FOREGROUND-ATTENTION IN NEURAL DECODING: GUIDING LOOP-ENC-DEC TO RECONSTRUCT VISUAL STIMULUS IMAGES FROM FMRI

Anonymous authors
Paper under double-blind review

ABSTRACT

The reconstruction of visual stimulus images from functional Magnetic Resonance Imaging (fMRI) has received extensive attention in recent years, which provides a possibility to interpret the human brain. Due to the high-dimensional and high-noise characteristics of fMRI data, how to extract stable, reliable and useful information from fMRI data for image reconstruction has become a challenging problem. Inspired by the mechanism of human visual attention, in this paper, we propose a novel method of reconstructing visual stimulus images, which first decodes the distribution of visual attention from fMRI, and then reconstructs the visual images guided by visual attention. We define visual attention as foreground attention (F-attention). Because the human brain is strongly wound into sulci and gyri, some spatially adjacent voxels are not connected in practice. Therefore, it is necessary to consider the global information when decoding fMRI, so we introduce the self-attention module for capturing global information into the process of decoding F-attention. In addition, in order to obtain more loss constraints in the training process of encoder-decoder, we also propose a new training strategy called Loop-Enc-Dec. The experimental results show that the F-attention decoder decodes the visual attention from fMRI successfully, and the Loop-Enc-Dec guided by F-attention can also well reconstruct the visual stimulus images.

1 INTRODUCTION

In recent years, reconstructing visual stimulus images from fMRI has gradually gained attention, which provides the possibility of "mind reading" in the future (Fig. 1). Existing work has shown that there is a certain mapping relationship between visual stimuli and brain activity (Poldrack & Farah, 2015), which provides us with the possibility and basis for reconstructing visual stimuli from fMRI data. The fMRI data which is collected from the human brain records the variations in blood oxygen level dependent (BOLD) and reflects the activity of nerves in human brain. Through the analysis of fMRI data, the correlation between brain activity and visual tasks can be explored, helping us to understand human visual mechanism better.

Prior studies. In recent years, there has been a lot of research in this field. The main methods can be roughly divided into the following two categories: **Linear regression models**, which encode fMRI data into image pixel values, and finally achieve image reconstruction. **Deep learing models**, such as DCNN and GAN model frameworks.

For the first type of methods, Yoichi Miyawaki et al. applied a multi-scale linear weighting model to predict the pixel value of each image and obtain the reconstruction results of the black and white simple images (Miyawaki et al., 2008). Fujiwara et al. proposed to use Bayesian canonical correlation analysis (BCCA) to build a reconstruction model, which extracts image information from measured data and reconstruct images (Fujiwara et al., 2013). Marcel AJ van Gerve et al. adopted a hierarchical generative model composed of Boltzmann machines with restricted conditions, which explores the reconstruction of feature hierarchies based on learning (Van Gerven et al., 2010). Yu et al. advanced a correlation network framework that could be flexibly combined with diverse pattern representation models. They revealed the effective connectivity of human brain and reconstructed the visual stimulus images (Yu & Zheng, 2017; Yu et al., 2018).

For the second type of methods, Guohua Shen et al. proposed an end-to-end deep convolution model to reconstruct visual stimulus images directly by using fMRI data, and at the same time they used GAN framework to optimize the training of the model (Shen et al., 2019a). In response to the current problem of the lack of data sets used to reconstruct stimulus images from fMRI data, Roman Beliy et al. suggest an approach for self-supervised training on unlabeled fMRI data (Beliy et al., 2019). Yunfeng Lin et al. proposed a model called DCNN-GAN by combining a reconstruction network and a GAN module. They utilized the CNN for hierarchical feature extraction and the DCNN-GAN to reconstruct more realistic images (Lin et al., 2019). Tao Fang et al. defined a Shape-Semantic GAN, considering the functional differences of different visual cortex areas, and reconstruct visual stimulus images under the guidance of shape and semantic information (Fang et al., 2020). The ideas of DCNN and GAN have also been applied to various other model frameworks to achieve the reconstruction of visual stimulus images (Seeliger et al., 2018; VanRullen & Reddy, 2019; Zhang et al., 2020; Qiao et al., 2020).

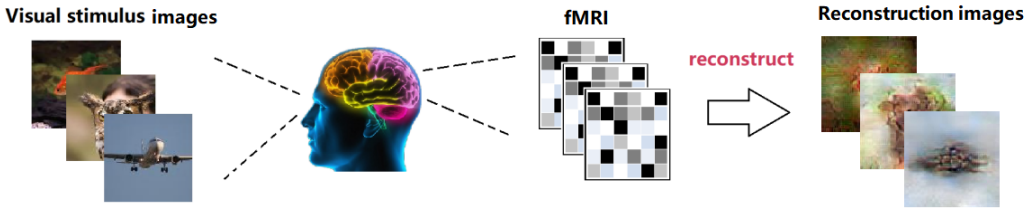

Figure 1: The work in this paper. The purpose of this article is to extract useful information from the visual cortex of the human brain, and build a model to reconstruct visual stimulus images from fMRI data by using the uesful information.

Our contributions. Because of the complex neural information in the visual cortex, how to extract valuable information from high-dimensional fMRI data has been a challenging problem. Inspired by human visual attention mechanism, we introduce visual attention to this work for the first time. We first decode the human visual attention from a small amount of fMRI data successfully, and then use visual attention to guide the work of reconstructing visual stimulus images from fMRI. Existing studies have shown that the human visual system is more inclined to focus on prominent objects, and the neural representation of these prominent objects in the brain is more obvious (Ungerleider & G, 2000; Braun et al., 2001). The distribution of attention leads to an information bottleneck, only the most prominent objects are allowed to appear in the inferior temporal cortex, especially the ventral visual stream that encodes the identity of the object. The visual attention mechanism is crucial for simulating the neural response of the higher visual system (Poggio & Anselmi, 2016; Khosla et al., 2020). Based on the above conclusions, we propose Foreground-attention (F-attention). We abstract the distribution of human visual attention as the most prominent foreground object distribution area in the images. Through the constructed F-attention decoder, the fMRI corresponding to the natural stimulus images can be decoded into visual attention distribution (Fig. 2a).

Because human brain is strongly wound into sulci and gyri, some spatially adjacent voxels are not directly connected in the human brain (Tallinen et al., 2014), the correlation between voxels is not only determined by location. Therefore, we need to consider global information when decoding fMRI, so we first introduce the self-attention module (Vaswani et al., 2017; Wang et al., 2018; Zhang et al., 2019) into the fMRI decoding process, the self-attention module can capture global information of the input data and expand the receptive field (Fig. 3c).

At the same time, we propose a new encoding and decoding framework called Loop-Enc-Dec (Fig. 2b,c). The first step we pre-train the fMRI encoder on the data set to realize the encoding process from images to fMRI. The second step we train the end-to-end encoder-decoder model under the guidance of F-attention, and input the images decoded by the decoder into the fMRI encoder for encoding to add the re-encoding constraint in the loss function, that is called "Loop-Enc-Dec". In the training process, natural pictures without fMRI data are also added for end-to-end self-enc-dec training, which increases the stability and generalization of the model. After experimental evaluation, the performance of Loop-Enc-Dec under the guidance of F-attention is better than that of mainstream methods. The main contributions of this article are as follows:

- As far as we know, we first introduced the visual attention mechanism to the work of reconstructing visual stimulus images from fMRI. According to neuroscience research, we abstract visual attention as the attention distribution of foreground objects (F-attention), and build a visual attention decoder to decode the human visual attention distribution from fMRI successfully.
- To our best knowledge, we first introduced the self-attention module into the work of fMRI decoding to capture global information of fMRI data.
- We propose a new enc-dec framework called Loop-Enc-Dec, guided by F-attention, which adds more loss constraints to the training of the image reconstruction model. Evaluating the quality of reconstructed images, our method outperforms previous works.

2 METHODS

In this section, we first introduce the data set used in the experiments, and then give a clear definition of the problem to be solved in this paper. Next we introduce the details of F-attention and Loop-Enc-Dec framework proposed in this article, as well as the step-by-step training strategy for the model. Finally, our evaluation method is given.

2.1 fMRI DATA SET

We use a publicly available benchmark data sets (Horikawa & Kamitani, 2017), which is widely used in the work of visual stimulus image reconstruction. In the image presentation experiment, fMRI signals were measured while subjects viewed a sequence of object images from ImageNet (Deng et al., 2009). The image presentation experiment consisted of two sessions: the training image session and the testing image session. In the training image session, 1200 images from 150 object categories (8 images from each category) were each presented once. In the testing image session, 50 images from 50 object categories (one image from each category) were each presented 35 times. The data collectors performed their analysis for each combination of feature types/layers and brain regions of interest (ROIs; V1–V4, the lateral occipital complex (LOC), fusiform face area (FFA), parahippocampal place area (PPA), lower visual cortex (LVC; V1–V3), higher visual cortex (HVC; covering regions around LOC, FFA and PPA) and the entire visual cortex (VC; covering all of the visual subareas listed above), the voxel size is 3×3×3 mm^3) (Horikawa & Kamitani, 2017).

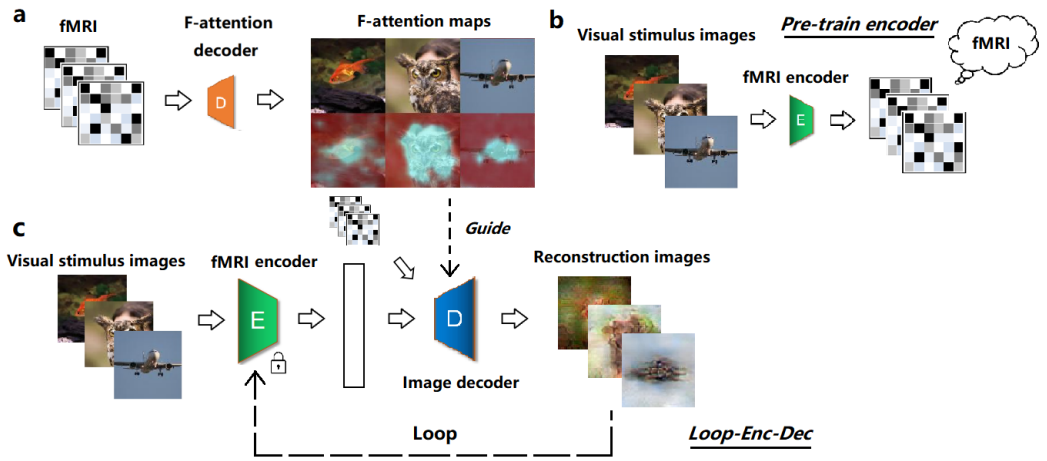

Figure 2: The overall model framework of the paper. (a) The overall structure of F-attention decoder. (b)(c) The Loop-Enc-Dec. We first pre-train a fMRI encoder to encode images to fMRI, and then use an image decoder guided by F-attention to reconstruct images, during the training, decoded images loop back to fMRI encoder.

2.2 PROBLEM DEFINITION

The problem we need to solve in this paper is to reconstruct the visual stimulus images from the neural representation fMRI of the human brain (Fig. 1). While the subjects are seeing the visual stimulus images, the fMRI data corresponding to the human brain is collected. After preprocessing, the fMRI is used as input to construct a model to reconstruct the visual stimulus images seen by the subjects. The visual stimulus image is defined as y, and the corresponding fMRI is defined as x, then the problem is simplified to construct a reconstruction model f,

$$y_i = f(x_i)|y_i \in M, x_i \in F \tag{1}$$

M stands for visual stimulus image data set, F stands for fMRI data set.

2.3 PROPOSED MODEL

Summarizing the proposed model. We first trian the F-attention decoder on fMRI data(Fig. 2a), then pre-train the image encoder(Fig. 2b), next we fix the weight of the encoder, train the decoder under the guidance of F-attention(Fig. 2c). When testing, we first use fMRI as the input of F-attention decoder to decode the distribution of F-attention, then input the fMRI into the image decoder to reconstruct the visual stimulus images guided by corresponding F-attention.

2.3.1 F-ATTENTION

In order to get more useful information from fMRI data to guide the work of visual stimulus image reconstruction, we introduced the human visual attention mechanism into the work. We abstracted the human visual attention as foreground attention (F-attention) based on neuroscience research (Poggio & Anselmi, 2016; Khosla et al., 2020). By constructing a F-attention decoder, we decode human visual attention distribution from fMRI data.

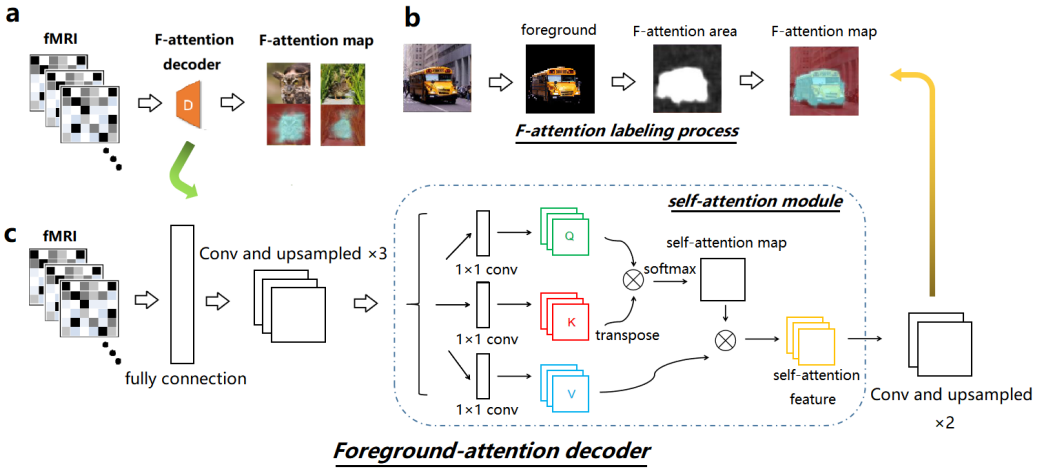

Figure 3: Foreground-attention. (a)The overall structure of F-attention decoder, through the F-attention decoder, we can decode the human visual attention distribution from the visual neural representation of the brain. (b)The human visual attention distribution labeling process of training set. (c)The detailed structure of F-attention decoder.

Visual attention labeling. For the purpose of obtaining the human visual attention distribution of the training set, we first label the train data (Fig. 3b). According to the definition of F-attention, we first extract the prominent objects in the natural images, and then perform the binary value on the images. Through the transformation operation, the main distribution areas of F-attention are obtained. Then we assigned attention weights to different regions in the images (the bluer the color, the higher the attention weights, in the train set, the weights of the foreground object areas are 1, the background areas are 0), the final F-attention maps are obtained as the human visual attention labels of the training set.

F-attention decoder. Then we construct F-attention-decoder to realize the decoding process from fMRI data to human visual attention(Fig. 3c). First, the preprocessed fMRI data is integrated into a vector, then the vector is sequentially connected to a fully connection module, convolution and upsampling modules, a self-attention module, convolution and upsampling modules, and we finally get the F-attention map. The self-attention module calculation formula is as follows:

$$Self - Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

In the training process, we use the mean square error loss (MSE) to constrain. After 400 rounds of iterations, the error decreases and stabilizes. The loss function is as follows:

$$Loss_{F-attention}(FA_i, FA_i^*) \propto \sum (FA_i - FA_i^*)^2 \tag{3}$$

Where FA_i represents the visual attention distribution corresponding to the images in the training set, and FA_i^* stands for the visual attention distribution decoded by the F-attention decoder.

2.3.2 LOOP-ENC-DEC

The Loop-Enc-Dec reconstruction model we proposed is divided into two steps for training, pre-training encoder (Fig. 2b) and training Loop-Enc-Decoder guided by F-attention (Fig. 2c).

Pre-train fMRI encoder. The fMRI encoder realizes the encoding process from natural images to fMRI. The structure of the encoder is mainly composed of residual network, convolution and full connection. The front end of the encoder is used as ResNet50 network (He et al., 2016) pre-trained on the ImageNet dataset, this part can extract image features (Fig. 4). The loss function is composed of the weighted sum of the mean square error (MSE) and the cosine similarity, as shown below:

$$Loss_{fMRI} \propto \alpha \sum (x_i - x_i^*)^2 + \beta \sum cos\left(\angle (x_i, x_i^*)\right) \tag{4}$$

Where α and β are hyperparameters, x_i is the original fMRI label, and x_i^* is the encoding result. In this paper, α is assigned a value of 0.85 and β is -0.15. This loss function is also used in the training process of Loop-Enc-Dec.

Train Loop-Enc-Decoder guided by F-attention. We fix the weight of pre-trained fMRI encoder, then connect an image decoder to form an end-to-end enc-dec model. The so-called 'Loop' idea is to re-input the reconstructed images into the encoder for cyclic encoding in the process of training Enc-Dec, increasing the fMRI encoding loss constraint, so that the original end-to-end model is transformed into a loop structure. The image decoder is composed of full connection, convolution and upsampling. At the same time, F-attention is used as a weight to multiply the convolution feature to guide the training process of the model (Fig. 4).

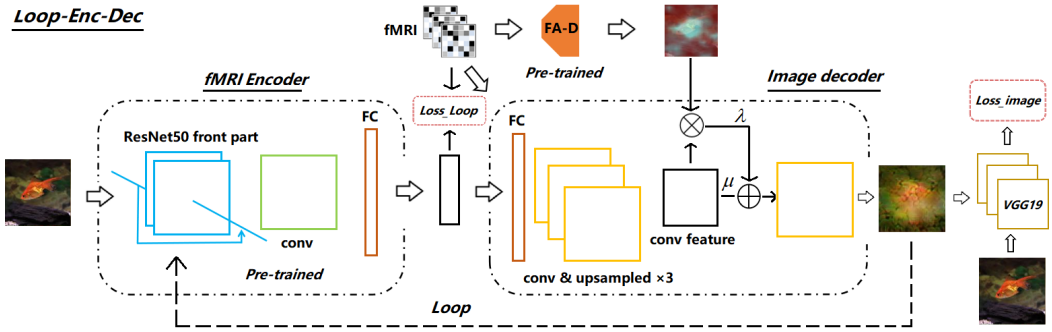

Figure 4: Detailed structure of Loop-Enc-Dec. We fix the network parameters of the pre-trained encoder, and train the Image decoder under the guidance of F-attention to realize the work of visual stimulus image reconstruction.

The F-attention (FA) decoded by pre-training F-attention decoder and the convolution features ($ConvF$) are used as a Hadamard product, the output is weighted λ and added to the original input

convolution feature weighted μ. Then we complete the guidance of the attention weight, as shown in the following formula:

$$Feature_{F-attention} = \lambda FA \circ ConvF + \mu ConvF \tag{5}$$

In the training process, λ is assigned a value of 0.7 and μ is 0.3. We also add unlabeled ImageNet images for end-to-end unsupervised self-enc-dec to enhance the generalization ability of the model. The loss function in the training process consists of three parts, as follows:

$$Loss_{Loop-Enc-Dec} = Loss_{image} + Loss_{Loop} + Loss_{imageNet} \tag{6}$$

Details of image decoder training strategy and the loss function. During training, the image decoder has two inputs, one is the original fMRI data, and the other is the output of pre-trained fMRI encoder. These two inputs are decoded by the image decoder to generate two images, which we named $img_{fMRI}, img_{feature}$, then we loop $img_{fMRI}, img_{feature}$ back into the fMRI encoder to encode two different results, which we names $fea_{imgfMRI}, fea_{imgfea}$. And we named the image obtained by self-enc-dec of ImageNet data $img_{imageNet}$. The three parts of the loss function are defined as follows:

$$Loss_{image} \propto \sum [V(img_{fMRI}) - V(img)]^2 + \sum [V(img_{feature}) - V(img)]^2 \tag{7}$$

$$Loss_{Loop} = Loss_{fMRI}(fea_{imgfMRI}, x) + Loss_{fMRI}(fea_{imgfea}, x) \tag{8}$$

$$Loss_{imageNet} \propto \sum [V(img_{ImageNet}) - V(imgNet)]^2 \tag{9}$$

Where V represents the feature extractor composed of pre-trained $VGG19$ (Simonyan & Zisserman, 2014), img is the original stimulus image, and $imgNet$ is the imageNet image without fMRI tags.

2.4 Evaluation method

This paper uses a quantitative method to evaluate the reconstructed visual stimulus images. The evaluation method is called pairwise similarity comparison (Shen et al., 2019b). There are a total of 50 natural stimulus images in the test set. For each reconstructed image, it is calculated SSIM (Wang et al., 2004) value separately from the 50 natural stimulus images. We define the SSIM between the reconstructed image and the corresponding original image as $SSIM_{pair}$, and the SSIM with the remaining 49 as $SSIM_{random}$, then $SSIM_{pair}$ and $SSIM_{random}$ compare the SSIM values in pairs, a total of 50×49=2450 comparisons, we count the correct of times ($SSIM_{pair}$ is greater than $SSIM_{random}$), and calculate the correct rate(Shen et al., 2019b).

3 Results

In the following, we will present the F-attention decoding results, visual stimulus image reconstruction results, some ablation experiments involving the model, and describe some details of the training process, finally verify some conclusions in neuroscience. All our experiments are performed on the data set Horikawa2017 (Horikawa & Kamitani, 2017). There are a total of 1200 images in the training set and a total of 50 images in the test set. There is no overlap in image categories between the training set and the test set.

3.1 F-attention decodeing results

We use entire visual cortex (VC) of fMRI data labeled with visual attention distribution for training. A total of 280 rounds of experiments are performed. The learning rate is set by using the SGDR method (Loshchilov & Hutter, 2016). According to SGDR, we divide the 280 rounds of training into four small rounds. The learning rate of each small round rises to the maximum value of 0.03 through a linear function, and then decays to the minimum value of 4e-5 through cosine annealing. Each small round repeats this operation to realize the periodic restart of the learning rate and improve the convergence speed and generalization of the model, the details are shown in the appendix. The results of decoding the F-attention distribution from fMRI are shown in Fig. 5, the more blue the

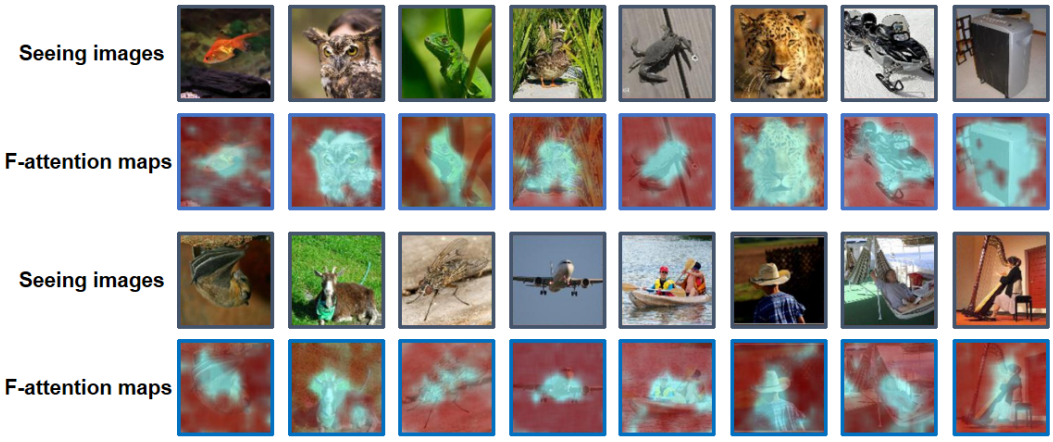

Figure 5: F-attention map decoding results of test data set. The more blue the area, the higher the weight of visual attention. The distribution of F-attention in the images can be accurately found.

area, the higher the weight of visual attention. From the results, it can be seen that the F-attention decoder we design can decode the distribution of visual attention successfully, whether the object is an animal, a man-made object, or a human, such as the fourth image in the second row, the duck in the grass can be accurately found, the seventh image in the fourth row, visual attention distribution area of the person lying on the hammock can also be accurately found.

3.2 Visual stimulus images reconstruction results

In this part, we present the reconstruction results of our method and compare the results with other existing works (Shen et al., 2019a; Beliy et al., 2019; Fang et al., 2020), the comparison results are shown in Fig. 6. Through the evaluation method of pairwise similarity comparison mentioned above, we quantitatively evaluate the results of the experiment. It can be seen from the experimental results that our method can effectively reconstruct the images seen by humans from fMRI, and better decode the shape and direction of the prominent object. Compared with the previous work (reconstruction results of the visual stimulus images shown in Fang et al. (2020)), some improvements have been obtained. From a subjective perspective, our results can better pay attention to the distribution of foreground objects in the images. For example, for the reconstruction of the first goldfish image, the position and color of the goldfish in the image are more accurate. From a quantitative comparison point of view, our method experiment results achieve 68.9% by pairwise SSIM comparison, slightly better than 68.4% (Fang et al., 2020), and better than 66.8% (Fang et al., 2020) and 63.8% (Shen et al., 2019a).

3.3 Model ablation experiments

Here, we conduct ablation experiments to discuss the role of F-attention and Loop-Enc-Dec in reconstruction of visual stimulus images, the results are shown in Fig. 7. First, we give the results obtained using the complete method of this paper (the second row of Fig. 7), the pairwise SSIM value is 68.9%. Then on the basis of the complete method, the F-attention guidance is removed from the training of the decoder, and only the Loop-Enc-Dec structure is used for step-by-step training. First, the fMRI encoder is pre-trained, then the Loop-Enc-Dec is trained. We finally use the image decoder to test and get the reconstruction results (the third row of Fig. 7), the pairwise SSIM value is 66.7%. Then we remove the $Loss_{Loop}$ on the basis of the complete method, just use the image loss of end-to-end encoding, and at the same time train the image decoder under the guidance of F-attention and obtain the reconstruction results (the fourth row of Fig. 7), The value of pairwise SSIM is 65.0%. Finally, on the basis of the complete method, we remove the F-attention guidance and the $Loss_{Loop}$, and only use the end-to-end enc-dec for step-by-step training. The image reconstruction results are shown in the fifth row of Fig. 7, the pairwise SSIM value is 62.5%.

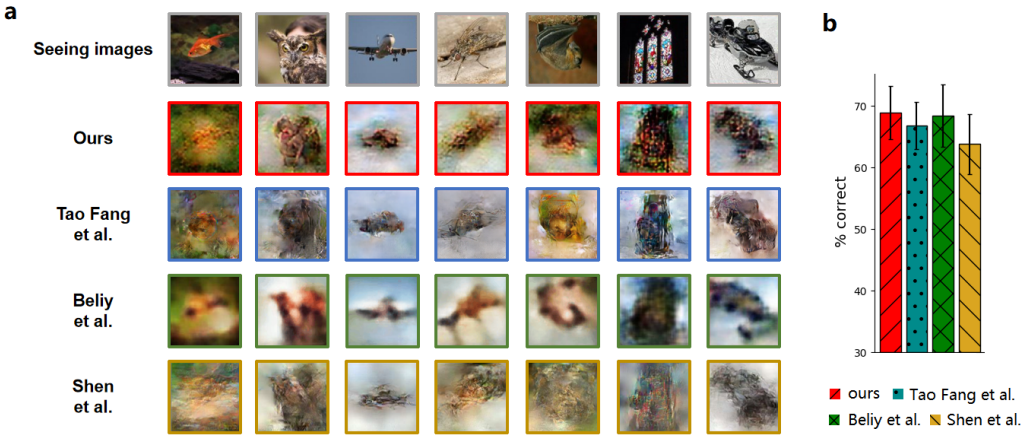

Figure 6: Reconstruction results of visual stimulus images compared with other methods. (a) Visual stimulus images reconstructed by different methods. (b) Result comparison with pairwise similarity.

Comparing the ablation experimental results, it is not difficult to see that our proposed F-attention and $Loss_{Loop}$ have played a positive role in the image reconstruction work. Comparing the results in the second row with the results in the fifth row, it can be seen that under the combined effect of F-attention and $Loss_{Loop}$, the image reconstruction quality has been greatly improved, including the shapes and colors of the prominent objects in the images, 68.9% is better than 62.5% in pairwise SSIM quantification.

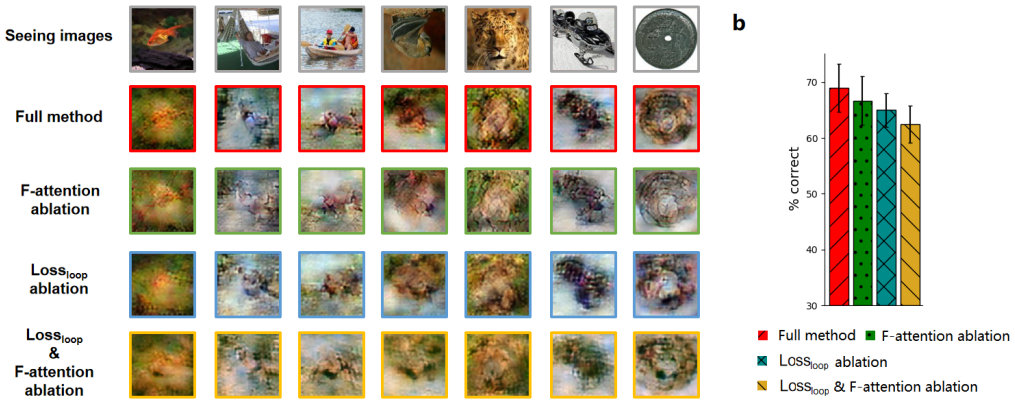

Figure 7: Model ablation experiments. (a) Visual stimulus images reconstructed by full method, full method without F-attention, full method without $Loss_{loop}$, full method without F-attention and $Loss_{Loop}$. (b) Comparing ablation results by pairwise SSIM.

Comparing the ablation experimental results separately. Comparing the results of the second row and the third row, it can be seen that F-attention plays a certain guiding role in the model, making the reconstruction of foreground objects in the images more prominent, such as the goldfish in the water in the first column, and the people lying in a hammock in the second column, pairwise SSIM quantitative comparison shows that 68.9% is better than 66.7%. The guiding effect of F-attention can also be obtained by comparing the results of fourth rows and fifth rows. The reconstruction results of foreground objects of fourth row are more prominent, and pairwise SSIM quantitative comparison shows that 65.0% is better than 62.5%. Comparing the results of the second row and the fourth row, it can be seen that the $Loss_{Loop}$ has made the details of the reconstruction images more clear and accurate. For example, the face of the leopard in the fifth column is clearer, and the shape and details of stone pan are more accurate in the seventh column. The pairwise SSIM quantitative

comparison shows that 68.9% is better than 65.0%. The constraint effect of $Loss_{Loop}$ can also be obtained by comparing the results of the third row and the fifth row. The details and quality of the image reconstruction in the third row are better. The pairwise SSIM quantitative comparison shows that 66.7% is better than 62.5%. According to the above ablation experiments, we proved the guiding role of F-attention in image reconstruction and the constraint role of $Loss_{Loop}$ in the training process.

3.4 Validation of neuroscience research

We verify some results of neuroscience research through experiments in the following. It is generally believed that in the human visual cortex, the low-level visual (LVC) cortex often pays attention to the shape and contour information of objects, and has significant directional sensitivity, while the high-level visual cortex (HVC) may correspond to the category information, movement information and other higher-level information of objects (Carandini et al., 2005; DiCarlo et al., 2012; Bracci et al., 2017; Zeman et al., 2020). We reconstruct the visual stimulus images with the fMRI data of the low-level visual cortex and the high-level visual cortex respectively. The results are shown in the Fig. 8, the pairwise SSIM value of the images reconstructed from LVC is 64.7%, and HVC is 55.7%. By comparing the experimental results, we can find that the LVC reconstruction results largely restore the shape and direction information of the foreground objects in the images, while the HVC reconstruction results lose the shape and direction information of the foreground objects. Quantitatively, the pairwise SSIM value of LVC is also better than HVC. This results verify that low-level visual cortex (LVC) pays more attention to the shape and contour information of objects than high-level visual cortex (HVC).

Figure 8: Validation of neuroscience research. (a) Visual stimulus images reconstructed from different visual cortex areas, including VC, HVC and LVC. (b) Result comparison with pairwise SSIM.

4 Conclusion

In this paper, inspired by human visual attention, we define the Foreground-attention distribution of people when they view natural images, and successfully decode the visual attention distribution (Foreground-attention) from the fMRI data of human brain. We achieve the decoding and interpretation of the visual attention of the human brain. The results show that areas with higher attention are concentrated on prominent objects in the images, such as animals and animal faces in the images, while lower attention is given to messy background information. This result is in line with the research conclusions on human brain visual attention (Ungerleider & G, 2000; Braun et al., 2001; Poggio & Anselmi, 2016; Khosla et al., 2020). At the same time, we propose a new enc-dec training strategy called Loop-Enc-Dec, and apply the decoded F-attention to guide it, then reconstruct the visual stimulus images from the fMRI data successfully. Compared with the previous work, the reconstructed visual images in this paper are more realistic and get a higher score on pairwise SSIM. From the perspective of neuroscience, we verify the following conclusions through experiments: The low-level visual cortex pays more attention to the shape of objects in the scene and has significant directionality, while the high-level visual cortex is insensitive to shape and direction information.

REFERENCES

Roman Beliy, Guy Gaziv, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. From voxels to pixels and back: Self-supervision in natural-image reconstruction from fmri. *Advances in Neural Information Processing Systems*, 32:6517–6527, 2019.

Stefania Bracci, J Brendan Ritchie, and Hans Op de Beeck. On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia*, 105:153–164, 2017.

Jochen Braun, Christof Koch, Joel L Davis, et al. *Visual attention and cortical circuits*. MIT Press, 2001.

Matteo Carandini, Jonathan B Demb, Valerio Mante, David J Tolhurst, Yang Dan, Bruno A Olshausen, Jack L Gallant, and Nicole C Rust. Do we know what the early visual system does? *Journal of Neuroscience*, 25(46):10577–10597, 2005.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

Tao Fang, Yu Qi, and Gang Pan. Reconstructing perceptive images from brain activity by shape-semantic gan. *Advances in Neural Information Processing Systems*, 33, 2020.

Yusuke Fujiwara, Yoichi Miyawaki, and Yukiyasu Kamitani. Modular encoding and decoding models derived from bayesian canonical correlation analysis. *Neural computation*, 25(4):979–1005, 2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):1–15, 2017.

Meenakshi Khosla, Gia Ngo, Keith Jamison, Amy Kuceyeski, and Mert Sabuncu. Neural encoding with visual attention. *Advances in Neural Information Processing Systems*, 33, 2020.

Yunfeng Lin, Jiangbei Li, and Hanjing Wang. Dcnn-gan: Reconstructing realistic image from fmri. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pp. 1–6. IEEE, 2019.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–929, 2008.

Tomaso Poggio and Fabio Anselmi. *Visual cortex and deep networks: learning invariant representations*. MIT Press, 2016.

Russell A Poldrack and Martha J Farah. Progress and challenges in probing the human brain. *Nature*, 526(7573):371–379, 2015.

Kai Qiao, Jian Chen, Linyuan Wang, Chi Zhang, Li Tong, and Bin Yan. Biggan-based bayesian reconstruction of natural images from human brain activity. *Neuroscience*, 444:92–105, 2020.

Katja Seeliger, Umut Güçlü, Luca Ambrogioni, Yagmur Güçlütürk, and Marcel AJ van Gerven. Generative adversarial networks for reconstructing natural images from brain activity. *NeuroImage*, 181:775–785, 2018.

Guohua Shen, Kshitij Dwivedi, Kei Majima, Tomoyasu Horikawa, and Yukiyasu Kamitani. End-to-end deep image reconstruction from human brain activity. *Frontiers in Computational Neuroscience*, 13:21, 2019a.

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLoS computational biology*, 15(1):e1006633, 2019b.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Tuomas Tallinen, Jun Young Chung, John S Biggins, and L Mahadevan. Gyrification from constrained cortical expansion. *Proceedings of the National Academy of Sciences*, 111(35):12667–12672, 2014.

Sabine Kastner Ungerleider and Leslie G. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, 23(1):315–341, 2000.

Marcel AJ Van Gerven, Floris P De Lange, and Tom Heskes. Neural decoding with hierarchical generative models. *Neural computation*, 22(12):3127–3142, 2010.

Rufin VanRullen and Leila Reddy. Reconstructing faces from fmri patterns using deep generative neural networks. *Communications biology*, 2(1):1–10, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Siyu Yu and Nanning Zheng. A correlation network framework for revealing brain connections in brain decoding. *IEEE Workshop on Nonlinear Circuit Networks*, 2017.

Siyu Yu, Nanning Zheng, Yongqiang Ma, Hao Wu, and Badong Chen. A novel brain decoding method: A correlation network framework for revealing brain connections. *IEEE Transactions on Cognitive and Developmental Systems*, 11(1):95–106, 2018.

Astrid A Zeman, J Brendan Ritchie, Stefania Bracci, and Hans Op de Beeck. Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. *Scientific reports*, 10(1):1–12, 2020.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pp. 7354–7363. PMLR, 2019.

Yichen Zhang, Shanshan Jia, Yajing Zheng, Zhaofei Yu, Yonghong Tian, Siwei Ma, Tiejun Huang, and Jian K Liu. Reconstruction of natural visual scenes from neural spikes with deep neural networks. *Neural Networks*, 125:19–30, 2020.

A APPENDIX

A.1 DETAILS OF MODELS AND TRAINING

In this paper, the details of the F-attention decoder we designed are as follows, the input fMRI data is a one-dimensional vector, and the fully connected layer is connected, the output size is 512*10*10, then reshape it into the feature of size (10, 10, 512). The convolutions has step size of 2 and a kernel size of 4. The number of self-attention channels is 64. In Loop-Enc-Dec, the front-end ResNet50 feature extractor in the fMRI decoder selects 'conv3-block4-out' as the feature output, the output feature size is 14*14*512, followed by two layers of residual networks, one layer of max-pooling and convolution (the size of the convolution kernel is 3, the step size is 1, padding is 'same'), and finally outputs the encoded results through a layer of fully connection. In order to eliminate the checkerboard effect that appears in the reconstructed image, in image decoder, we set the size of the convolution kernel to be twice the convolution step length, and use linear interpolation for each upsampling to expand the feature to four times the original.

A.2 CHANGES IN LEARNING RATE USING SGDR

The linear function and cosine annealing function used in each round are as follows:

$$lr_{SGDR} = \begin{cases} \frac{(E_{cur}+1)*l_{max}}{E_w+1}, E_{cur} < E_w \\ l_{min} + \frac{1}{2}(l_{max} - l_{min}) * (1 + cos(\frac{E_{cur}-E_w}{E_{per}-E_w})), E_{cur} \geq E_{warm} \end{cases} \tag{10}$$

E_{cur} counts how many epochs have been performed since the last restart, E_w represents the 'warm' epochs of each small round , E_{per} represents the total epochs of each small round, l_{max} represents the maximum learning rate, and l_{min} represents the termination learning rate.

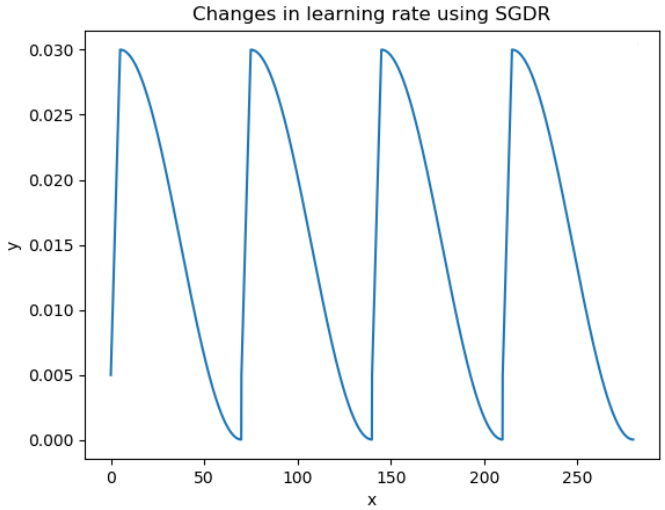

Figure 9: Changes in learning rate using SGDR

A total of 280 rounds of experiments were performed. The learning rate is set using the SGDR method (Loshchilov & Hutter, 2016). According to SGDR, we divide the 280 rounds of training into four small rounds. The learning rate of each small round rises to the maximum value of 0.03 through a linear function, and then decays to the minimum value of 4e-5 through cosine annealing. Each small round repeats this operation to realize the periodic restart of the learning rate and improve the convergence speed and generalization of the model.