Kay et al. (2019). Fieldmaps acquired in each session were phase-unwrapped using the FSL utility prelude (version 2.0) with flags -s -t 0. We then regularized the fieldmaps by performing 3D local linear regression using an Epanechnikov kernel with radius 5 mm. We used values in the magnitude component of the fieldmap as weights in the regression in order to improve robustness of the field estimates. This regularization procedure removes noise from the fieldmaps and imposes spatial smoothness. Finally, we linearly interpolated the fieldmaps over time, producing an estimate of the field strength for each functional volume acquired.

For functional data, we discarded the data points of the first 18 s in the main experiment, the first 14 s in the LO localizer experiment, and the first 6 s in the hMT+ localizer experiment. This procedure ensures a 2 s blank was kept before the first task block in all three experiments.

The functional data were initially volume-based pre-processed by performing one temporal and one spatial resampling. The temporal resampling realized slice time correction by executing one cubic interpolation for each voxel's time series. The spatial resampling was performed for EPI distortion and head motion correction. The regularized time-interpolated field maps were used to correct EPI spatial distortion. Rigid-body motion parameters were then estimated from the undistorted EPI volumes with the SPM5 utility spm_realign (using the first EPI volume as the reference). Finally, the spatial resampling was achieved by one cubic interpolation on each slice-time-corrected volume (the transformation for correcting distortion and the transformation for correcting motion are concatenated such that a single interpolation is performed).

We co-registered the average of the pre-processed functional volumes obtained in a scan session to the T1 volume (rigid-body transformation). In the estimation of the co-registration alignment, we used a manually defined 3D ellipse to focus the cost metric on brain regions that are unaffected by gross susceptibility effects (e.g., near the ear canals). The final result of the co-registration is a transformation that indicates how to map the EPI data to the subject-native anatomy.

With the anatomical co-registration complete, the functional data were re-analyzed using surface-based pre-processing. The reason for this two-stage approach is that the volume-based pre-processing is necessary to generate the high-quality undistorted functional volume that is used to determine the registration of the functional data to the anatomical data. It is only after this registration is obtained that the surface-based pre-processing can proceed.

In surface-based pre-processing, the exact same procedures associated with volume-based pre-processing are performed, except that the final spatial interpolation is performed at the locations of the vertices of the intermediate gray matter surfaces. Thus, the only difference between volume- and surface-based pre-processing is that the data are prepared either on a regular 3D grid (volume) or an irregular manifold of densely spaced vertices (surface). The entire surface-based pre-processing ultimately reduces to a single temporal resampling (to deal with slice acquisition times) and a single spatial resampling (to deal with EPI distortion, head motion, and registration to anatomy). Performing just two simple pre-processing operations has the benefit of avoiding unnecessary interpolation and maximally preserving spatial resolution (Kang et al., 2007; Kay and Yeatman, 2017; Kay et al., 2019). After this pre-processing, time-series data for each vertex of the cortical surfaces were ultimately produced.

## General Linear Modeling

We estimated the vertex responses (i.e., beta estimates from GLM modeling) of all stimulus trials in the main experiment using the GLMdenoise method (Kay et al., 2013). All blank trials were modeled as a single predictor. This analysis yielded beta estimations of 241 conditions (120 images × 2 trials + 1 blank trial). Notably, we treated two presentations of the same image as two distinct predictors in order to calculate the consistency of the response patterns across the two trials.

## Region-of-Interest Definitions

Based on the retinotopic experiment, we calculated the population receptive field (pRF) (http://kendrickkay.net/analyzePRF) of each vertex and defined low-level visual areas (V1–V4) based on the pRF maps. To define LO, we first selected vertices that show significantly higher responses to intact images than scrambled images (two-tails *t*-test, $p < 0.05$, uncorrected). In addition, hMT+ was defined as the area that shows significantly higher responses to moving than static dots (two-tails *t*-test, $P < 0.05$, uncorrected). The intersection vertices between LO and hMT+ were then removed from LO.

## Vertex Selection

To further select task-related vertices in each ROI (**Figure 2A**), we performed a searchlight analysis on flattened 2D cortical surfaces (Chen et al., 2011). For each vertex, we defined a 2D searchlight disk with 3 mm radius. The geodesic distance between two vertices was approximated by the length of the shortest path between them on the flattened surface. Given the vertices in the disk, we calculated the representational dissimilarity matrices (RDM) of all RE images for each of the two presentation trials. The two RDMs were then compared (Spearman's R) to show the consistency of activity patterns across the two trials. Here rank-correlation (e.g., Spearman's R) is used as it was recommended

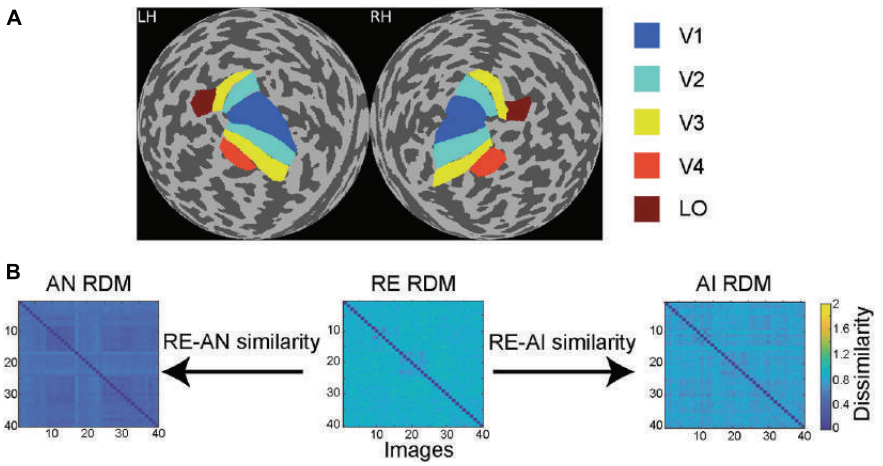


**FIGURE 2 | (A)** Regions of interest (ROIs) in a sample subject. Through retinotopic mapping and functional localizer experiments, we identified five ROIs—V1, V2, V3, V4 and lateral occipital (LO) cortex—in both left (LH) and right (RH) hemispheres. **(B)** Calculation of RE-AN and RE-AI similarity. For each CNN layer or brain ROI, three RDMs are calculated with respect to the three types of images. We then calculate the Spearman correlation between the AN and the RE RDMs, obtaining the RE-AN similarity. Similarly, we can calculate the RE-AI similarity.

when comparing two RDMs (Kriegeskorte et al., 2008; Nili et al., 2014).

The 200 vertices (100 vertices from each hemisphere) with the highest correlation values were selected in each ROI for further analysis (**Figure 3**). Note that vertex selection was only based on the responses to the RE images and did not involve any response data for the AN and the AI images. We also selected a total of 400 vertices in each area and we found our results held. The results are shown in **Supplementary Figure 2**.

## Representational Similarity Analysis

We applied RSA separately to the activity in the CNN and the brain.

### RSA on CNN Layers and Brain ROIs

For one CNN layer, we computed the representational dissimilarity between every pair of the RE images, yielding a 40 × 40 RDM (i.e., $RDM_{RE}$) for the RE images. Similarly, we obtained the other two RDMs each for the AN (i.e., $RDM_{AN}$) and the AI images (i.e., $RDM_{AI}$). We then calculated the similarity between the three RDMs as follows:

$$R_{RE-AN} = corr(RDM_{RE}, RDM_{AN}), \quad (2)$$

$$R_{RE-AI} = corr(RDM_{RE}, RDM_{AI}), \quad (3)$$

This calculation generated one RE-AN similarity value and one RE-AI similarity value for that CNN layer (see **Figure 2B**). We repeated the same analysis above on the human brain except that we used the activity of vertices in a brain ROI.

In a given ROI or AlexNet layer, we first resampled 80% voxels or artificial neurons without replacement (**Supplementary Figure 5**). In each sample, we calculated RE, AI, and AN RDM, and calculated the difference between RE-AI similarity and RE-AN similarity, obtaining one difference value. This was done 1000 times, yielding 1000 different values as the baseline distribution for RE-AI and RE-AN difference. This method is used for examining the relative difference between the RE-AN and the RE-AN similarities.

To construct the null hypotheses for the absolute RE-AN and the RE-AI similarities, in each voxel or artificial neuron sample, we further permuted the image labels with respect to their corresponding activities for the RE images (**Supplementary Figure 6**). In other words, an image label may be paired with a wrong activity pattern. We then recalculated the RE-AN and the RE-AI similarities. In this way, 1000 RE-AN and 1000 RE-AI similarity values were generated. The two distributions consisting of 1000 values were regarded as the null hypothesis distributions of RE-AN or RE-AI, respectively.

In addition, the Mann-Kendall test was applied to assess the monotonic upward or downward trend of the RE-AN similarities over CNN layers. The Mann-Kendall test can be used in place of a parametric linear regression analysis, which can be used to test if the slope of the estimated linear regression line is different from zero.

In order to verify the statistical effectiveness of the fMRI experimental results of the three subjects, we used the G*Power tool (Faul et al., 2009) to re-analyze our experimental results. For each ROI, we carried out a paired *t*-test (i.e., "means: difference between two dependent means (matched pairs)" in G*Power) on the RE-AI similarities and the RE-AN similarities of the three subjects. We calculated three RE-AI/RE-AN difference values (i.e., the height difference between blue and red bars in **Figure 3**), each for one subject. The effect size was determined from the mean and SD of the difference values. We first set the type of power analysis to "*post hoc*: compute achieved power – given α, sample size, and effect size" to estimate the statistical power given $N = 3$. The statistical power (1-β error probability, α error probability was set to 0.05) was then calculated. We then set the type of power analysis to "*a priori*: compute required sample

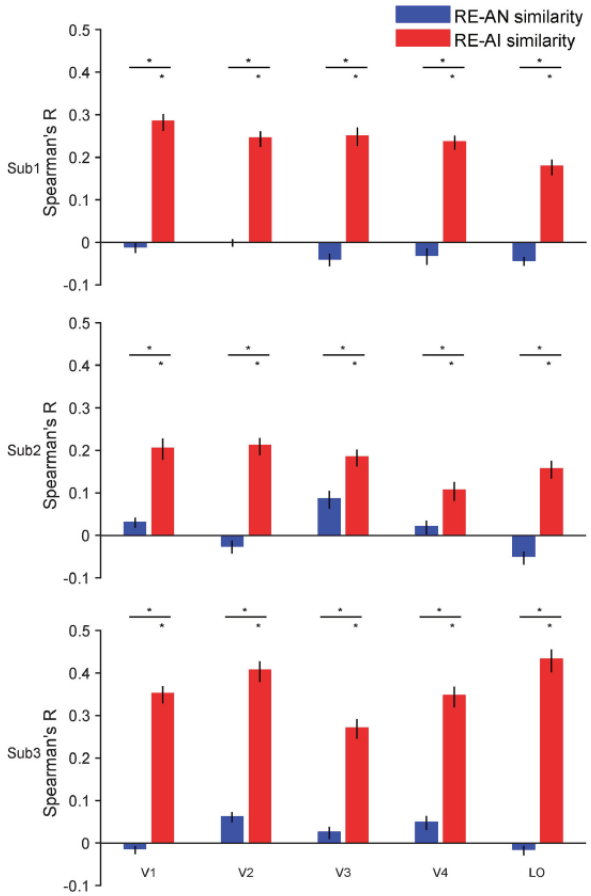


**FIGURE 3 |** RE-AI and RE-AN similarities in the human brain. Three subplots indicate the three human subjects. In all five brain ROIs, the RE-AI (red bars) similarities are substantially higher than the RE-AN (blue bars) similarities. Error bars are 95% confidence intervals of similarity values by bootstrapping vertices in one brain ROI (see Methods). The black asterisks above bars indicate that the similarity values are significantly different from null hypotheses (permutation test, $p < 0.05$, see Methods).

size – given α, power, and effect size," and calculated the estimated minimum required sample size to achieve a statistical power of 0.8 with the current statistics.

### Searchlight RSA

We also performed a surface-based searchlight analysis in order to show the cortical topology of the RE-AN and the RE-AI similarity values. For each vertex, the same 2D searchlight disk was defined as above. We then repeated the same RSA on the brain, producing two cortical maps with respect to the RE-AN and RE-AI similarity values.

## Forward Encoding Modeling

Here, forward encoding models assume that the activity of a voxel in the brain can be modeled as the linear combination of the activity of artificial neurons in CNNs. Thus, forward encoding modeling can bridge the representations of the two systems. Thus, forward encoding modeling can bridge the representations of the two systems. This is also the typical approach in existing related works (Güçlü and van Gerven, 2015; Kell et al., 2018).

We first trained the forward encoding models only based on the RE images data in the brain and the CNN. For the response sequence $y = \{y_1, \ldots, y_d\}^{\mathrm{T}}$ of one vertex to the 40 RE images, it is expressed as Eq. (4):

$$\mathrm{y} = \mathrm{Xw}, \tag{4}$$

**X** is an $m$-by-($n$+1) matrix, where $m$ is the number of training images (i.e., 40), and $n$ is the number of units in one CNN layer. The last column of **X** is a constant vector with all elements equal to 1. **w** is an ($n$+1)-by−1 unknown weighting matrix to solve. Because the number of training samples $m$ was less than the number of units $n$ in all CNN layers, we imposed an additional sparse constraint on the forward encoding models to avoid overfitting:

$$\min_{\mathrm{w}} ||\mathrm{w}||_0 \quad \text{subject to} \quad \mathrm{y} = \mathrm{Xw}, \tag{5}$$

Sparse coding has been widely suggested and used in both neuroscience and computer vision (Vinje and Gallant, 2000; Cox and Savoy, 2003). We used the regularized orthogonal matching pursuit (ROMP) method to solve the sparse representation problem. ROMP is a greedy method developed by Needell D and R Vershynin (Needell and Vershynin, 2009) for sparse recovery. Features for prediction can be automatically selected to avoid overfitting. For the selected 200 vertices in each human ROI, we established 8 forward encoding models corresponding to the 8 CNN layers. This approach yielded a total of 40 forward encoding models (5 ROIs × 8 layers) for one subject.

Based on the train forward encoding models, we calculated the Pearson correlation between the empirically measured and model-predicted response patterns evoked by the adversarial images. To test the prediction accuracy against null hypotheses, we randomized the image labels and performed permutation tests as described above. Specifically, we resampled 80% vertices in a brain ROI 1000 times without replacement and in each sample recalculated the mean response prediction accuracy, resulting in a bootstrapped accuracy distribution with 1000 mean response prediction accuracy values (**Supplementary Figure 7**). The upper and lower bounds of the 95% confidence intervals were derived from the bootstrapped distribution. Similarly, we compared the bootstrapped distributions of two types of adversarial images to derive the statistical difference between the RE-AI and the RE-AN similarity.

# RESULTS

## Dissociable Neural Representations of Adversarial Images in AlexNet and the Human Brain

### Human Brain

For one brain ROI, we calculated the representational dissimilarity matrix (i.e., 40 x 40 RDM) for each of the

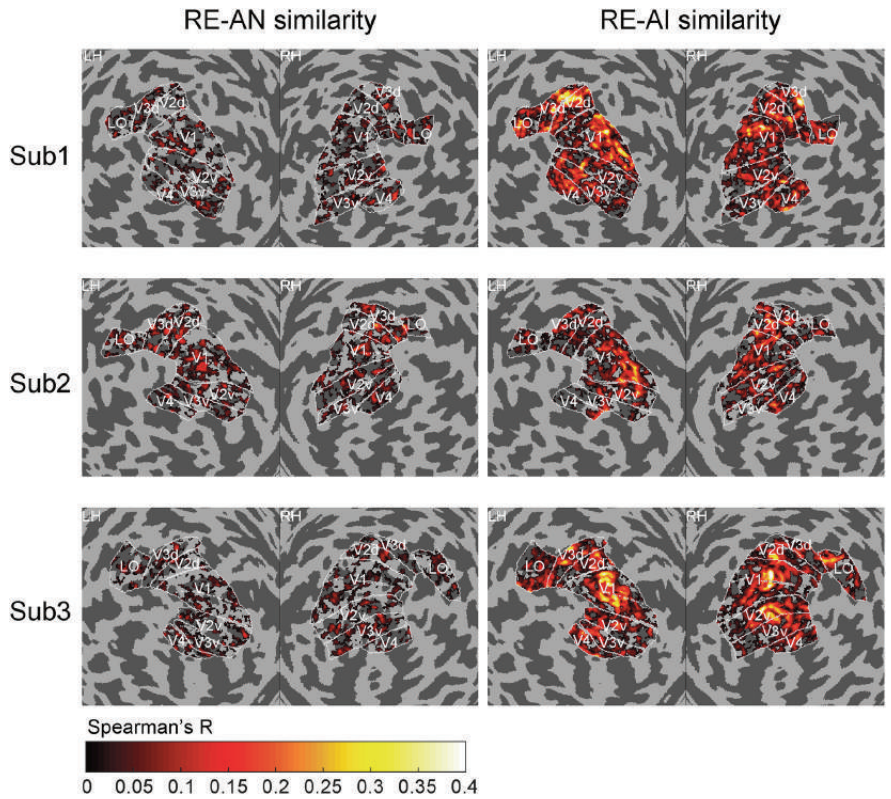


**FIGURE 4 |** Cortical topology of RE-AI and RE-AN similarities. The RE-AI similarities are overall higher than the RE-AN similarities across all early visual areas in the human brain.

three image types. We then calculated the RE-AN similarity—the correlation between the RDM of the RE images and that of the AN images, and the RE-AI similarity between the RE images and the AI images.

We made three major observations. First, the RE-AI similarities were significantly higher than null hypotheses in almost all ROIs in the three subjects (red bars in **Figure 3**, permutation test, all $p$-values $< 0.005$, see Methods for the deviation of null hypotheses). Conversely, this was not true for the RE-AN similarities (blue bar in **Figure 3**, permutation test, only four $p$-values $< 0.05$ in 3 subjects x 5 ROI = 15 tests). Third and more importantly, we found significantly higher RE-AI similarities than the RE-AN similarities in all ROIs (**Figure 3**, bootstrap test, all $p$-values $< 0.0001$). These results suggest that the neural representations of the AI images, compared with the AN images, are much more similar to that of the corresponding RE images. Notably, this representational structure is also consistent with the perceptual similarity of the three types of images in humans. In other words, the neural representations of all images in the human brain largely echo their perceptual similarity.

In addition, the results of the statistical power analysis showed that the final average power (1-β error probability, α error probability was set to 0.05, $N = 3$) across five ROIs for the paired $t$-test on RE-AI similarities and RE-AN similarities of the three subjects equaled 0.818 (V1:0.911, V2: 0.998, V3:0.744, V4:0.673, LO:0.764). And the average minimum required sample size was 2.921 (V1:2.623, V2:2.131, V3:3.209, V4:3.514, LO:3.129, the power was set to 0.8). In other words, the number of subjects can meet the minimum statistical power.

We also performed a searchlight analysis to examine the cortical topology of the neural representations. The searchlight analysis used the same calculation as above (see Methods). We replicated the results (see **Figure 4**) and found a distributed pattern of higher RE-AI similarities in the early human visual cortex. In addition, we expanded our searchlight analysis for broader regions (see **Supplementary Figure 3**) and obtained the qualitatively same main results.

## AlexNet

We repeated our analyses above in AlexNet and again made three observations. First, the RE-AI similarities were higher than null hypotheses across all layers (**Figure 5**, permutation test, all $p$-values $< 0.001$), and the RE-AI similarities declined from low to high layers (Mann–Kendall test, $p = 0.009$). Second, the RE-AN similarities were initially low ($p$-values $> 0.05$ in layers 1–2) but then dramatically increased (Mann–Kendall test, $p < 0.001$) and became higher than the null hypotheses from layer 3 (all $p$-values $< 0.05$ in layers 3–8). Third and most importantly, we found that the RE-AN similarities were not higher than the RE-AI similarities in all intermediate layers (i.e., layers 1–7, bootstrap

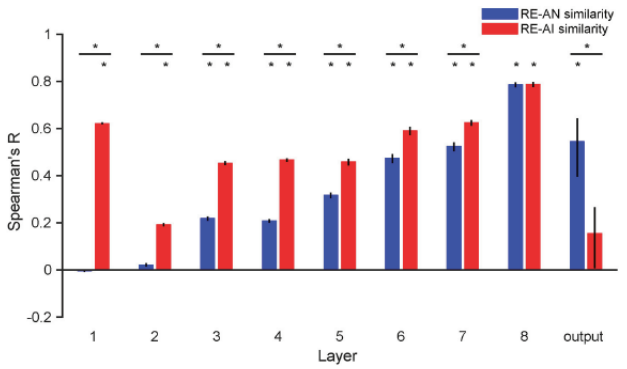


**FIGURE 5 |** RE-AN and RE-AI similarities across layers in AlexNet. the RE-AN similarity increases and the RE-AI similarities decline along the processing hierarchy. The RE-AN similarities are not higher than the RE-AI similarities in all representational layers (i.e., layer 1–7). Error bars indicate 95% bootstrapped confidence intervals (see Methods).

test, all $p$-values $< 0.05$, layer 7, $p = 0.375$) except the output layer (i.e., layer 8, $p < 0.05$).

These results are surprising because it suggests that neural representations of the AI images, compared with the AN images, are more similar to the representations of the RE images. However, the output labels of the AN images are similar to those of the corresponding RE images in AlexNet. In other words, there exists substantial inconsistency between the representational similarity and perceptual similarity in AlexNet. We emphasize that, assuming that in order for two images look similar, there must be at least some neural populations somewhere in a visual system that represents them similarly. But, astonishingly, we found no perception-compatible neural representations in any representational layer. Also, the transformation from layer 7 to the output layer is critical and eventually renders the RE-AN similarity higher than the RE-AI similarity in the output layer. This is idiosyncratic because AlexNet does not implement effective neural codes of objects in representational layers beforehand but the last transformation reverses the relative RDM similarity of the three types of images. This is drastically different from the human brain that forms correct neural codes in all early visual areas.

## Forward Encoding Modeling Bridges Responses in AlexNet and Human Visual Cortex

The RSA above mainly focuses on the comparisons across image types within one visual system. We next used forward encoding modeling to directly bridge neural representations across the two systems. Forward encoding models assume that the activity of a voxel in the brain can be modeled as the linear combination of the activity of multiple artificial neurons in CNNs. Following this approach, we trained a total of 40 (5 ROIs x 8 layers) forward encoding models for one subject using regular images. We then tested how well these trained forward encoding models can generalize to the corresponding adversarial images. The rationale is that, if the brain and AlexNet process images in a similar fashion, the forward encoding models trained on the RE images should transfer to the adversarial images, and vice versa if not.

We made two major findings here. First, almost all trained encoding models successfully generalized to the AI images (**Figure 6**, warm color bars, permutation test, $p$-values $< 0.05$ for 113 out of the 120 models for three subjects) but not to the AN images (**Figure 6**, cold color bars, permutation test, $p$-values $> 0.05$ for 111 out of the 120 models). Second, the forward encoding models exhibited much stronger predictive power on the AI images than the AN images (bootstrap test, all $p$-values $< 0.05$, except the encoding model based on layer 8 for LO in subject 2, $p = 0.11$). These results suggest that the functional correspondence between AlexNet and the human brain only holds when processing RE and AI images but not AN images. This result is also consonant with the RSA above and demonstrates that both systems treat RE and AI images similarly, but AN images very differently. But again, note that AlexNet exhibits the opposite behavioral pattern of human vision.

# DISCUSSION AND CONCLUSION

Given that current CNNs still fall short in many tasks, we use adversarial images to probe the functional differences between a prototypical CNN—AlexNet, and the human visual system. We make three major findings. First, the representations of AI images, compared with AN images, are more similar to the representations of corresponding RE images. These representational patterns in the brain are consistent with human percepts (i.e., perceptual similarity). Second, we discover a representation-perception disassociation in all intermediate layers in AlexNet. Third, we use forward encoding modeling to link neural activity in both systems. Results show that the processing of RE and AI images are quite similar but both are significantly different from AN images. Overall, these observations demonstrate the capacity and limit of the similarities between current CNNs and human vision.

## Abnormal Neural Representations of Adversarial Images in CNNs

To what extent neural representations reflect physical or perceived properties of stimuli is a key question in modern vision science. In the human brain, researchers have found that early visual processing mainly processes low-level physical properties of stimuli, and late visual processing mainly supports high-level categorical perception (Grill-Spector and Malach, 2004). We ask a similar question here—to what extent neural representations in CNNs or the human brain reflect their conscious perception.

One might argue that the representation-perception disassociation in AlexNet is trivial, given that we already know that AlexNet exhibits opposite behavioral patterns compared to human vision. But we believe thorough quantifications of their neural representations in both systems are still of great value. First, neural representations do not necessarily follow our conscious perception, and numerous neuroscience studies have shown disassociated neural activity and perception

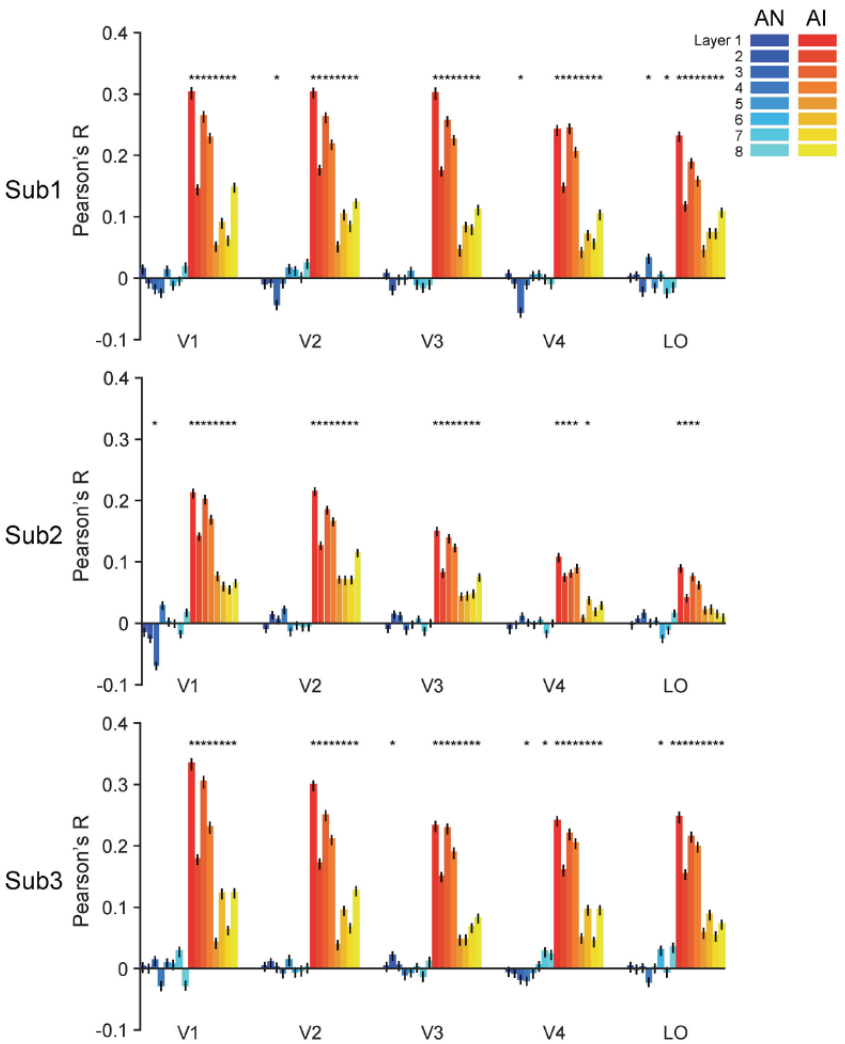


**FIGURE 6 |** Accuracy of forward encoding models trained on RE images and then tested on adversarial images. After the models are fully trained on the RE images, we input the adversarial images as inputs to the models can predict corresponding brain responses. The *y*-axis indicates the Pearson correlation between the brain responses predicted by the models and the real brain responses. The generalizability of forward encoding models indicates the processing similarity between the RE and AN (cool colors) or AI (warm colors) images. Error bars indicate 95% bootstrapped confidence intervals (see Methods).

in both the primate or human brain in many cases, such as visual illusion, binocular rivalry, visual masking (Serre, 2019). The question of representation-perception association lies at the center of the neuroscience of consciousness and should also be explicitly addressed in AI research. Second, whether representation and perception are consistent or not highly depends on processing hierarchy, which again needs to be carefully quantified across visual areas in the human brain and layers in CNNs. Here, we found no similar representations of AN and regular images in any intermediate layer in AlexNet even though they "look" similar. This is analogous to the scenario that we cannot decode any similar representational patterns of two images throughout a subject's brain, although the subject behaviorally reports the two images are similar.

## Adversarial Images as a Tool to Probe Functional Differences Between the CNN and Human Vision

In computer vision, adversarial images impose problems on the real-life applications of artificial systems (i.e., adversarial attack) (Yuan et al., 2017). Several theories have been proposed to explain the phenomenon of adversarial images (Akhtar and Mian, 2018). For example, one possible explanation is that CNNs are forced to behave linearly in high dimensional spaces, rendering them vulnerable to adversarial attacks (Goodfellow et al., 2014b). Besides, flatness (Fawzi et al., 2016) and large local curvature of the decision boundaries (Moosavi-Dezfooli et al., 2017), as well as low flexibility of the networks (Fawzi et al., 2018) are all possible reasons. (Szegedy et al., 2013) has suggested that current CNNs

are essentially complex nonlinear classifiers, and this discriminative modeling approach does not consider generative distributions of data. We will further address this issue in the next section.

In this study, we focused on one particular utility of adversarial images—to test the dissimilarities between CNNs and the human brain. Note that although the effects of adversarial images indicate the deficiencies of current CNNs, we do not object to the approach to use CNNs as a reference to understand the mechanisms of the brain. Our study here fits the broad interests in comparing CNNs and the human brain in various aspects. We differ from other studies just because we focus on their differences. We do acknowledge that it is quite valuable to demonstrate functional similarities between the two systems. But we believe that revealing their differences, as an alternative approach, might further foster our understandings of how to improve the design of CNNs. This is similar to the logic of using ideal observer analysis in vision science. Although we know human visual behavior is not optimal in many situations, the comparison to an ideal observer is still meaningful as it can reveal some critical mechanisms of human visual processing. Also, we want to emphasize that mimicking the human brain is not the only way or even may not be the best way to improve CNN performance. Here, we only suggest a potential route given that current CNNs still fall short in many visual tasks as compared to humans.

Some recent efforts have been devoted to addressing CNN-human differences. For example, Rajalingham et al. (2018) found that CNNs explain human (or non-human primate) rapid object recognition behavior at the level of category but not individual images. CNNs better explain the ventral stream than the dorsal stream (Wen et al., 2017). To further examine their differences, people have created some unnatural stimuli/tasks, and our work on adversarial images follows this line of research. The rationale is that, if CNNs are similar to humans, they should exhibit the same capability in both ordinary and unnatural circumstances. A few studies adopted some other manipulations (Flesch et al., 2018; Rajalingham et al., 2018), such as manipulation of image noise (Geirhos et al., 2018) and distortion (Dodge and Karam, 2017).

## Possible Caveats of CNNs in the Processing of Adversarial Images

Why CNNs and human vision behave differently on adversarial images, especially on AN images? We want to highlight three reasons and discuss the potential route to circumvent them.

First, current CNNs are trained to match the classification labels generated by humans. This approach is a discriminative modeling approach that characterizes the probability of $p(\text{class} \mid \text{image})$. Note that natural images only occupy a low-dimensional manifold in the entire image space. Under this framework, there must exist a set of artificial images in the image space that fulfills a classifier but does not belong to any distribution of real images. Humans cannot recognize AN images because humans do not merely rely on discriminative classifiers but instead perform Bayesian inference and take into consideration both likelihood $p(\text{image}|\ \text{class})$ and prior experience $p(\text{class})$. One approach to overcome this is to build generative deep models to learn latent distributions of images, such as variational autoencoders (Kingma and Welling, 2013) and generative adversarial networks (Goodfellow et al., 2014a).

Another advantage of deep generative models is to explicitly model the uncertainty in sensory processing and decision. It has been well-established in cognitive neuroscience that the human brain computes not only form a categorical perceptual decision, but also a full posterior distribution over all possible hidden causes given a visual input (Knill and Pouget, 2004; Wandell et al., 2007; Pouget et al., 2013). This posterior distribution is also propagated to downstream decision units and influences other aspects of behavior.

Third, more recurrent and feedback connections are needed. Numerous studies have shown the critical role of top-down processing in a wide range of visual tasks, including recognition (Bar, 2003; Ullman et al., 2016), tracking (Cavanagh and Alvarez, 2005), as well as other cognitive domains, such as memory (Zanto et al., 2011), language comprehension (Zekveld et al., 2006) and decision making (Fenske et al., 2006; Rahnev, 2017). In our results, the responses in the human visual cortex likely reflect the combination of feedforward and feedback effects whereas the activity in most CNNs only reflects feedforward inputs from earlier layers. A recent study has shown that recurrence is necessary to predict neural dynamics in the human brain using CNN features (Engel et al., 1994).

# CONCLUDING REMARKS

In the present study, we compared neural representations of adversarial images in AlexNet and the human visual system. Using RSA and forward encoding modeling, we found that the neural representations of RE and AI images are similar in both systems but AN images were idiosyncratically processed in AlexNet. These findings open a new avenue to help design CNN architectures to achieve brain-like computation.

# DISCLOSURE STATEMENT

All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (Henan Provincial People's Hospital) and with the Helsinki Declaration of 1975, as revised in 2008 (5). Informed consent was obtained from all patients for being included in the study.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Henan Provincial People's Hospital. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

CZ, R-YZ, LT, and BY designed the research. CZ, X-HD, L-YW, G-EH, and LT collected the data. CZ and R-YZ analyzed the data and wrote the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by the National Key Research and Development Plan of China under Grant 2017YFB1002502.

## ACKNOWLEDGMENTS

We thank Pinglei Bao, Feitong Yang, Baolin Liu, and Huafu Chen for their invaluable comments on the manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2021.677925/full#supplementary-material

## REFERENCES

Agrawal, P., Stansbury, D., Malik, J., and Gallant, J. L. (2014). Pixels to voxels: modeling visual representation in the human brain. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1407.5104 (accessed February 23, 2021).

Akhtar, N., and Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: a survey. *IEEE Access.* 6, 14410–14430. doi: 10.1109/access.2018.2807385

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *J. Cogn. Neurosci.* 15, 600–609. doi: 10.1162/089892903321662976

Blakemore, C., Carpenter, R. H., and Georgeson, M. A. (1970). Lateral inhibition between orientation detectors in the human visual system. *Nature* 228, 37–39. doi: 10.1038/228037a0

Cavanagh, P., and Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends Cogn. Sci.* 9, 349–354. doi: 10.1016/j.tics.2005.05.009

Chen, Y., Namburi, P., Elliott, L. T., Heinzle, J., Soon, C. S., Chee, M. W., et al. (2011). Cortical surface-based searchlight decoding. *Neuroimage* 56, 582–592. doi: 10.1016/j.neuroimage.2010.07.035

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* 6:27755.

Cox, D. D., and Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI)"brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19, 261–270. doi: 10.1016/s1053-8119(03)00049-1

Deng, J., Dong, W., Socher, R., Li, L., Kai, L., and Li, F.-F. (2009). "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 248–255.

Dodge, S., and Karam, L. (2017). "Can the early human visual system compete with Deep Neural Networks?," in *Proceedings of the IEEE International Conference on Computer Vision Workshop*, (Venice: IEEE), 2798–2804.

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E.-J., et al. (1994). fMRI of human visual cortex. *Nature* 369, 525–525.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/brm.41.4.1149

Fawzi, A., Fawzi, O., and Frossard, P. (2018). Analysis of classifiers' robustness to adversarial perturbations. *Mach. Learn.* 107, 481–508. doi: 10.1007/s10994-017-5663-3

Fawzi, A., Moosavi-Dezfooli, S.-M., and Frossard, P. (2016). "Robustness of classifiers: from adversarial to random noise," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, (Barcelona: Curran Associates Inc), 1632–1640.

Fenske, M. J., Aminoff, E., Gronau, N., and Bar, M. (2006). "Chapter 1 Top-down facilitation of visual object recognition: object-based and context-based contributions," in *Progress in Brain Research*, eds S. Martinez-Conde, S. L. Macknik, L. M. Martinez, J. M. Alonso, and P. U. Tse (Amsterdam: Elsevier), 3–21. doi: 10.1016/s0079-6123(06)55001-0

Flesch, T., Balaguer, J., Dekker, R., Nili, H., and Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proc. Natl. Acad. Sci.* 115, 10313–10322.

Geirhos, R., Temme, C. R. M., Rauber, J., Schuett, H. H., Bethge, M., and Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1808.08750 (accessed February 23, 2021).

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014a). "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, (Cambridge, MA: MIT Press), 2672–2680.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014b). Explaining and harnessing adversarial examples. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1412.6572 (accessed February 23, 2021).

Grill-Spector, K., and Malach, R. (2004). The human visual cortex. *Annu. Rev. Neurosci.* 27, 649–677.

Güçlü, U., and van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35, 10005–10014. doi: 10.1523/jneurosci.5023-14.2015

Güçlü, U., and van Gerven, M. A. J. (2017). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *Neuroimage* 145, 329–336. doi: 10.1016/j.neuroimage.2015.12.036

Hong, H., Yamins, D. L., Majaj, N. J., and Dicarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* 19, 613–622. doi: 10.1038/nn.4247

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

Jozwik, K. M., Kriegeskorte, N., and Mur, M. (2016). Visual features as stepping stones toward semantics: explaining object similarity in IT and perception with non-negative least squares. *Neuropsychologia* 83, 201–226. doi: 10.1016/j.neuropsychologia.2015.10.023

Kang, X., Yund, E. W., Herron, T. J., and Woods, D. L. (2007). Improving the resolution of functional brain imaging: analyzing functional data in anatomical

space. *Magn. Reson. Imaging* 25, 1070–1078. doi: 10.1016/j.mri.2006.12.005

Kay, K., Jamison, K. W., Vizioli, L., Zhang, R., Margalit, E., and Ugurbil, K. (2019). A critical assessment of data quality and venous effects in sub-millimeter fMRI. *Neuroimage* 189, 847–869. doi: 10.1016/j.neuroimage.2019.02.006

Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F., and Wandell, B. A. (2013). GLMdenoise: a fast, automated technique for denoising task-based fMRI data. *Front. Neurosci.* 7:247. doi: 10.3389/fnins.2013.00247

Kay, K. N., and Yeatman, J. D. (2017). Bottom-up and top-down computations in word-and face-selective cortex. *eLife* 6:e22341.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., and Mcdermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* 98, 630–644. doi: 10.1016/j.neuron.2018.03.044

Khaligh-Razavi, S.-M., Henriksson, L., Kay, K., and Kriegeskorte, N. (2017). Fixed versus mixed RSA: explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *J. Math. Psychol.* 76, 184–197. doi: 10.1016/j.jmp.2016.10.007

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1312.6114 (accessed February 23, 2021).

Knill, D. C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719. doi: 10.1016/j.tins.2004.10.007

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126–1141. doi: 10.1016/j.neuron.2008.10.043

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* 25, 1097–1105.

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

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., Frossard, P., and Soatto, S. (2017). Analysis of universal adversarial perturbations. *arXiv* [Preprint]. Available online at: https://arxiv.org/pdf/1705.09554.pdf (accessed February 23, 2021).

Needell, D., and Vershynin, R. (2009). Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Found. Computat. Math.* 9, 317–334. doi: 10.1007/s10208-008-9031-3

Nguyen, A., Yosinski, J., and Clune, J. (2015). "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Boston, MA: IEEE), 427–436.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computat. Biol.* 10:e1003553. doi: 10.1371/journal.pcbi.1003553

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* 16, 1170–1178. doi: 10.1038/nn.3495

Rahnev, D. (2017). Top-down control of perceptual decision making by the prefrontal cortex. *Curr. Direct. Psychol. Sci.* 26, 464–469. doi: 10.1177/0963721417709807

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., and 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. *J. Neurosci.* 38, 7255–7269. doi: 10.1523/jneurosci.0388-18.2018

Serre, T. (2019). Deep learning: the good, the bad, and the ugly. *Annu. Rev. Vis. Sci.* 5, 399–426. doi: 10.1146/annurev-vision-091718-014951

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Intriguing properties of neural networks. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1312.6199 (accessed February 23, 2021).

Ullman, S., Assif, L., Fetaya, E., and Harari, D. (2016). Atoms of recognition in human and computer vision. *Proc. Natl. Acad. Sci. U.S.A.* 113, 2744–2749.

Vinje, W. E., and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287, 1273–1276. doi: 10.1126/science.287.5456.1273

Wandell, B. A., Dumoulin, S. O., and Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron* 56, 366–383. doi: 10.1016/j.neuron.2007.10.012

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., and Liu, Z. (2017). Neural encoding and decoding with deep learning for dynamic natural vision. *Cereb. Cortex* 28, 4136–4160. doi: 10.1093/cercor/bhx268

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and Dicarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Yamins, D. L. K., and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365. doi: 10.1038/nn.4244

Yuan, X., He, P., Zhu, Q., Bhat, R. R., and Li, X. (2017). Adversarial examples: attacks and defenses for deep learning. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1712.07107 (accessed February 23, 2021).

Zanto, T. P., Rubens, M. T., Thangavel, A., and Gazzaley, A. (2011). Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nat. Neurosci.* 14:656. doi: 10.1038/nn.2773

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., and Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. *Neuroimage* 32, 1826–1836. doi: 10.1016/j.neuroimage.2006.04.199

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.