



Teaching CORnet human fMRI representations for enhanced model-brain alignment

Zitong Lu¹ · Yile Wang²

Received: 23 October 2024 / Revised: 24 March 2025 / Accepted: 1 April 2025
© The Author(s) 2025

Abstract

Deep convolutional neural networks (DCNNs) have demonstrated excellent performance in object recognition and have been found to share some similarities with brain visual processing. However, the substantial gap between DCNNs and human visual perception still exists. Functional magnetic resonance imaging (fMRI) as a widely used technique in cognitive neuroscience can record neural activation in the human visual cortex during the process of visual perception. Can we teach DCNNs human fMRI signals to achieve a more brain-like model? To answer this question, this study proposed ReAlnet-fMRI, a model based on the SOTA vision model CORnet but optimized using human fMRI data through a multi-layer encoding-based alignment framework. This framework has been shown to effectively enable the model to learn human brain representations. The fMRI-optimized ReAlnet-fMRI exhibited higher similarity to the human brain than both CORnet and the control model in within- and across-subject as well as within- and across-modality model-brain (fMRI and EEG) alignment evaluations. Additionally, we conducted an in-depth analysis to investigate how the internal representations of ReAlnet-fMRI differ from CORnet in encoding various object dimensions. These findings provide the possibility of enhancing the brain-likeness of visual models by integrating human neural data, helping to bridge the gap between computer vision and visual neuroscience.

Keywords Brain-inspired neural network · Object recognition · Neural alignment · Human brain-like model

Introduction

Deep convolutional neural networks (DCNNs) in computer vision have rapidly advanced, achieving and even surpassing human performance in object recognition (Lecun et al. 2015). These advancements not only drive the development of artificial intelligence (AI) but have also garnered significant interest from researchers in cognitive neuroscience. Numerous studies have observed that DCNNs not only internally represent basic visual features, such as orientation, position, shape, and texture (Yosinski et al. 2015; Zeiler and Fergus 2014), but also capture important object attributes found in primate and human neuroimaging studies,

such as animacy, spikiness, and real-world size (Bao et al. 2020; Cichy et al. 2014; Coggan and Tong 2023; Khaligh-Razavi et al. 2018; Konkle and Caramazza 2013; Lu and Golomb 2023b; Yargholi and de Beeck 2023). Additionally, some studies that directly compare CNNs with human brain representations have found that DCNNs exhibit hierarchical visual processing similar to the human visual system, both temporally and spatially (Cichy et al. 2016; Güçlü and van Gerven 2015; Kietzmann et al. 2019; Lu and Golomb 2023a; Yamins et al. 2014). However, the presence of significant representational similarities does not imply that the two systems are highly alike. Both representational similarity analysis and regression-based measurements indicate that there remain substantial differences between DCNNs and the human brain in visual perception.

Simply expanding the training dataset or increasing the number of layers within DCNNs cannot enhance model-brain alignment (Schrimpf et al. 2020). Some studies have attempted to modify the model's architecture by constructing dual-way pathways models (Bai et al. 2017; Choi et al. 2023; Han and Sereno 2022, 2023; Sun et al.

✉ Zitong Lu
lu.2637@osu.edu

¹ Department of Psychology, The Ohio State University, Columbus 43210, USA

² Department of Neuroscience, The University of Texas at Dallas, Richardson, USA

2017), adding feedback pathways (Konkle and Alvarez 2023), incorporating topographic constraints (Finzi et al. 2022; Lee et al. 2020; Lu et al. 2023; Margalit et al. 2023) to simulate how the brain works. A recent study has proposed a more brain-like recurrent DCNN called CORnet (Kubilius et al. 2018, 2019). Although CORnet includes only four convolutional layers, it achieves higher similarity to the brain than most other vision models with many more layers (Kubilius et al. 2019), as measured by BrainScore (Schrumpf et al. 2020), a platform for evaluating the similarity between models and primate brains. However, despite CORnet being considered a more brain-like visual system model, it is still solely trained on images and has not been truly optimized using neural data. This raises the question: can we further refine CORnet, which was originally trained only on images, by using neural data to make it more brain-like?

To address this question, researchers have developed two main approaches to optimize DCNNs using neural data by refining the training process to bridge them closer to brain representations. One approach is the similarity-based method, which optimizes a similarity loss to make the model's representations more similar to neural activity (Dapello et al. 2023; Federer et al. 2020; Li et al. 2019; Pirlot et al. 2022). This method has been used in studies to optimize DCNNs using neural activity from the mouse V1, or monkey V1 or IT. The other approach is the encoding-based method, which involves training DCNNs with an additional task to generate real brain neural activity (Safarani et al. 2021; Shao et al. 2024), thus enabling the model to learn more brain-like internal representations. This method not only utilizes invasive neural activity from monkeys but also has recently been applied using non-invasive fMRI signals from humans to optimize ResNet. However, these studies have some limitations. Firstly, they often align a specific model layer with a specific brain region's neural activity, yet the correspondence between model layers and different brain regions is neither one-to-one nor well-understood. Secondly, these studies tend to focus more on whether the optimized model exhibits higher robustness rather than on model-brain alignment, which is the focus of our research.

A recent study proposed a novel multi-layer encoding alignment framework to effectively optimize CORnet using human EEG signals (Lu et al. 2024). This human EEG-optimized model, called ReAlnet, was found to be significantly more similar to the human brain in terms of not only EEG but also fMRI and behavior. This may be one of the most effective methods for using neural data to optimize models, as it does not require specifying direct correspondences between different model layers and brain processing stages. While this demonstrates the robustness of the framework with EEG data, it remains

unknown whether the same framework can be effectively applied to other neural modalities such as fMRI. fMRI and EEG differ fundamentally: EEG captures fast temporal dynamics of neural activity, while fMRI provides spatially resolved but temporally coarse signals. Additionally, fMRI datasets are often smaller and higher-dimensional than EEG datasets, introducing unique challenges for model training. Successfully extending this framework to fMRI would validate its flexibility and highlight its generalizability to diverse neural data types. Therefore, the key question of this study is: Can we enhancing model-brain alignment by extending this framework to let CORnet learn human visual processing from fMRI data without specifying correspondences between its internal convolutional layers and different visual regions of the human brain?

In this study, we adapted CORnet to learn human fMRI representations using a modified multi-layer encoding alignment framework and proposed this fMRI-optimized model as ReAlnet-fMRI. We trained three personalized ReAlnet-fMRI models based on fMRI signals from three individual human subjects when they viewed natural images. We evaluated the model-brain alignment using within- and cross-subject within-modality fMRI data, as well as cross-subject across-modality EEG data. All evaluations suggest that ReAlnet-fMRIs exhibit enhanced model-brain alignment. Additionally, further internal representational analysis of both the purely image-trained CORnet and our fMRI-aligned ReAlnet-fMRIs revealed representational differences between these networks. Our study offers notable contributions to both cognitive neuroscience and artificial intelligence fields: (1) Our image-to-fMRI encoding-based alignment framework can simultaneously optimize multiple layers of the deep convolutional neural network without specifying correspondences between its internal convolutional layers and different visual regions of the human brain. (2) Our fMRI-optimized ReAlnet-fMRIs effectively learned human brain representations and exhibited substantial improvements in aligning with human brain representations across different subjects and neuroimaging modalities (human fMRI and EEG). (3) We provided not only comprehensive evaluations on model-brain alignment but also detailed analyses of models' internal representations to help us understand how the internal representations of ReAlnet-fMRI differ from CORnet trained purely on images.

Methods

Here we first introduce the datasets we used in our study, including two fMRI datasets and one EEG dataset. Then, we describe the basic CORnet architecture and the training process of how we combined human fMRI signals to optimize CORnet to achieve our brain-aligned ReAlnet-fMRI models. Finally, we introduce the evaluation methods for measuring representational similarity between models and human brains and behaviors, and also the details of internal representational analysis on models we conducted to explore what ReAlnet-fMRI learned from human fMRI representations.

Human neuroimaging datasets

fMRI dataset for model training The fMRI data originates from (Shen et al. 2019). This *Shen fMRI dataset* recorded human brain fMRI signals from three subjects while they focused on the center of the screen viewing natural images from ImageNet. We applied the training set from *Shen fMRI dataset*, which comprises fMRI signals of each subject viewing 1,200 images (from 150 object categories, 8 images per category) with each image being viewed 5 times and averaged the fMRI signals across the repeated trials to obtain more stable brain activity for each image observation to train our ReAlnet-fMRIs (1,200 samples for training).

Here, we selected the voxels from the entire visual cortex to obtain fMRI signals. The visual cortex (VC) was defined as the combination of early visual areas (V1, V2, V3, and V4) and the higher visual cortex (HVC), a contiguous region covering LOC, FFA, and PPA, as delineated based on functional localizer data (Shen et al. 2019). V1–V4 were identified through retinotopy experiments, while the HVC was identified using voxels showing significantly higher activation for intact object, face, or scene images compared to scrambled images. And we applied principal component analysis (PCA) based on individual training data to reduce the total number of voxels in the visual cortex to 1,024 feature dimensions. Consequently, the training data corresponding to each human subject consisted of 1,200 samples \times 1,024 features.

fMRI dataset for model test (within-modality & within-subject) To evaluate whether ReAlnet-fMRIs shows higher similarity to human fMRI representations, we applied the test set from *Shen fMRI dataset* to test the within-modality and within-subject model-fMRI similarity. This test set comprises fMRI signals of same three subjects viewing 50 images (from 50 object categories from ImageNet but different from 150 training categories), 40 artificial

shape images, and 10 alphabetical letter images with each image being viewed 24, 20, and 12 times respectively.

Similar to what we did for the training set, we averaged the fMRI signals across the repeated trials to obtain more stable brain activity for each image observation. For the testing, we extracted signals from five regions-of-interest (ROIs) for subsequent comparison of model-fMRI similarity: V1, V2, V3, V4, and the lateral occipital complex (LOC). The inclusion of the LOC region is motivated by its role as the human brain's counterpart to the interior temporal cortex (IT) in non-human primates, a key region involved in high-level object recognition. This correspondence has been established in previous studies, which suggest that LOC in humans performs similar functions to IT in monkeys, particularly in processing high-level semantic and categorical object features (Grill-Spector et al. 2001; Khosla et al. 2022). In addition, V1–V4 were included to evaluate the model's ability to capture low- to mid-level visual representations, providing a complete hierarchical view of visual processing from early visual areas to high-level regions like LOC. This multi-regional comparison, including LOC, allows us to assess the extent to which ReAlnet-fMRI aligns with both lower- and higher-order visual representations in humans.

fMRI dataset for model test (within-modality & across-subject) Although our ReAlnet-fMRIs were trained on individual fMRI signals, we also would like to test whether these models learn more general brain representations across individuals than the original purely image-trained CORnet. Thus we selected another fMRI dataset (Horikawa and Kamitani 2017), *Horikawa fMRI dataset*, which included fMRI signals from five different subjects viewing natural images. Here, we used the test set of *Horikawa fMRI dataset* to evaluate the within-modality but across-subject model-fMRI similarity. This test set comprises fMRI signals of five subjects viewing 50 images (as same as images used in *Shen fMRI dataset's* test set) with each image being repeated 35 times. We also averaged the repeated trials and extracted signals from V1, V2, V3, V4, and LOC to calculate the model-fMRI similarity.

EEG dataset for model test (across-modality & across-subject) To further confirm that our ReAlnet-fMRIs learn more general human brain representations instead of just human fMRI representations, we need to apply across-modality human neuroimaging data to conduct the model-brain alignment evaluation. Here we obtained human EEG data from an EEG open dataset, *THINGS EEG2 dataset* (Gifford et al. 2022), including EEG data from 10 healthy human subjects

in a rapid serial visual presentation (RSVP) paradigm. Stimuli were images sized 500×500 pixels from *THINGS dataset* (Hebart et al. 2019), which consists of images of objects on a natural background from 1854 different object concepts. We applied the test set in *THINGS EEG2 dataset* to evaluate the across-modality and also across-subject model-EEG similarity. In this test set, each subject completed 16,000 trials with 200 images from 200 object concepts and 80 repeated trials per images. Subjects viewed one image per trial (100ms).

EEG data were collected using a 64-channel EASYCAP and a BrainVision actiCHamp amplifier. We used already pre-processed data from 17 channels (O1, Oz, O2, PO7, PO3, POz, PO4, PO8, P7, P5, P3, P1, Pz, P2) overlying occipital and parietal cortex. We re-epoched EEG data ranging from stimulus onset to 200ms after onset with a sample frequency of 100 Hz. Thus, the shape of our EEG data matrix for each trial is 17 channels \times 20 time points. Similar to fMRI, we averaged all the repeated trials for each image to obtain more stable EEG signals.

Model architecture and training

Basic architecture of ReAlnet-fMRI We have chosen the state-of-the-art CORnet-S model (Kubilius et al. 2018, 2019) as the foundational architecture for ReAlnet-fMRI. Both CORnet and ReAlnet consist of four visual layers (V1, V2, V4, and IT) and a category decoder layer. Layer V1

performs a 7×7 convolution with a stride of 2, followed by a 3×3 max pooling with a stride of 2, and another 3×3 convolution. Layer V2, V4, and IT each perform two 1×1 convolutions, a bottleneck-style 3×3 convolution with a stride of 2, and a 1×1 convolution. Apart from the initial Layer V1, the other three visual layers include recurrent connections, allowing outputs of a certain layer to be passed through the same layer several times (twice in Layer V2 and IT, and four times in Layer V4).

Image-to-fMRI encoding-based alignment framework In addition to the original CORnet structure, we have added an fMRI generation module designed to construct an image-to-fMRI encoding model for generating human fMRI signals from human visual cortex (Fig. 1A). Each visual layer is connected to a nonlinear $N \times 128$ layer-encoder (Enc-V1, Enc-V2, Enc-V4, and Enc-IT correspond to Layer V1, V2, V4, and IT) that processes through a fully connected network with a ReLU activation. These four layer-encoders are then directly concatenated to form an $N \times 512$ Multi-Layer Visual Encoder, which is subsequently connected to an $N \times 1024$ fMRI encoder through a linear layer to generate the predicted fMRI signals. Here N is the batch size. Therefore, we aim for ReAlnet-fMRI to not only perform the object classification task but also to generate human fMRI signals through the fMRI generation module to learn human fMRI representations when the human subject views the certain image. During this process of predicting human

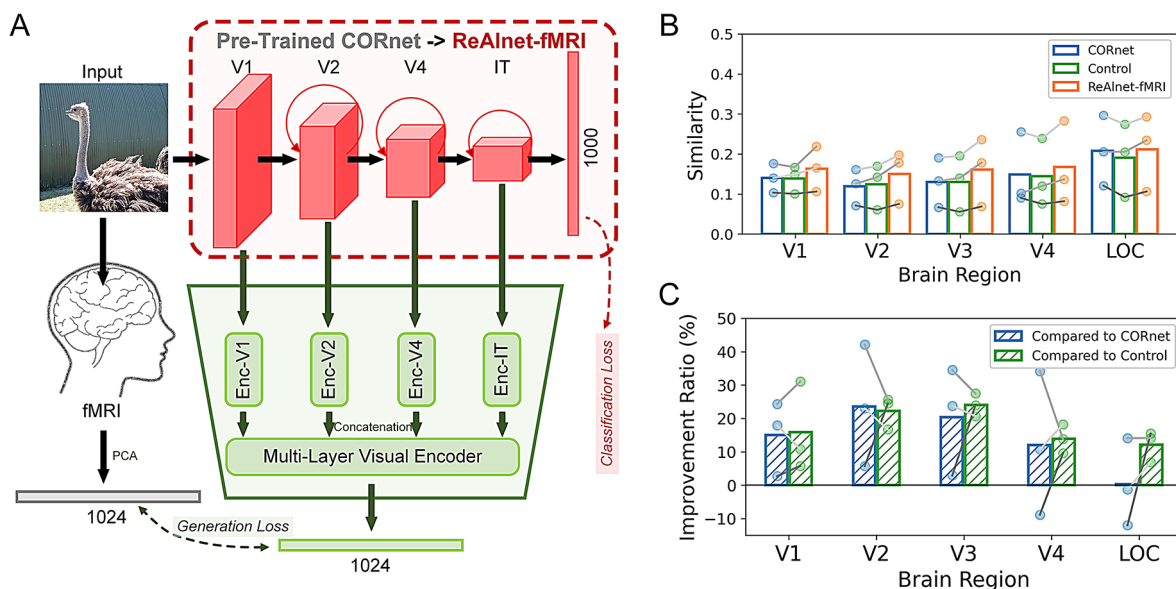


Fig. 1 Human fMRI-optimized ReAlnet-fMRI as a more human brain-like vision model. (A) An overview of ReAlnet-fMRI alignment framework. Adding an additional multi-layer encoder to an ImageNet pre-trained CORnet-S, the outputs contain the category classification results and the generated fMRI signals with two losses, a classification loss and a generation loss. (B) Within-subject representational similarity

between models (CORnet, Control, and ReAlnet-fMRIs) and human fMRI on natural images. (C) Similarity Improvement ratio of within-subject model-fMRI similarity on natural images of ReAlnet-fMRIs compared to other two models. Each circle dot indicates an individual ReAlnet-fMRI. Different shades of grey lines correspond to three different subjects in *Shen fMRI dataset*

fMRI activity, ReAlnet-fMRI's visual layers are poised to effectively extract features more aligned with neural representations.

Similar to the study of ReAlnet (Lu et al. 2024), the training loss L^A of this modified alignment framework consists of a classification loss and a generation loss with a parameter β that determines the relative weighting:

$$L^A = L^C + \beta \cdot L^G$$

L^C represents the standard categorical cross entropy loss for model predictions on ImageNet labels:

$$L^C = - \sum_{i=1}^N y_i \log(p_i)$$

Here, y_i represents the i -th image, and p_i represents the probability that model predicts the i -th image belongs to class i out of 1000 categories. However, some images in *Shen fMRI dataset* were not included in 1,000 categories in ImageNet 1,000 category version. Therefore, we adopt the same strategy as in previous studies (Dapello et al. 2023; Lu et al. 2024), using the labels obtained from the ImageNet pre-trained CORnet without neural alignment as the true labels to stabilize the classification performance of ReAlnet-fMRI.

L^G represents the generation loss including a mean squared error (MSE) loss and a contrastive loss between the generated and real fMRI signals. This contrastive loss is calculated based on the dissimilarity (1 minus Spearman correlation coefficient) between generated and real signals, aiming to bring the generated signals from the same image (positive pairs) closer to the corresponding real human fMRI signals and make the generated signals from different images (negative pairs) more distinct. L^G is calculated as followed:

$$L^G = \frac{1}{N} \sum_{i=1}^N (S_i - \hat{S}_i)^2 + 1 - \frac{1}{N} \sum_{i=1}^N \rho(S_i, \hat{S}_i) + \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j=1, j \neq i}^N \rho(S_i, \hat{S}_j)$$

Here, S_i and \hat{S}_i represent the generated and real fMRI signals corresponding to the i -th image.

Training procedures We trained 3 individual ReAlnet-fMRI independently based on 3 human subjects' fMRI data in *Shen fMRI dataset* independently. Each network was trained to minimize the alignment loss including both classification and generation losses with a static loss weight β of 10, 20, 30, 40 or 50 and a static training rate of 0.00002 for 5 epochs using the Adam optimizer. We used a batch size of 16, meaning the contrastive loss

computed dissimilarities of 256 pairs for each gradient step.

Additionally, the absence of correct ImageNet labels for many images in *Shen fMRI dataset* should indeed decrease the category classification performance on ImageNet. To better control, we trained a control model with $\beta=0$, called Control. We tested the classification accuracy on ImageNet and behavior similarity on Brain-Score (See *Model-Brain Similarity Measurement* section) of CORnet, Control, and ReAlnet-fMRIs at different β values (Figure S1). Although the classification accuracy of ReAlnet-fMRIs decreased compared with CORnet and showed a slight tendency to decrease as β increased, it hardly changed compared with Control. Also, there was no significant different on behavioral similarity between CORnet and ReAlnet-fMRIs when beta is larger than 10, and all ReAlnet-fMRIs showed higher behavioral similarity than Control. In the main text, we show the results of ReAlnet-fMRIs with $\beta=40$ (See the results of ReAlnet-fMRIs with $\beta=10, 20, 30, 50$ in Supplementary Figure S2, Figure S3, Figure S4, Figure S5, Figure S6, Figure S7, Figure S8, Figure S9, Figure S10, Figure S11, Figure S12, Figure S13).

Model-brain similarity measurement

Representational similarity analysis (RSA) (Kriegeskorte et al. 2008) is used for representational comparisons between models and human brain activity. First, we computed representational dissimilarity matrices (RDMs) for models and human fMRI or EEG signals. Then, we calculated Spearman correlation coefficients between model RDMs and human neural RDMs. All RSA analyses were implemented based on NeuroRA toolbox (Lu and Ku 2020).

Model-fMRI similarity To evaluate the within-subject model-fMRI similarity, the shape of each RDM (natural images, artificial shape images, or alphabetical letter images) is 50×50 , 40×40 , or 10×10 in *Shen fMRI dataset* test set. For fMRI RDMs, we calculated 1 minus Pearson correlation coefficient between voxel-wise activation patterns corresponding to any two images as the dissimilarity index in the RDM for each ROI and each subject. For model RDMs, we input 50 natural images, 40 artificial shape images, and 10 alphabetical letter images respectively into each model and obtained latent features from each visual layer. Then, we constructed each layer's RDM by calculating the dissimilarity using 1 minus Pearson correlation coefficient between flattened vectors of latent features corresponding to any two images. To compare the representations, we calculated the Spearman correlation coefficient as

the similarity index between layer-by-layer model RDMs and neural fMRI RDMs corresponding to different ROIs, assigning the final similarity for a certain brain region as the highest similarity result across model layers due to the lack of a clear correspondence between different model layers and brain regions. Similarly, to evaluate the across-subject model-fMRI similarity, the only difference was that we obtained 50×50 fMRI RDMs corresponding to 5 ROIs for each subject in *Horikawa fMRI dataset* test set.

Model-EEG similarity To evaluate the across-modality and across-subject model-EEG similarity, the shape of each RDM is 200×200 , corresponding to 200 images in THINGS EEG2 test set. For EEG RDMs, we applied timepoint-by-timepoint classification-based EEG decoding and used decoding accuracy between two image conditions as the dissimilarity index to construct EEG RDM for each timepoint and each subject. For model RDMs, we input 200 images into each model and obtained the final layer-by-layer model RDMs. To compare the representational similarity temporally, we calculated the Spearman correlation coefficient between layer-by-layer model RDMs and timepoint-by-timepoint neural EEG RDMs.

Model-behavior similarity We measured the model-behavior similarity based on Brain-Score, which is a framework evaluating how similar the model is to the primate visual system (Schrimpf et al. 2020). Here, we applied the behavioral benchmarks (including “Rajalingham-2018public-i2n” assessing the ability of recognizing core objects from visual images, even with various changes in position, size, viewing angle, and background of the objects (Rajalingham et al. 2018) and “Geirhos2021-error_consistency” measuring the similarity of errors made by ANN and human (Geirhos et al. 2021) from Brain-Score platform. We obtained the behavior similarity of both CORnet and ReAlnet-fMRI with different β values. For more detailed information about the behavioral benchmarks in Brain-Score, please refer to the original papers (Geirhos et al. 2021; Rajalingham et al. 2018; Schrimpf et al. 2020).

Model internal representational analysis

To assess the model’s encoding of different object features and explore which object dimension the fMRI-optimized ReAlnet-fMRI encodes more strongly or weakly compared to CORnet, we applied 49 object space dimensions from THINGS (Hebart et al. 2020). These dimensions were derived from large scale human similarity judgments for real-world images of 1854 object concepts by using a data-driven computational model. The model

captured most explained variance in similarity judgments and produced these 49 highly reproducible and meaningful object dimensions that reflect various conceptual and perceptual properties of all objects. Thus, these 49 object dimensions serve as the core dimensions in object space. For more detailed, please refer to the original THINGS paper (Hebart et al. 2020).

Here, our analysis is based on the 200 images in the test set of THINGS EEG2 dataset. These 200 images are involved in THINGS image dataset (but not involved in the model’s training in our study), and we could directly obtain the feature weights along 49 dimensions for all 200 images from the THINGS dataset website (<https://osf.io/jum2f/>). We employed an RDM-based partial Spearman correlation method for the analysis. Specifically, we first computed the RDM for the IT layer (which contained more higher-level information) of each model and 49 feature RDMs based on 200 images (calculating the absolute value differences in feature encoding strength on the same dimension between every two images as dissimilarity). Then, we computed the partial correlation between the model RDM and each feature RDM, regressing out the other 48 feature RDMs, and calculated the square of the partial correlation coefficient as the explained variance of the model by that object dimension. We also conducted the same analysis for the V1-V4 layers, and the corresponding results are shown in Figure S14.

Results

Within-modality & within-subject model-fMRI similarity

First, we would like to confirm that training the model using fMRI signals of humans viewing natural images can make the model’s representations more similar to human fMRI representation at the within-subject level. Based on the analysis of the 50 natural images in the test set from *Shen fMRI dataset*, we calculated the similarity between (1) fMRI RDMs based on 5 different ROIs and RDMs of CORnet based on 4 layers, (2) their fMRI RDMs and RDMs of Control, (3) their fMRI RDMs and RDMs of the subject-matched ReAlnet-fMRI. The results showed that ReAlnet-fMRIs were more similar to human fMRI representations across various visual brain regions compared to both CORnet and Control (Fig. 1B).

We also analyzed the improvement ratio of similarity (compared to CORnet: $(\text{ReAlnet-fMRI} - \text{CORnet})/\text{CORnet}$; compared to Control: $(\text{ReAlnet-fMRI} - \text{Control})/\text{Control}$) (Fig. 1C). The average improvement ratio exceeded 10%, with the highest improvement ratio reaching 43%. Although

there were 1–2 instances where ReAlnet-fMRIs exhibited lower similarity to human V4 and LOC representations compared to CORnet, they were still higher than Control. These few instances of negative improvement are likely due to the lack of ImageNet category labels for some images during the training of ReAlnet-fMRI. Nonetheless, our image-to-fMRI encoding-based alignment framework demonstrated a significant capability to enhance model-fMRI alignment that ReAlnet-fMRIs show significantly higher alignment compared to CORnet and Control.

Furthermore, can ReAlnet-fMRI, which is trained to learn fMRI representations corresponding to natural images, also learn brain representations when humans view other categories of images? To investigate this, we selected two other parts of data from the test set of *Shen fMRI dataset*, which includes 40 artificial shape images and 10 alphabetical letter images. The results showed that for simple shape and letter images, ReAlnet-fMRIs still exhibited higher similarity to human brain representations (Fig. 2A–B). In V4 and LOC brain regions, we observed instances where ReAlnet-fMRIs had lower similarity compared to CORnet and/or Control. This result is primarily driven by the characteristics of the image type. For stimuli such as artificial shapes and letters, which lack rich semantic content, it is inherently challenging to improve alignment in regions like LOC that specialize in high-level semantic processing. Instead, ReAlnet-fMRIs

demonstrate consistent alignment improvement in regions such as V1–V4, which process the low- and mid-level features more prevalent in these stimuli.

It is important to note that the alignment improvements of ReAlnet-fMRIs are strongly influenced by the type of stimuli. For images rich in semantic content, such as natural images, substantial improvements are observed from low- to high-level visual regions. In contrast, for stimuli dominated by low-level features, the improvements are largely restricted to regions specialized in low- and mid-level processing.

Within-modality & across-subject model-fMRI similarity

To further confirm that our individually fMRI-optimized ReAlnet-fMRIs learn not only individual-specific but also more general brain representations across individuals, compared to the original purely image-trained CORnet, we conducted the across-subject model-fMRI similarity analysis on the test set from *Horikawa fMRI dataset*. The results were consistent with the within-subject (subject-matched) similarity results from *Shen fMRI dataset* above, showing that three ReAlnet-fMRI models exhibited higher similarity to the fMRI representations of all five subjects in *Horikawa fMRI dataset* (Fig. 3A). The lower representational similarity of ReAlnet-fMRI compared to CORnet in the LOC

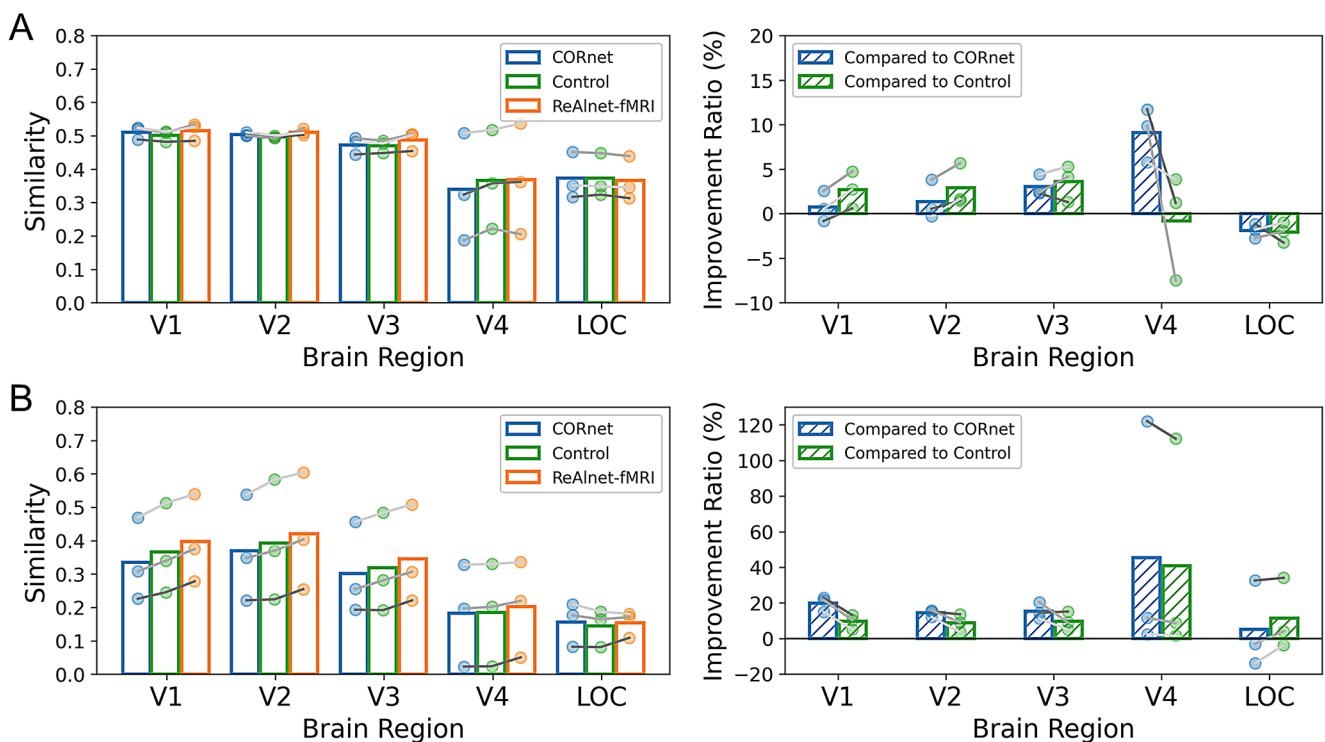


Fig. 2 Within-subject model-fMRI similarity and similarity improvement ratio on (A) artificial shape images and (B) alphabetical letter images. Each circle dot indicates an individual ReAlnet-fMRI. Different shades of grey lines correspond to three different subjects in *Shen fMRI dataset*

region is likely due to the performance limitations of the ReAlnet-fMRI model in encoding high-level visual features. However, it is worth noting that ReAlnet-fMRI still demonstrates significant improvements over the Control model, indicating that the framework effectively enhances alignment with neural data overall.

Additionally, we combined the results of the improvement ratio of three ReAlnet-fMRIs compared to Control tested on two datasets (within- and across-subject) together in Fig. 3B. The combined results further highlight the model's generalization performance across different datasets and subjects. The trends indicate that the similarity improvement across datasets is consistent: for example, the model that exhibits the highest improvement in similarity to V1 when tested on the Shen fMRI dataset also shows the highest improvement in similarity to V1 when tested on the Horikawa dataset.

Across-modality & across-subject model-EEG similarity

Although we observed that ReAlnet-fMRIs exhibit higher similarity to human fMRI representations, it is important to note that ReAlnet-fMRIs are trained based on fMRI signals. This raises the question: are ReAlnet-fMRIs learning just the fMRI representations, or are they capturing broader human brain representations in visual perception? If it is the latter, we should be able to observe that ReAlnet-fMRIs also show higher similarity to across-modality human EEG representations compared to CORnet. To test this, we conducted an across-modality and across subject model-EEG similarity analysis using the EEG data from 10 subjects in the test set of *THINGS EEG2 dataset*. The results showed that ReAlnet-fMRIs have significantly higher similarity to human EEG neural dynamics across all four visual layers than both CORnet and Control without human neural alignment (Fig. 4A).

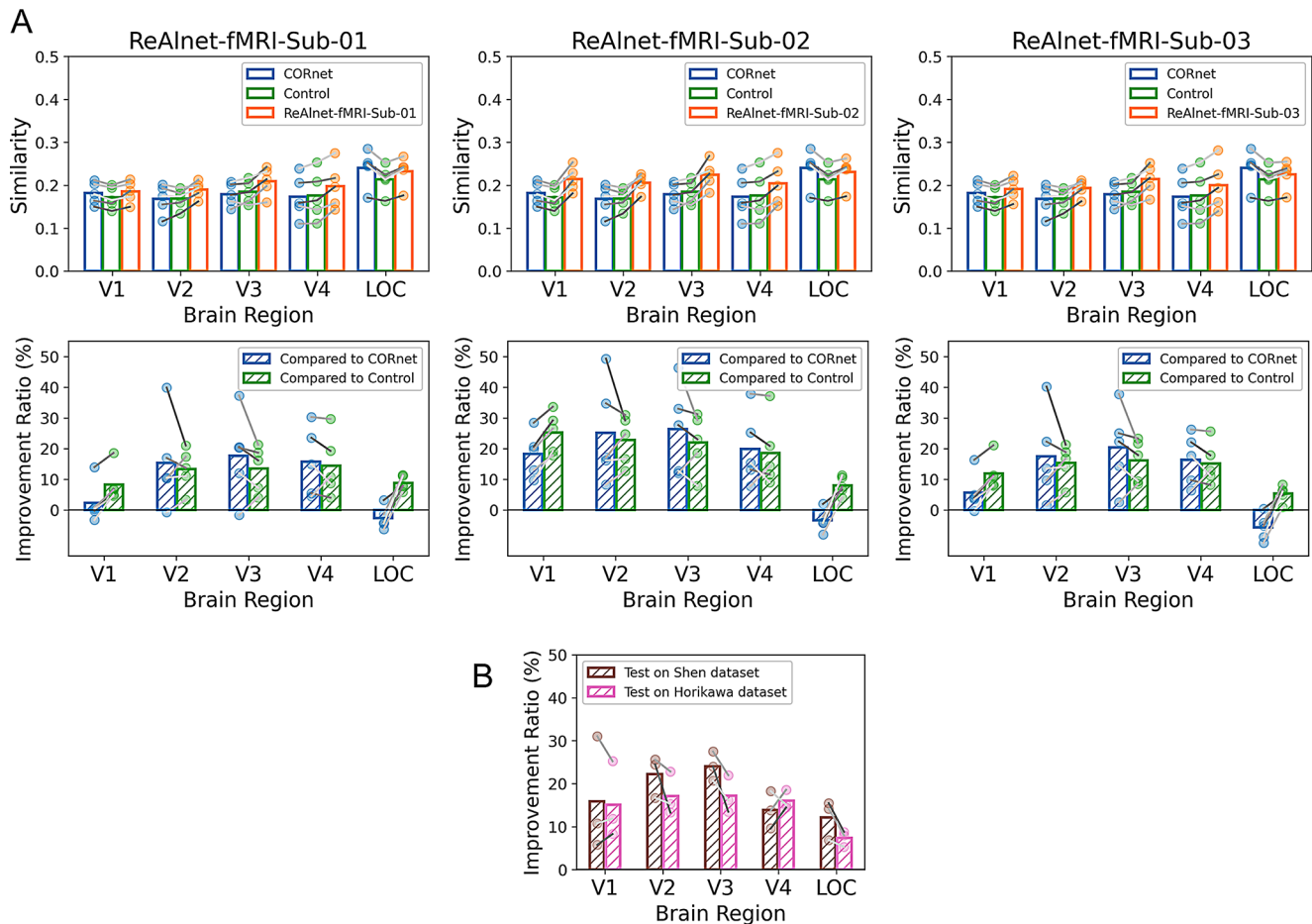


Fig. 3 (A) Across-subject model-fMRI similarity and similarity improvement ratio. Each circle dot indicates a subject from *Horikawa fMRI dataset*. Different shades of grey lines correspond to three different subjects in *Horikawa fMRI dataset*. (B) Similarity improvement ratio of ReAlnet-fMRI compared to the Control (averaging the results

across ten participants from *Horikawa fMRI dataset*) across both *Shen fMRI* and *Horikawa fMRI* datasets. Different shades of grey lines correspond to three personalized ReAlnet-fMRIs trained on *Shen fMRI dataset*

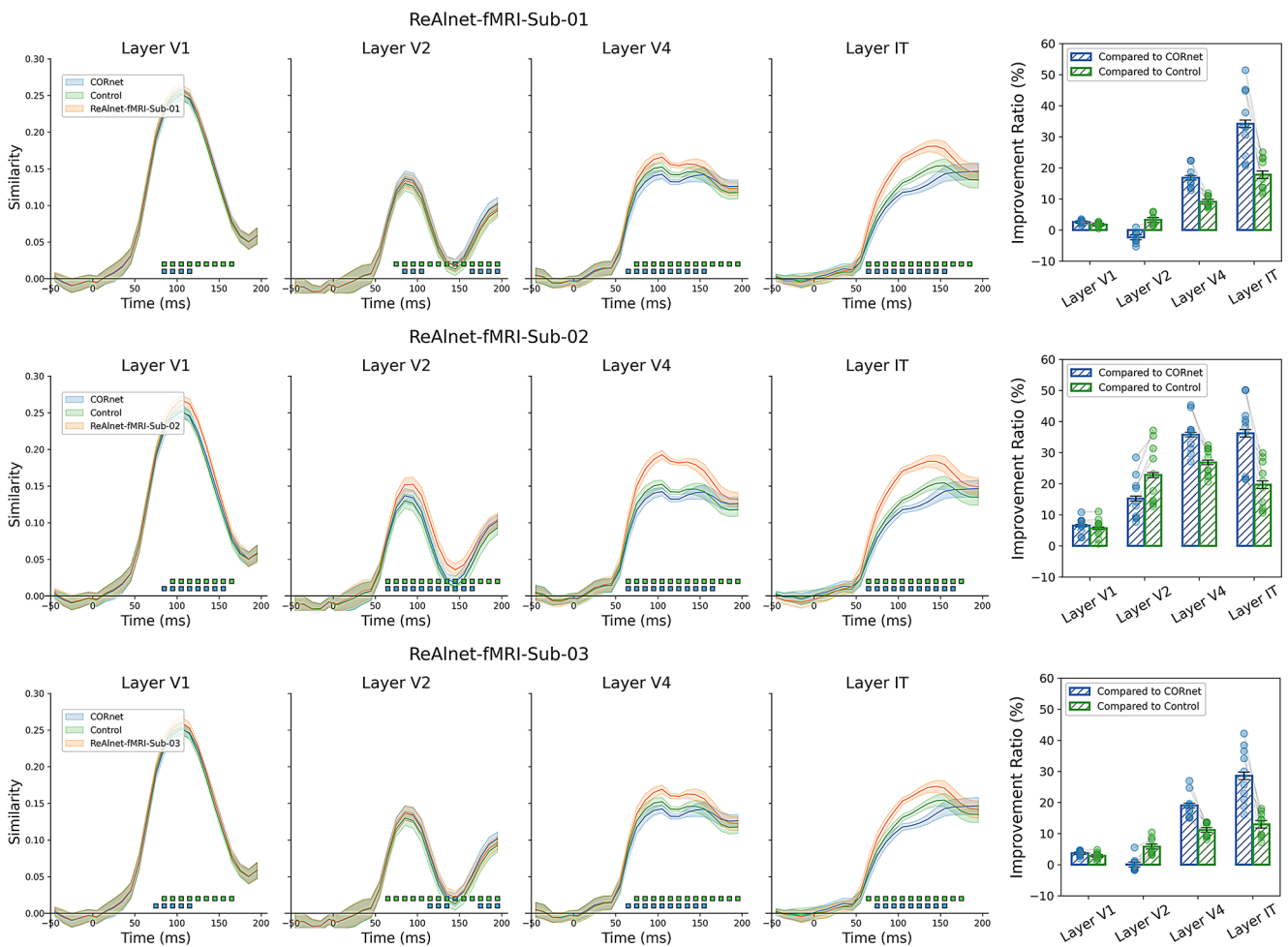


Fig. 4 Across-subject model-EEG similarity. Left shows the temporal similarity results. Blue and green square dots with black outlines at the bottom indicate the timepoints where ReAlnet-fMRI vs. CORnet and

ReAlnet-fMRI vs. Control were significantly different $p < .05$. Shaded area reflects \pm SEM. Right shows the similarity improvement ratio. Each circle dot indicates a subject from *THINGS EEG2* dataset

This indicates that ReAlnet-fMRIs learn broader, cross-modality human brain representations from human fMRI signals, not just fMRI representations.

Furthermore, the representational similarity of ReAlnet-fMRI is notably higher than that of CORnet and Control in higher layers (Layer V4 and IT), while this difference is less pronounced in lower layers (Layer V1 and V2) (Fig. 4B). This could differ from the results in Fig. 3 showing that, compared with lower-level cortical areas (layer V1 and V2), in higher-level cortical areas (layer V4 and LOC), the representational similarity of ReAlnet-fMRI is not notably better than that of CORnet and Control. We would like to emphasize that there is no strict one-to-one mapping between the layers of a model and brain regions. For example, the model's layer V1 is not equivalent to the brain's V1 region. Thus, the differences observed between Figs. 3 and 4 could be due to the distinct definitions and modalities used in these analyses.

Internal representational analysis

Given that our image-to-fMRI encoding-based alignment framework can indeed help CORnet learn human visual representations from human fMRI signals and improve model-brain alignment, the next question is: what are the differences in visual information encoding between the fMRI-optimized ReAlnet-fMRIs and CORnet? Specially, what aspects of information encoding are enhanced by the brain fMRI data in ReAlnet-fMRIs? We conducted an in-depth analysis of the IT layer of both ReAlnet-fMRIs and CORnet to examine their encoding of 49 object dimensions. These internal representational analysis results showed that both ReAlnet-fMRIs and CORnet encode a wide range of object visual features, with stronger encoding observed in dimensions of food-related, animal-related, artificial/hard, electronic/technology, and body/parts (Fig. 5A). Further calculating the differences between ReAlnet-fMRIs and CORnet, we observed that ReAlnet-fMRIs exhibit stronger

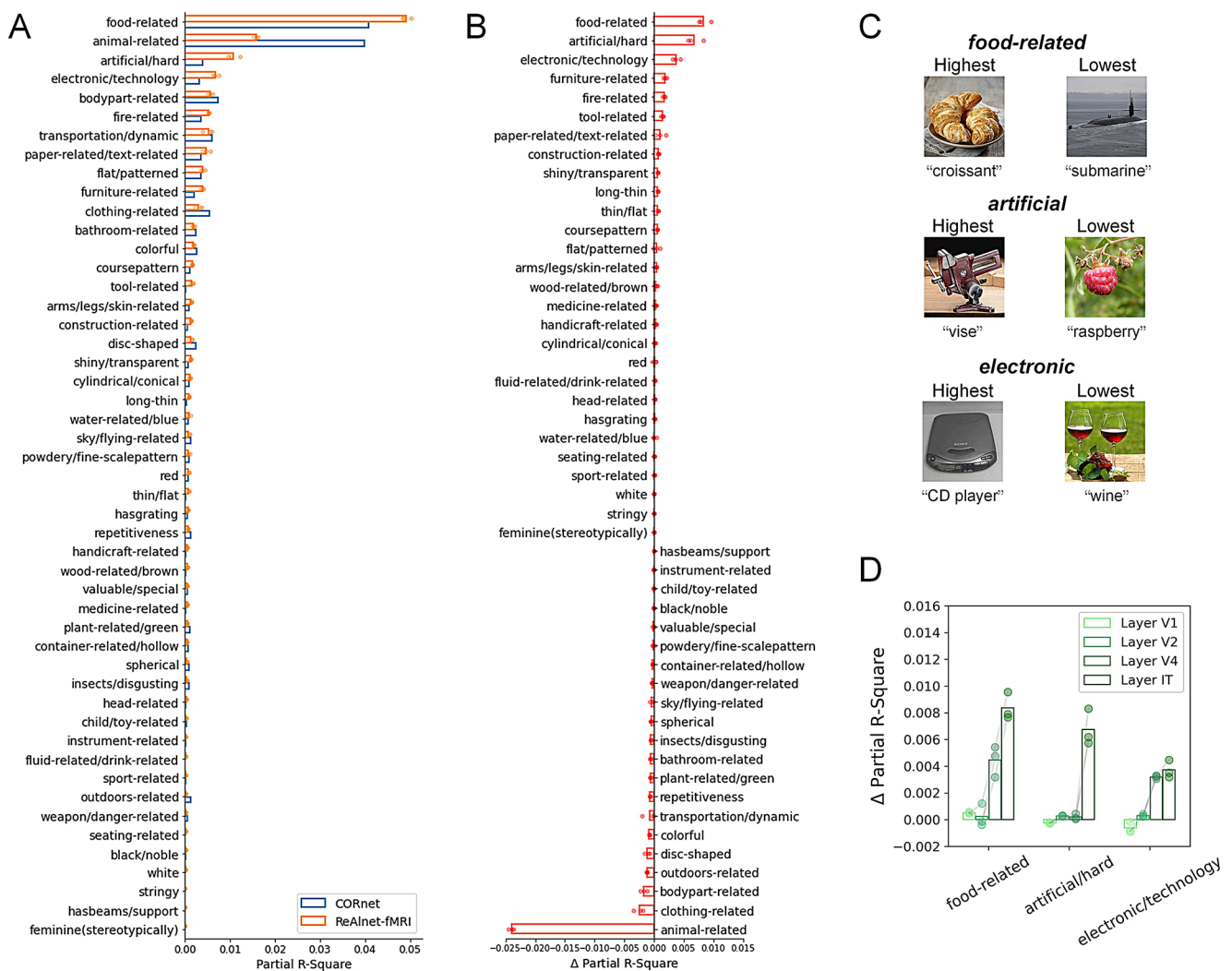


Fig. 5 Internal representations in ReAlnet-fMRIs and CORnet. **(A)** Partial r-square of each object dimension in ReAlnet-fMRIs and CORnet. **(B)** The difference of partial r-square between ReAlnet-fMRIs and CORnet. Each circle dot indicates an individual ReAlnet-fMRI. **(C)**

processing in food-related, artificial/hard, and electronics/technology dimensions, while showing weaker processing in animal-related dimensions compared to CORnet (Fig. 5B). Figure 5C provides some example images with their concepts showing the highest and lowest representations of three top features (food-related, artificial/hard, and electronic/technology). It is important to note that these differences do not imply that ReAlnet-fMRIs or CORnet fail to encode certain object dimensions. Instead, they suggest that models trained solely on image data may not capture as much information related to food, artificial objects, and electronics as models optimized with human neural data. Conversely, they might overly capture animal-related information. More importantly, these results also indicate that the brain encompasses more food-related, artificial/hard, and electronic/technology information, which is effectively

captured and learned by our ReAlnet-fMRIs. Further analysis across all layers in the model suggests that Layer IT shows the highest representational enhancement on these object features (Fig. 5D).

Why does CORnet exhibit weaker encoding of such information? On one hand, CORnet is trained on the 1000-category classification task of ImageNet, an object recognition benchmark that does not specifically require the model to capture higher-level information, such as food-related or artificial/hard object information. Perhaps, we can improve DCNNs by augmenting the original ImageNet 1000 classification task with additional tasks that require more fine-grained recognition, but are common in everyday human experience, such as food and tool categorizations. This may enhance the model's encoding of these important features. On the other hand, information such as food is

often learned through multiple sensory systems in real life, including taste, smell, motion, and even language, rather than through vision alone. CORnet, being a vision-only model, cannot learn these comprehensive encoding patterns as the human brain does. Many human fMRI studies have discovered the evidence of food-related and artificial object information encoding in human brains (Cichy et al. 2014; Jain et al. 2023; Khosla et al. 2022), making it reasonable that ReAlnet-fMRIs, optimized with human fMRI signals, exhibits stronger encoding of such information and more brain-like representations.

Discussion

Our study aimed to teach the current state-of-the-art vision model, CORnet, human fMRI representations by training the model based on an image-to-fMRI encoding-based alignment framework. These human fMRI-optimized models, ReAlnet-fMRIs, enhanced the model's alignment with human brain representations. We evaluated this at different levels, and the results from multiple experiments based on multiple neuroimaging datasets confirmed that this alignment improvement was observed not only in within-modality and within-subject model-fMRI comparisons but also in across-subject model-fMRI and across-modality model-EEG comparisons. This indicates that ReAlnet-fMRIs, to some extent, captured the way how human brains process visual information and utilized brain fMRI data to optimize their weights in a brain-like manner within our alignment training process. This optimization process demonstrated strong generalization across images, image categories, brain imaging modalities, and individuals, ultimately resulting in a comprehensive enhance model-brain alignment.

This study builds upon the multi-layer encoding alignment framework introduced in (Lu et al. 2024) and extends its application to fMRI data, addressing a critical gap in our understanding of its generalizability across neural modalities. While EEG and fMRI both provide valuable insights into neural activity, their fundamental differences in temporal and spatial resolution pose unique challenges for model optimization. One of the key contributions of this study is the validation of the framework's adaptability to a neural modality with distinct characteristics. fMRI data's spatial resolution and higher dimensionality required modifications to the framework, such as PCA for dimensionality reduction, to ensure effective training. By successfully applying the framework to fMRI data, this study demonstrates its flexibility and robustness, establishing it as a versatile tool for model-brain alignment research.

Additionally, we analyzed the differences in internal representations between fMRI-optimized ReAlnet-fMRIs and the purely image-trained CORnet. Interestingly, we observed stronger encoding of information related to food and other several features in ReAlnet-fMRIs. These differences highlight the distinct ways the brain processes visual information compared to models trained purely on image-based pattern recognition.

Why does CORnet exhibit weaker encoding of such information? On one hand, CORnet is trained on the 1000-category classification task of ImageNet, an object recognition benchmark that does not specifically require the model to capture higher-level information, such as food-related or artificial/hard object information. Perhaps, we can improve DCNNs by augmenting the original ImageNet 1000 classification task with additional tasks that require more fine-grained recognition, but are common in everyday human experience, such as food and tool categorizations. This may enhance the model's encoding of these important features. On the other hand, information such as food is often learned through multiple sensory systems in real life, including taste, smell, motion, and even language, rather than through vision alone. CORnet, being a vision-only model, cannot learn these comprehensive encoding patterns as the human brain does. Many human fMRI studies have discovered the evidence of food-related and artificial object information encoding in human brains (Cichy et al. 2014; Jain et al. 2023; Khosla et al. 2022), making it reasonable that ReAlnet-fMRIs, optimized with human fMRI signals, exhibits stronger encoding of such information and more brain-like representations.

Due to the lack of clear correspondence between different model layers and brain regions, we proposed this multi-layer encoding-based alignment method. Interestingly, the model-brain similarity results show different improvement patterns across fMRI and EEG: while model-fMRI similarity improvements were reduced in higher-level brain areas, model-EEG similarity improvements were more pronounced in higher-level model layers. Several neurophysiological and methodological factors may contribute to this divergence. First, the number of voxels varies across brain regions, with early visual areas such as V1, V2, and V3 encompassing more voxels than higher-level areas like V4 and LOC. As a result, during the training, the model may have been more strongly guided by signals from early visual cortex, leading to a greater emphasis on lower-level features. Although the model ultimately learned brain-like visual representations across multiple levels, its exposure was biased toward early visual signals. Second, EEG inherently has lower spatial resolution and is less sensitive to fine-grained, low-level visual features such as orientation

or spatial frequency. Instead, EEG is often better at capturing categorical, high-level information like faces, objects, or scenes. Indeed, previous EEG decoding studies have shown that decoding accuracy is higher for categorical-level features than for low-level visual properties (higher decoding accuracy) than low-level visual features (Bae and Luck 2018; Khaligh-Razavi et al. 2018; Lu and Golomb 2023a, b; Lu and Ku 2023; Wang et al. 2022). This may help explain why ReAlnet-fMRI - although trained on fMRI - shows stronger alignment with high-level EEG signals in later model layers such as V4 and IT layers. These observed differences suggest that representational alignment results could be influenced and interpreted in light of the characteristics of the neural modality being used, which provides insights to future studies in model-brain alignment. Looking ahead, integrating multiple neural data types (e.g., fMRI and EEG) may offer a more comprehensive path toward optimizing neural networks to better learn the full range of human visual representations.

To achieve more brain-like visual models, besides approaches like our current work that directly use neural data for model optimization (Dapello et al. 2023; Federer et al. 2020; Li et al. 2019; Lu et al. 2024; Pirlot et al. 2022; Shao et al. 2024), there are other strategies as well. These include modifying the model's architecture by constructing dual-way pathways models similar to the brain's visual pathways (Bai et al. 2017; Choi et al. 2023; Han and Sereno 2022, 2023; Sun et al. 2017), adding feedback pathways (Konkle and Alvarez 2023), incorporating topographic constraints (Finzi et al. 2022; Lee et al. 2020; Lu et al. 2023; Margalit et al. 2023), or changing the model's training tasks by using self-supervised training or training the model in a richer 3D environment (Konkle and Alvarez 2022; Prince et al. 2023). These different approaches to realizing brain-inspired AI are not mutually exclusive; in fact, they may complement each other. Current researchers from both computer science and neuroscience are exploring various angles to word toward this goal.

With the increasing adoption of artificial neural networks as effective tools to study human cognitive mechanisms, cognitive psychologists and neuroscientists have employed these models in two ways: Using computational models to extract specific features from stimuli to identify how the human brain encodes these features (Cichy et al. 2016; Cichy and Kaiser 2019; McMahon et al. 2023). Leveraging pre-trained models to perform reverse engineering manipulations to infer cognitive processing mechanisms (Kanwisher et al. 2023; Lu and Ku 2023). These approaches highlight the need for models that closely mirror human brain processing as foundational tools for studying cognitive functions. Our ReAlnet-fMRI demonstrates, across multiple dimensions, that it aligns more closely with human

brain representations compared to models trained purely on images. This makes it a more effective computational tool for current cognitive computational neuroscience research.

Additionally, we acknowledge a key limitation of our work: due to the mismatch between the images in the current fMRI dataset and the ImageNet dataset, ReAlnet-fMRIs did not show improved performance on object classification tasks. As a result, we cannot yet provide evidence that our work improves model robustness. However, a related study using similar fMRI-optimized frameworks have already demonstrated that this approach can significantly enhance model robustness (Shao et al. 2024). Therefore, we are confident that our alignment framework has the potential to not only serve as a better tool for understanding the brain but also improve model robustness in future work, provided the datasets are more suitable for alignment.

For the future directions, from a data perspective, the primary limitations of our current study stem from (1) the relatively smaller sample size of neural datasets compared to image datasets with vast samples, and (2) the lack of shared labels between different datasets, such as the absence of some ImageNet category labels for images used in human neuroimaging studies. These limitations restrict further enhancement of ReAlnet-fMRI's similarity to the human brain and reduce its classification performance on ImageNet. From a technical perspective, future research may need to focus on (1) more effectively learning the alignment of models with the human brain using small-sample neural data, and (2) employing self-supervised or unsupervised learning methods that do not require category labels for model training.

In future work, we hope to extend this multi-layer alignment framework to broader domains, such as language models, auditory models, and even cross-modal models. Certainly, these extended applications will also necessitate corresponding neural data collection efforts. Additionally, beyond focusing on model-brain alignment, it is worth testing whether models optimized using neural data show performance improvements in tasks where traditional models perform poorly. For instance, testing the robustness of models under different noise conditions to see if there is an improvement (some studies have already found brain-aligned models to have higher adversarial robustness) (Dapello et al. 2023; Shao et al. 2024), and evaluating whether the models exhibit stronger abstraction and generalization capabilities.

In summary, we utilized a neural alignment framework capable of synchronously optimizing multiple layers of CORnet and trained three individualized ReAlnet-fMRI models based on this framework. Through a series of detailed and comprehensive evaluations, we demonstrated that ReAlnet-fMRIs are more human brain-like visual

models that exhibit significantly enhance model-brain alignment. In addition, the training process of learning fMRI representations enabled ReAlnet-fMRIs to show stronger encoding in object dimensions such as food-related features compared to CORnet. We hope that our work helps bridge the gap between AI in computer vision and human visual neuroscience, thereby achieving more brain-like intelligence and understanding the brain from an AI perspective.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11571-025-10252-y>.

Acknowledgements We thank Julie Golomb for helpful discussions. We thank the Ohio Supercomputer Center and Georgia Stuart for providing the essential computing resources and support.

Author contributions Zitong Lu: Conceptualization, Methodology, Visualization, Investigation, Writing- Original draft preparation. Yile Wang: Methodology, Investigation, Writing- Reviewing and Editin.

Data availability The code that support the findings of this study will be openly available in GitHub at <https://github.com/ZitongLu1996/ReAlnet>.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bae GY, Luck SJ (2018) Dissociable decoding of Spatial attention and working memory from EEG oscillations and sustained potentials. *J Neurosci* 38(2):409–422
- Bai S, Li Z, Hou J (2017) Learning two-pathway convolutional neural networks for categorizing scene images. *Multimedia Tools Appl* 76(15):16145–16162
- Bao P, She L, McGill M, Tsao DY (2020) A map of object space in primate inferotemporal cortex. *Nature* 583(7814):103–108
- Choi M, Han K, Wang X, Zhang Y, Liu Z (2023) A Dual-Stream neural network explains the functional segregation of dorsal and ventral visual pathways in human brains. *Adv Neural Inform Process Syst (NeurIPS)*, 36. <http://arxiv.org/abs/2310.13849>
- Cichy RM, Kaiser D (2019) Deep neural networks as scientific models. *Trends Cogn Sci* 23(4):305–317
- Cichy RM, Pantazis D, Oliva A (2014) Resolving human object recognition in space and time. *Nat Neurosci* 17(3):455–462
- Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A (2016) Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci Rep* 6(1):1–13
- Coggan DD, Tong F (2023) Spikiness and animacy as potential organizing principles of human ventral visual cortex. *Cereb Cortex* 33(13):8194–8217
- Dapello J, Kar K, Schrimpf M, Geary RB, Ferguson M, Cox DD, DiCarlo JJ (2023) Aligning Model and Macaque Inferior Temporal Cortex Representations Improves Model-to-Human Behavioral Alignment and Adversarial Robustness. *International Conference on Learning Representations (ICLR)*, 12
- Federer C, Xu H, Fyshe A, Zylberberg J (2020) Improved object recognition using neural networks trained to mimic the brain's statistical properties. *Neural Netw* 131:103–114
- Finzi D, Margalit E, Kay K, Yamins DLK, Grill-Spector K (2022) Topographic DCNNs trained on a single self-supervised task capture the functional organization of cortex into visual processing streams. *NeurIPS 2022 Workshop SVRHM*
- Geirhos R, Narayanappa K, Mitzkus B, Thieringer T, Bethge M, Wichmann FA, Brendel W (2021) Partial success in closing the gap between human and machine vision. *Adv Neural Inf Process Syst* 34:23885–23899. <https://github.com/bethgelab/model-vs-human/>
- Gifford AT, Dwivedi K, Roig G, Cichy RM (2022) A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage* 264:119754
- Grill-Spector K, Kourtzi Z, Kanwisher N (2001) The lateral occipital complex and its role in object recognition. *Vision Res* 41(10–11):1409–1422
- Güçlü U, van Gerven MAJ (2015) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J Neurosci* 35(27):10005–10014
- Han Z, Sereno A (2022) Modeling the ventral and dorsal cortical visual pathways using artificial neural networks. *Neural Comput* 34(1):138–171
- Han Z, Sereno A (2023) Identifying and localizing multiple objects using artificial ventral and dorsal cortical visual pathways. *Neural Comput* 35(2):249–275
- Hebart MN, Dickter AH, Kidder A, Kwok WY, Corriveau A, Van Wicklin C, Baker CI (2019) THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS ONE* 14(10):1–24
- Hebart MN, Zheng CY, Pereira F, Baker CI (2020) Revealing the multi-dimensional mental representations of natural objects underlying human similarity judgements. *Nat Hum Behav* 4(11):1173–1185
- Horikawa T, Kamitani Y (2017) Generic decoding of seen and imagined objects using hierarchical visual features. *Nat Commun* 8:15037
- Jain N, Wang A, Henderson MM, Lin R, Prince JS, Tarr MJ, Wehbe L (2023) Selectivity for food in human ventral visual cortex. *Commun Biology* 6(1):1–14
- Kanwisher N, Khosla M, Dobs K (2023) Using artificial neural networks to ask 'why' questions of Minds and brains. *Trends Neurosci* 46(3):240–254
- Khaligh-Razavi S-M, Cichy RM, Pantazis D, Oliva A (2018) Tracking the Spatiotemporal neural dynamics of Real-world object size and animacy in the human brain. *J Cogn Neurosci* 30(11):1559–1576
- Khosla M, Ratan Murty NA, Kanwisher N (2022) A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Curr Biol* 32(19):4159–4171e9
- Kietzmann TC, Spoerer CJ, Sörensen LKA, Cichy RM, Hauk O, Kriegeskorte N (2019) Recurrence is required to capture the

- representational dynamics of the human visual system. *Proc Natl Acad Sci USA* 116(43):21854–21863
- Konkle T, Alvarez GA (2022) A self-supervised domain-general learning framework for human ventral stream representation. *Nat Commun* 13(1):1–12
- Konkle T, Alvarez G (2023) Cognitive steering in deep neural networks via Long-Range modulatory feedback connections. *Adv Neural Inform Process Syst (NeurIPS)*, 36
- Konkle T, Caramazza A (2013) Tripartite organization of the ventral stream by animacy and object size. *J Neurosci* 33(25):10235–10242
- Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci*, 4
- Kubilius J, Schrimpf M, Nayebi A, Bear D, Yamins DLK, DiCarlo JJ (2018) CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *BioRxiv*
- Kubilius J, Schrimpf M, Kar K, Rajalingham R, Hong H, Majaj NJ, Issa EB, Bashivan P, Prescott-Roy J, Schmidt K, Nayebi A, Bear D, Yamins DLK, DiCarlo JJ (2019) Brain-Like object recognition with High-Performing shallow recurrent ANNs. *Adv Neural Inform Process Syst (NeurIPS)*, 32
- Lecun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444
- Lee H, Margalit E, Jozwik KM, Cohen MA, Kanwisher N, Yamins DLK, DiCarlo JJ (2020) Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *BioRxiv*
- Li Z, Brendel W, Walker EY, Cobos E, Muhammad T, Reimer J, Bethge M, Sinz FH, Pitkow X, Tolias AS (2019) Learning from brains how to regularize machines. *Adv Neural Inform Process Syst (NeurIPS)*, 32
- Lu Z, Golomb JD (2023a) Generate your neural signals from mine: individual-to-individual EEG converters. *Proceedings of the Annual Meeting of the Cognitive Science Society* 45
- Lu Z, Golomb JD (2023b) Human EEG and artificial neural networks reveal disentangled representations of object real-world size in natural images. *BioRxiv*
- Lu Z, Ku Y (2020) NeuroRA: A python toolbox of representational analysis from Multi-Modal neural data. *Front Neuroinformatics* 14:61
- Lu Z, Ku Y (2023) Bridging the gap between EEG and DCNNs reveals a fatigue mechanism of facial repetition suppression. *IScience* 26(12):108501
- Lu Z, Doerig A, Bosch V, Krahmer B, Kaiser D, Cichy RM, Kietzmann TC (2023) End-to-end topographic networks as models of cortical map formation and human visual behaviour: moving beyond convolutions. *ArXiv*
- Lu Z, Wang Y, Golomb JD (2024) Achieving More Human Brain-Like Vision via Human Neural Representational Alignment. *ArXiv*
- Margalit E, Lee H, Finzi D, DiCarlo JJ, Grill-Spector K, Yamins DLK (2023) A Unifying Principle for the Functional Organization of Visual Cortex. *BioRxiv*
- McMahon E, Bonner MF, Isik L (2023) Hierarchical organization of social action features along the lateral visual pathway. *Curr Biol* 33(23):5035–5047e8
- Pirlot C, Gerum RC, Efird C, Zylberberg J, Fyshe A (2022) Improving the Accuracy and Robustness of CNNs Using a Deep CCA Neural Data Regularizer. *ArXiv*
- Prince JS, Alvarez GA, Konkle T (2023) A contrastive coding account of category selectivity in the ventral visual stream. *BioRxiv*
- Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ (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(33):7255–7269
- Safarani S, Nix A, Willeke K, Cadena SA, Restivo K, Denfield G, Tolias AS, Sinz FH (2021) Towards robust vision by multi-task learning on monkey visual cortex. *Advances in Neural Information Processing Systems (NeurIPS)*, 34
- Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, Kar K, Bashivan P, Prescott-Roy J, Geiger F, Schmidt K, Yamins DLK, DiCarlo JJ (2020) Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *BioRxiv*
- Shao Z, Ma L, Li B, Beck DM (2024) Leveraging the Human Ventral Visual Stream to Improve Neural Network Robustness. *ArXiv*
- Shen G, Horikawa T, Majima K, Kamitani Y (2019) Deep image reconstruction from human brain activity. *PLoS Comput Biol*, 15(1), e1006633
- Sun T, Wang Y, Yang J, Hu X (2017) Convolution neural networks with two pathways for image style recognition. *IEEE Trans Image Process* 26(9):4102–4113
- Wang R, Janini D, Konkle T (2022) Mid-level feature differences support early animacy and object size distinctions: evidence from electroencephalography decoding. *J Cogn Neurosci* 34(9):1670–1680
- Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci USA* 111(23):8619–8624
- Yargholi E, de Baeck HO (2023) Category Trumps shape as an organizational principle of object space in the human occipitotemporal cortex. *J Neurosci* 43(16):2960–2972
- Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) Understanding Neural Networks Through Deep Visualization. *ArXiv*
- Zeiler M, Fergus R (2014) Visualizing and Understanding Convolutional Networks. *Proceedings of 13th European Conference on Computer Vision*, 8689, 808–833

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.