Semantics-Guided Hierarchical Feature Encoding Generative Adversarial Network for Visual Image Reconstruction From Brain Activity

Lu Meng and Chuanhao Yang

***Abstract*— The utilization of deep learning techniques for decoding visual perception images from brain activity recorded by functional magnetic resonance imaging (fMRI) has garnered considerable attention in recent research. However, reconstructed images from previous studies still suffer from low quality or unreliability. Moreover, the complexity inherent to fMRI data, characterized by high dimensionality and low signal-to-noise ratio, poses significant challenges in extracting meaningful visual information for perceptual reconstruction. In this regard, we proposes a novel neural decoding model, named the hierarchical semantic generative adversarial network (HS-GAN), inspired by the hierarchical encoding of the visual cortex and the homology theory of convolutional neural networks (CNNs), which is capable of reconstructing perceptual images from fMRI data by leveraging the hierarchical and semantic representations. The experimental results demonstrate that HS-GAN achieved the best performance on Horikawa2017 dataset (histogram similarity: 0.447, SSIM-Acc: 78.9%, Peceptual-Acc: 95.38%, AlexNet(2): 96.24% and AlexNet(5): 94.82%) over existing advanced methods, indicating improved naturalness and fidelity of the reconstructed image. The versatility of the HS-GAN was also highlighted, as it demonstrated promising generalization capabilities in reconstructing handwritten digits, achieving the highest SSIM (0.783±0.038), thus extending its application beyond training solely on natural images.**

***Index Terms*— Visual decoding, image reconstruction, generative adversarial network, fMRI.**

Manuscript received 1 October 2023; revised 28 December 2023 and 6 March 2024; accepted 7 March 2024. Date of publication 18 March 2024; date of current version 25 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62073061, in part by the Fundamental Research Funds for the Central Universities under Grant N2204009, in part by the Non-Profit Central Research Institute Fund of Chinese Academy of Medical Sciences under Grant 2022-JKCS-21, in part by Chongqing Science and Health Joint Medical Research Project under Grant 2023MSXM137, and in part by Liaoning Provincial Natural Science Foundation Joint Fund for Medical-Industrial Crossover under Grant 2022-YGJC-31. *(Corresponding author: Lu Meng.)*

The authors are with the College of Information Science and Engineering, Northeastern University, Shenyang 110819, China (e-mail: menglu1982@gmail.com; 2100877@stu.neu.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2024.3377698

I. Introduction

THE human visual system serves as a crucial sensory organ for acquiring external information [1], making the decoding of brain vision a compelling topic in the field of neuroscience. Functional magnetic resonance imaging (fMRI) [2] is an effective non-invasive method for recording brain activities, and its popularity in visual decoding studies is steadily increasing. Visual stimulus decoding encompasses three distinct tasks: image classification, stimuli recognition, and perceived reconstruction [3]. Among these tasks, reconstructing perceived images is the most challenging, as it requires efficient utilization of the limited information available in fMRI data.

Previous studies have demonstrated the existence of a mapping between cerebral cortical activity and stimulus images [4], enabling the decoding of perceptual images from fMRI data [5], [6], [7], [8]. Several approaches have been explored for perceptual image reconstruction, including machine learning methods, convolutional neural network (CNN) methods, and generative deep learning methods. Machine learning methods employ linear models to map fMRI voxels to handcrafted features (local image structure, Gabor filter features) for visual reconstruction [6]. However, these linear mapping-based approaches are primarily suited for simple stimulus images, such as domino patterns [7], handwritten numbers [8], and English letters [9], and may fall short in reconstructing complex natural images. Notably, researchers have discovered a strong correlation between CNN features and brain activity in the visual cortex [10], leading to the adoption of CNNs for recovering natural images from fMRI. These methods involve linearly mapping fMRI voxels to CNN features and then converting the corresponding features back into images through a decoder. For instance, Wen et al. linearly mapped fMRI signals to specific CNN layer features and utilized a decoding network for video frame reconstruction [11]. Bely et al. [12] devised an encoder-decoder framework based on CNNs to address the scarcity of fMRI data, where the encoder maps stimulus images to fMRI voxel space, and the decoder performs the reverse mapping. The combination of encoder and decoder enables the use of self-supervision for training. Kai et al. proposed a reconstruction model based on visual attention guidance, inspired by the mechanism of human visual attention. By decoding visual attention distribution from fMRI signals, and then reconstructing perceptual images under its guidance [13].

With the development of image generation models, many researches have begun to utilize deep generative models to reconstruct stimulus images, such as Variational Autoencoder (VAE) [14], Generative Adversarial Network (GAN) [15], and Latent Diffusion Model (LDM) [16]. These methods typically pre-train a deep generative model on large-scale datasets and then use linear regression or neural networks to learn the mapping of fMRI signals to latent feature vectors of the generative model. In this way, during the inference stage, the corresponding stimulus image can be reconstructed based on the latent feature vectors predicted by fMRI. For example, Ozcelik et al. [17] used ridge regression to decode latent variables from fMRI patterns for pre-training Instance-GAN to generate images with similar semantics to visual stimuli. With the help of deep generation network, images of different complexity can be reconstructed, such as faces [18], [19], single object-centered images [20], [21] and complex scene images [22], [23]. In particular, since the publication of the latent diffusion model, many visual reconstruction methods based on it have emerged [24], [25], [26], [27], [28], [29], which can reconstruct high-quality complex scene images by utilizing the powerful generative capabilities of the latent diffusion model. Although these methods based on deep generative models have achieved impressive naturalness of reconstructed images, they have several inherent problems: (1) The application of pre-trained generative model is favorable to enhance the reconstruction quality, but the generated image is generally inconsistent with the original image semantics. (2) There is no guarantee that the generated image contains low-level features of the visual stimulus, i.e., the reconstruction commonly fails to match the real image. (3) Even with random noise as input, these models can generate high-quality images, resulting in unreliable decoding.) However, for visual reconstruction task, more emphasis should be placed on consistency with the original image compared to the diversity of the generated image. Therefore, the perfect reconstruction of visual stimulation remains to be explored.

To address the problems of the above methods and make the reconstructed image as consistent as possible with the original image, it is necessary to consider how to send more low-level visual features into the reconstruction space, and how to adequately utilize the limited information in the fMRI signals to guide the generator to restore the complex colors and textures of the natural image. The works of Horikawa and Kamitani [30] identified homology between the visual cortex and deep neural networks (DNNs) in hierarchical representation. This discovery established that DNN features can serve as proxies for the hierarchical representation of human vision, which can be translated from fMRI signals. Fang et al. [31] further emphasized that lower visual cortex (LVC) exhibits a higher correlation with low-level image features, while higher visual cortexes (HVC) display stronger correlations with image semantic features. Incorporating information from different visual cortex areas has proven beneficial in enhancing visual decoding performance. However, previous studies merely employed layer-specific DNN features and disregarded the relationship between visual features at various levels of the stimulus image and the visual cortex. Consequently, this limitation resulted in insufficient visual decoding and hindered the model's generalization ability.

Building on these insights, we introduce a novel decoding framework called the hierarchical semantic generative adversarial network (HS-GAN) to reconstruct corresponding perceptual images from fMRI recordings. Drawing inspiration from the hierarchical encoding of the visual cortex, our approach involves constructing an image encoding network that extracts different levels of visual features (hierarchical encoder) from stimulus images and supplements semantic features (semantic encoder), which are then compressed into low-dimensional latent vectors. To preserve more fine-grained details during visual reconstruction, we devise a generative network with skip connections to restore the corresponding visual stimuli from these latent representations. Additionally, we integrate self-attention modules into the generator, enabling the model to effectively leverage important visual information contained in the latent vectors at different levels. To account for the potential nonlinearity of fMRI data, we design a neural decoder with residual connectivity, which efficiently learns the mapping of fMRI to DNN features without overfitting. Given the limited number of fMRI-image pairing samples, we divide the model training into two stages. Initially, the model is trained on an additional large natural dataset in the first stage to incorporate prior knowledge, thereby enhancing reconstruction quality. Subsequently, in the second stage, we solely train the neural decoder to learn the transformation from fMRI voxels to perceptual image visual and semantic features. During the inference stage, the neural decoder is employed to predict corresponding latent representations of perceptual images from test fMRI patterns, which are then fed to the generator to obtain the final reconstructed images. Our primary contributions can be summarized as follows:

- We propose a hierarchical semantic-guided visual reconstruction framework, which successfully decodes hierarchical visual and semantic representations of stimulus images from fMRI patterns. This approach maximizes the utilization of limited visual information in fMRI data, leading to improved reconstruction quality.
- The design of our generator, incorporating skip connections and attention modules, facilitates the recovery of perceptual images from low-dimensional representation vectors, further enhancing the fidelity of the reconstructed images.
- We introduce a neural decoder with residual connectivity, effectively learning the mapping of fMRI to DNN features and bolstering the accuracy of fMRI decoding. Additionally, we introduce a reconstruction loss in the training process of the neural decoder.
- Through extensive validation on two distinct datasets, our model achieves state-of-the-art performance, confirming the efficacy of the proposed approach

II. Related Work

A. Visual Reconstruction From fMRI

The existing approaches to visual reconstruction can be roughly divided into two groups. The first one emphasizes that the reconstructed images are similar to the original images in pixel space. Since the reconstructed image is expected to be consistent with the original image, this type of approach focuses on network design and training strategies, and train their own generative model from scratch. Shen et al. [20] designed an end-to-end DNN generative model to directly learn the mapping from fMRI voxels to images. Moreover, the discriminator and comparator were employed during the generator training process to introduce adversarial loss and perceptual loss. In the same year, Shen et al. [21] employed a linear decoder to decode fMRI into DNN features, and then optimized the pixel values of the image using feature loss to minimize the difference between its DNN features and the DNN features decoded from the fMRI pattern. The method of Beliy et al. [12] consists of an encoder E and a decoder D, where E converts the image to the corresponding fMRI voxels while D maps the fMRI to its corresponding image space. Two combined networks *E*-*D* and *D*-*E* were constructed by stacking E and D back-to-back for unsupervised training on unpaired images and fMRI data. Fang et al. [31] used linear models and shallow DNNs to decode shape features and category features of stimulus images from fMRI, respectively, which were then used as conditional information to train a GAN generator. Kai et al. [13] begin by predicting salient regions in the image (foreground attention) from the fMRI pattern, and then used it as a guide for the image decoder to recover the visual stimulus from the fMRI. A similar training strategy was used during training as in Beliy et al.

The second one focuses on the similarity of the reconstructed image to the original image in high-level semantic features. Such approaches typically synthesize the reconstructed content with the help of pre-trained generative models (e.g., instance-gan), guided by fMRI patterns. Ozcelik et al. [17] utilize ridge regression to decode conditional instance variables of an instance-GAN from fMRI patterns, which are then used as conditional guidance for pre-trained GAN to generate images with similar semantics to the visual stimuli. Chen et al. [25] first pre-trained a Mask Auto-Encoder (MAE) on an additional fMRI dataset, which was used to extract valid representations of fMRI voxels. Subsequently, the MAE-extracted features are utilized as the textual condition to fine-tune the pre-trained LDM to recover stimulus images. However, these methods can't guarantee that the reconstruction is semantically consistent with fMRI and thus lacks reliability. This is not suitable for some applications in real world scenarios, such as patient diagnosis.

B. Visual Information Processing

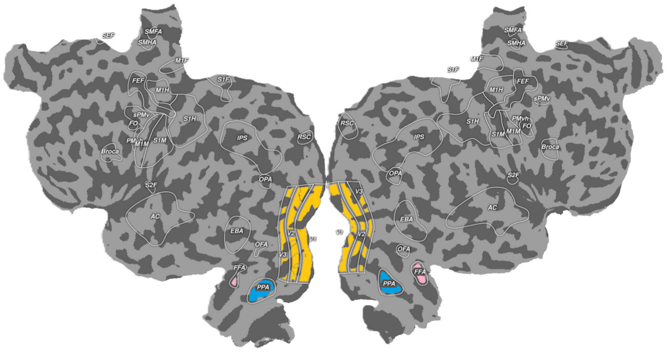

Fig. 1. The cortical surface map of the brain.

The processing of visual information can be divided into three different levels. Low-level processing, including the retina, lateral geniculate nucleus (LGN), and primary visual cortex (V1). This is the first step in visual processing, which focuses on perceiving the orientation, lines and edges of an image. Afterwards, mid-level processing, involving visual regions V2, V3, and V4. They extract shape, object and color features in the image, respectively. Finally, there is high-level processing. This step is accomplished by high-level visual areas such as fusiform face areas (FFA), lateral occipital (LOC), parahippocampal area (PPA), and medial temporal area (MT/V5). They show selective responses to face, object, place and movement. Based on the above conclusions, we should consider the relationship between different levels of visual regions and image features during fMRI decoding.

III. Method

A. Overview

Let (x, y) represent the {Image, fMRI} data pair, where $x \in \mathbf{R}^{H\times W\times C}$ represents the natural image, the H, W and C are the height, width and number of channels of x; $y \in \mathbf{R}^{L}$ represents the fMRI sample collected when the subject viewed image x and L denotes the dimension. Fig. 2a show that the reconstruction task is to recover the perceived images from fMRI recordings. The visual image reconstruction framework we proposed includes three key parts: image feature encoder, neural decoder, and GAN image generator (Fig. 2b). The image feature encoder E_{Φ} includes hierarchical encoder E_{η} and semantic encoder E_{ϵ}. For simplicity, we use $z_h = \{z_{h1}, z_{h2}, z_{h3}, z_{h4}\}$ to represent hierarchical latent vectors, where $z_h = E_{\eta}(x)$. In order to introduce category information into the reconstructed image, we use semantic encoder E_{ϵ} to obtain the semantic feature z_{sm} of original image to assist the generator G_{θ} in reconstructing the semantically meaningful image. Hierarchical latent vectors and semantic representations are concatenated and fed into the generator to reconstruct the perceived image. Let $\hat{x} = G_{\theta}(z_h, z_{sm})$ denotes the recovered image of x. Since training the generative model requires a large amount of data, we combine the image feature encoder and the generator to form an autoencoder structure, which allows for self-supervised learning using additional images. At the same time, we also introduce discriminator D_{ϕ} for confrontation training. Subsequently, we use the well-trained image feature encoder to guide the neural decoder D_{ψ} to learn the transformation from fMRI voxels to feature latent vectors, $\left(z_h^*, z_{sm}^*\right) = D_{\psi}(y)$. In this way, we can decode a rich set of image representations from fMRI. Finally, the natural image corresponding to the fMRI sample y is recovered by $x^* = G_{\theta}\left(z_h^*, z_{sm}^*\right)$. Note that the neural decoder predicts

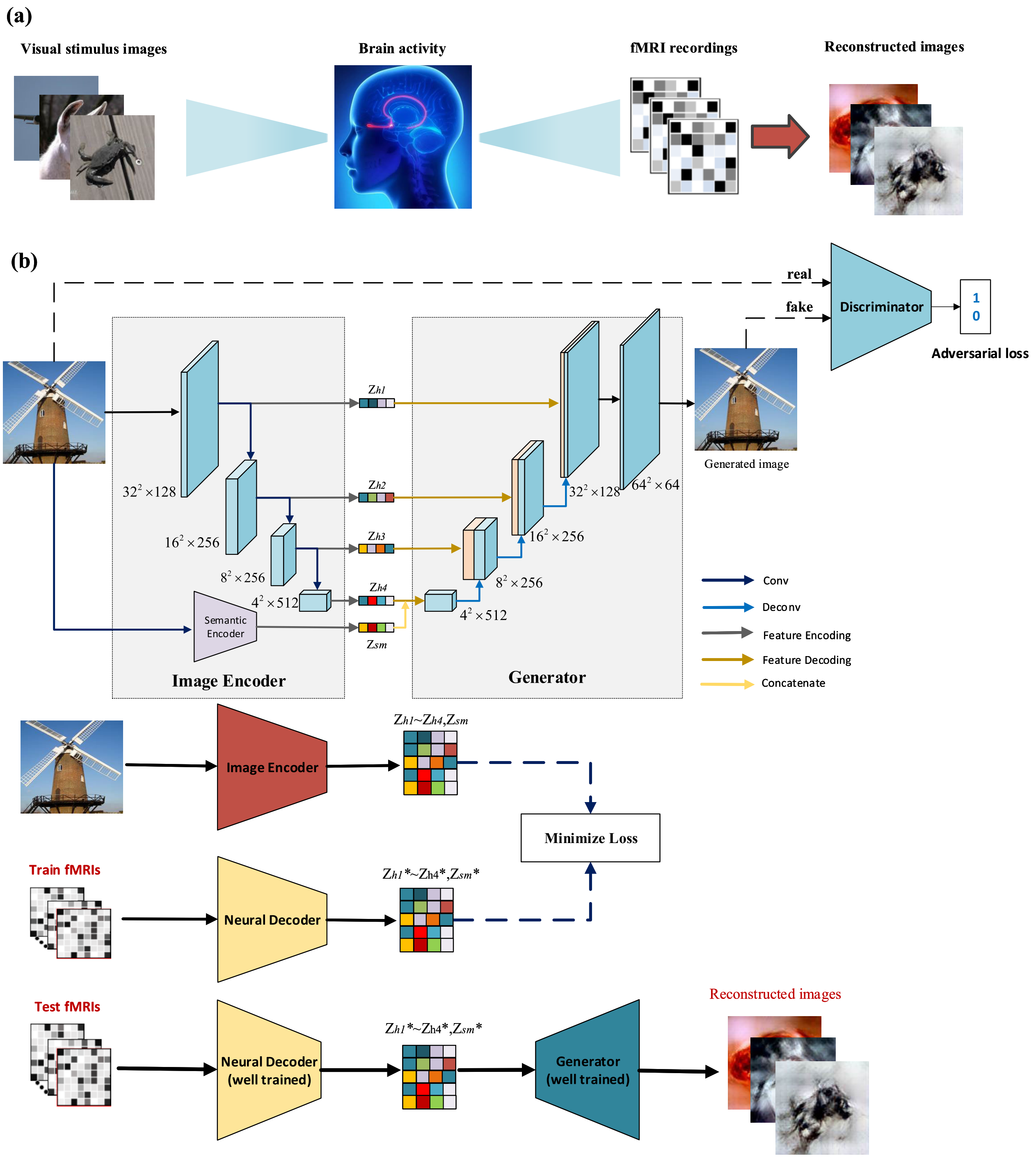

Fig. 2. The visual reconstruction framework proposed in this study. (a) Reconstruct the perceived images from fMRI recordings. (b) Review our overall framework.

hierarchical features using voxels from the entire visual cortex (VC), while for semantic feature it uses voxels from the HVC region, due to the fact that HVC shows more significant response to high-level image features [10]. This design takes into account the relationship between different levels of image features and visual areas.

B. Image Feature Encoder

The image feature encoder plays a crucial role in extracting visual features from images at different levels. It comprises two essential components: the hierarchical encoder and the semantic encoder (Fig. 3). Thus, our feature extraction module effectively preserves both low-level features and high-level semantic content, contributing to superior image reconstruction.

1) Hierarchical Encoder: As the backbone of the hierarchical encoder, we employ a pre-trained resnet-50 [32] deep network from ImageNet. Leveraging the residual connections in this network, we can retain certain low-level features while extracting high-level features from the image. As depicted in Fig. 3a, we utilize the conv1, layer1, layer2, and layer3 modules of the resnet-50 to obtain visual features at various levels of the

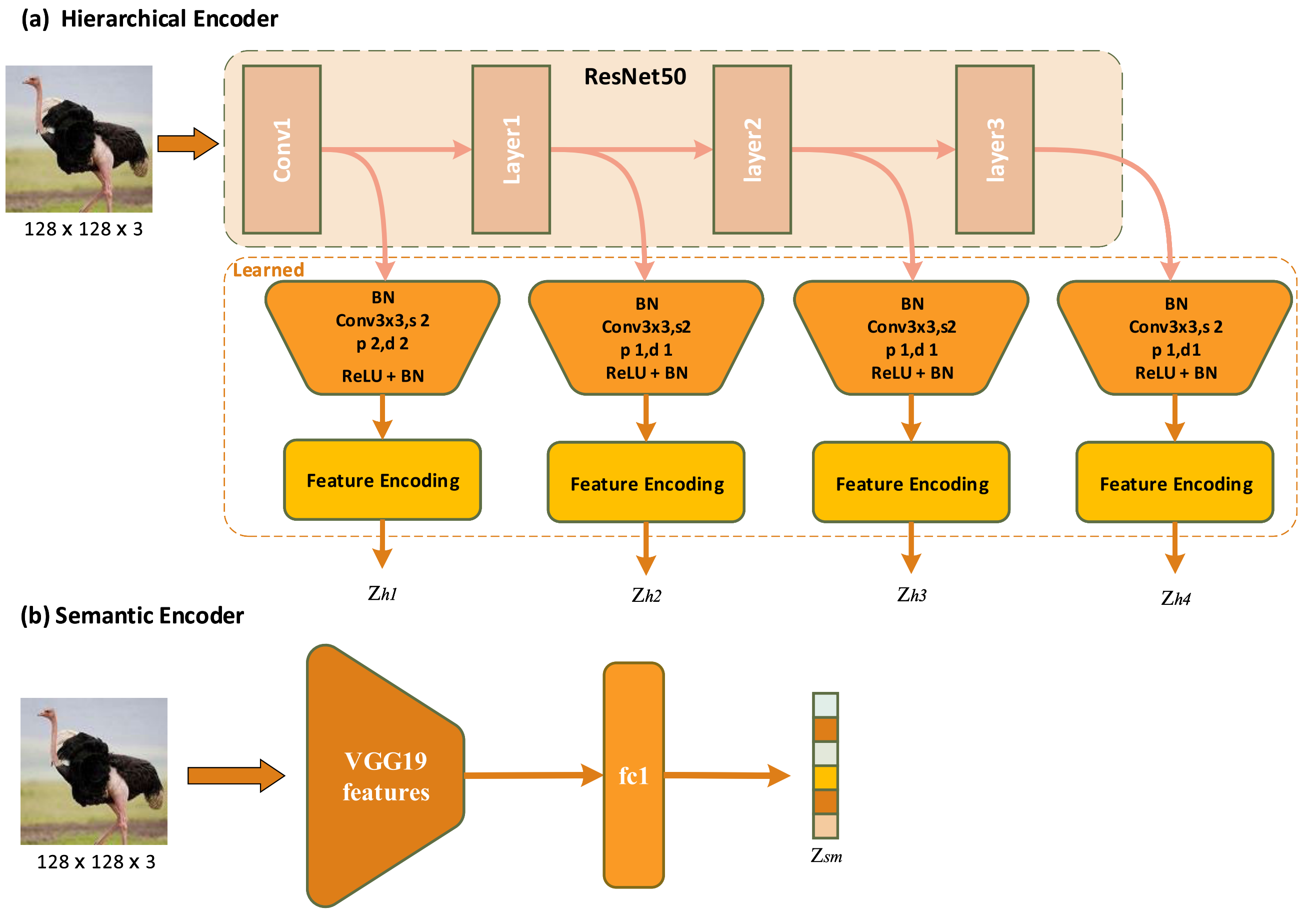

Fig. 3. Image encoder architecture. (a) Hierarchical encoder, where *s* denotes stride, *p* is padding, and *d* is dilation. (b) Semantic encoder.

input image. Since the weights of the resnet-50 network are fixed during training, convolutional modules are introduced to further process the extracted visual features. These features are then compressed into low-dimensional latent vectors using feature encoding blocks. These blocks consist of a convolutional layer and a global pooling layer to reduce the dimensionality of the feature maps, which are ultimately mapped to 1024-dimensional latent vectors through a fully connected (fc) layer.

2) Semantic Encoder: Since the reconstruction quality is positively correlated with the feature decoding accuracy, selecting a DNN with higher decoding accuracy theoretically achieves better reconstruction [21]. Based on this, we use the "brain-like" VGG-19 network [33] as the semantic encoder. Specifically, we use VGG-19 pre-trained on ILSVRC2012 [34] to construct the semantic encoding network, where VGG-19 has 19 convolutional layers and 3 fully connected (fc) layers. For the purpose of decreasing the computing cost and enhancing the decoding precision, the output of the first fc layer of VGG-19 is utilized as the semantic representation. In this way, the dimension of the feature vector is reduced to 4096 and the category information of the original image is preserved (Fig. 3b). The category information of the object assists the generator in reconstructing the underlying details of the stimulus image more accurately.

C. Image Generator

In order to retain more low-level details from the original images in the reconstructed images, we devise a hierarchical semantic GAN inspired by the U-Net [35] design principle (Fig. 2). This skip connection overcomes the limitation of traditional encoder-decoder models that tend to lose low-level features such as shape and texture due to the bottleneck structure during the extraction of high-level features. Consequently, our generator is adept at transferring more low-level details to the reconstruction space. The structure of the generator is shown in Fig. 4. Firstly, a fc layer is used to map the low-dimensional latent vectors into the image feature space, and then it is fed into the transposed convolution module for feature extraction and 2-fold up-sampling, which consists of a transposed convolution layer, a normalization layer and a ReLU activation. Furthermore, in the process of recovering perceived images from hierarchical latent vectors, the image generator must effectively leverage the information embedded within these feature vectors. To address this, we introduced self-attention modules into the image generator architecture, enabling it to emphasize crucial visual information while disregarding less relevant details. Finally, the extracted features are concatenated with the feature maps of the next step in the channel dimension to fuse the different levels of features. It should be noted that we concatenate the semantic vector z_{sm} with the hierarchical latent vector z_{h4} to constrain the generator to preserve the visual features

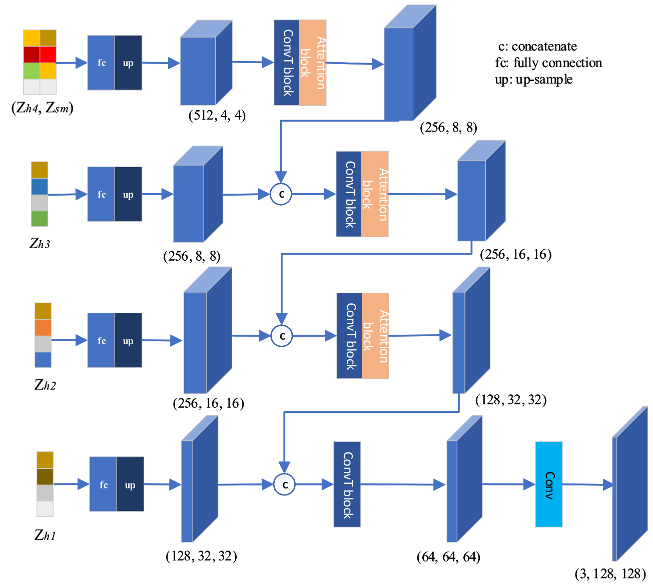

Fig. 4. The structure of image generator, where c denotes concatenation in the channel dimension and ConvT block includes a transpose convolution layer, batch normalization, and ReLu activation.

corresponding to the object category during reconstruction. The self-attention mechanism's calculation formula is as follows:

$$self-attention\left(Q,K,V\right)=softmax\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \quad (1)$$

where $d_k = 1$ and the definitions of Q, K, and V can be found in [36].

Model training. We use the combined loss of image loss, perceptual loss, and adversarial loss during generator training to enhance the recovery image quality. Where image loss is the Mean Square Error (MSE) of the reconstructed image and original image in pixels. Its equation is as follows:

$$\mathcal{L}_{img}=\frac{1}{N}\sum_{i=1}^{N}\left\|x_i-\hat{x}_i\right\|_2^2 \quad (2)$$

where x_i represents the real image, $\hat{x}_i = G_\theta\left(E_\Phi\left(x_i\right)\right)$ denotes the corresponding generated image, and N represents the sample size. We use the Learned Perceptual Image Patch Similarity (LPIPS) proposed in [37] as perceptual loss. It has been proved to achieve better reconstructed image quality [38]. This loss is defined as:

$$\mathcal{L}_{pl}=\frac{1}{N}\sum_{i=1}^{N}\left\|\Psi\left(x_i\right)-\Psi\left(\hat{x}_i\right)\right\|_2^2 \quad (3)$$

where $\Psi(\cdot)$ use the AlexNet [39] network, which is close to the structure of human visual cortex, as a feature extractor for the computation of perceptual loss. The last is adversarial loss, which can provide more natural image reconstruction. The adversarial loss formula is as follows:

$$\mathcal{L}_{adv}=-E\left[\log\left(D_\phi\left(G_\theta\left(z\right)\right)\right)\right] \quad (4)$$

where D_ϕ is the discriminator, which employs the convolutional layers to extract input image's features, and then feeds them into a full-connected layer and the sigmoid function to get the probability of being classified as a real image. $z = \{z_h, z_{sm}\}$ represents latent feature vectors. Finally, reconstruction loss $\mathcal{L}_{gen}$ is:

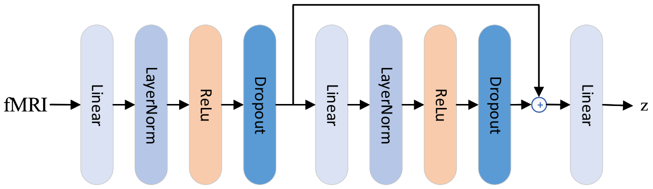

Fig. 5. The structure diagram of neural decoder.

$$\mathcal{L}_{gen}=\mathcal{L}_{img}+\lambda_1\mathcal{L}_{pl}+\lambda_2\mathcal{L}_{adv} \quad (5)$$

where λ_1 and λ_2 are hyperparameters representing the weights of perceptual loss and the adversarial loss, respectively. In order to balance different loss terms, it is necessary to choose appropriate parameters of λ_1 and λ_2. Specifically, we performed a grid search on the interval $[0.001, 10]$ for λ_1 and λ_2, and calculated the LPIPS values [37] of the different parameter models on the validation set. The experimental results indicate that the best reconstruction performance is obtained when $\lambda_1 = 1.0$ and $\lambda_2 = 0.01$, and the reconstructed images are closer to the original images in visual perception (achieving the lowest LPIPS). In addition, the discriminator training loss formula is as follows:

$$\mathcal{L}_{\text{dis}}=-E\left[\log D_\phi(x)\right]-E\left[\log\left(1-D_\phi\left(G_\theta(z)\right)\right)\right] \quad (6)$$

D. Neural Decoder

In this study, we employ the neural decoder to convert fMRI recordings into hierarchical latent vectors and semantic features, subsequently reconstructing the corresponding images through the generator. Existing approaches primarily rely on linear regression to establish the mapping from fMRI to DNN feature maps. However, it has been observed that fMRI signals may introduce nonlinearity when the stimulation duration is less than 4.2 seconds [40]. Additionally, during the image presentation experiment conducted by Horikawa and Kamitani [30], brain activity recordings of presented images were acquired without any rest intervals, introducing a form of nonlinearity in the fMRI data. Furthermore, under the assumption of a linear relationship between visual features and brain activity, simple decoding models are insufficient to model complex visual representations of the brain [17], [29], which leads to inadequate decoding of fMRI. In response to this, we devised a neural network with residual connections to effectively learn the mapping of fMRI to image features. In order to prevent overfitting, we incorporated LayerNorm and Dropout layers into the neural network, as depicted in Fig. 5.

In the training of neural decoder, we use the trained image feature encoder E_Φ to instruct the decoder D_ψ to learn the transformation of fMRI to image feature vectors, fixing the parameters of image generator during this process. We simultaneously minimize two loss functions: feature loss $\mathcal{L}_{feat}$ and reconstruction loss $\mathcal{L}_{gen}$. The feature loss includes MSE and cosine similarity to ensure that the vectors regressed by neural decoder are similar to the original feature vectors in both distance and direction. The feature loss term is defined as:

$$\mathcal{L}_{feat} = \mu\mathcal{L}_{mse} + (1-\mu)\mathcal{L}_{cosine} \tag{7}$$

where, $\mathcal{L}_{mse}$ and $\mathcal{L}_{cosine}$ are defined as follows:

$$\mathcal{L}_{mse} = \frac{1}{N}\sum_{i=1}^{N}\left\|z_i - z_i^*\right\|_2^2 \tag{8}$$

$$\mathcal{L}_{cosine} = \frac{1}{N}\sum_{i=1}^{N} 1 - \cos\left(\angle\left(z_i - z_i^*\right)\right) \tag{9}$$

where $z_i = E_\Phi(x_i)$ and $z_i^* = D_\psi(y_i)$. The reconstruction loss here is shown in equation (5). Therefore, optimize the parameters of D_ψ with the following objective:

$$\psi = \operatorname{argmin}\left(\mathcal{L}_{feat} + \mathcal{L}_{gen}\right) \tag{10}$$

During training, the empirical hyperparameter of the feature loss term is set consistently with the literature [12], *i.e.*, $\mu = 0.9$, with the difference that a reconstruction loss is additionally introduced. Note that due to the dimensional differences in fMRI data across subjects, we trained the decoder model separately for each subject. Finally, the perceptual image x^* corresponding to fMRI y can be obtained by $G_\theta\left(D_\psi(y)\right)$.

E. Self-Supervised Training

The image encoder and generator represent two vital components of our proposed framework. The image encoder extracts visual representations from input images, and the generator converts these representations back into corresponding images. To enhance the performance of both the encoder and generator, we jointly trained these two networks on an additional image dataset. Specifically, we randomly selected 40,000 images from the ILSVRC2012 [34] and resized them to 128 × 128 pixels for self-supervised learning in the context of the reconstruction framework. It is important to note that there is no overlap between the selected images and the training or test images in the fMRI dataset.

IV. Experimental Results

A. Experimental Implementation

1) Dataset: To evaluate the efficacy of our proposed method, we conducted experiments on two publicly available fMRI datasets: Horikawa2017 [30] and vanGerven2010 [41].

vanGerven2010 dataset: This dataset comprises visual stimuli of the numbers 6 and 9 selected from the MNIST dataset, totaling 100 grayscale images with a resolution of 28 × 28. The choice of these specific numbers is due to their substantial dissimilarity. During the image display trials, fMRI data were collected from one subject while viewing the stimulus images, encompassing voxels in the V1, V2, and V3 regions of the visual cortex. For training purposes, we selected 90 {image, fMRI} data pairs from the dataset, while the remaining pairs were reserved for testing. For the reconstruction of handwritten digits, we trained the neural decoder using fMRI voxels from all of the above visual cortex regions.

Horikawa2017 dataset: In the image display trials of this dataset, fMRI signals were collected from five subjects while viewing a series of images randomly selected from the ImageNet dataset. The training trials comprised 1200 images belonging to 150 categories, and the test trials contained 50 images from different categories. Notably, the image categories in the test set did not overlap with those in the training set. During fMRI data collection, the training trials involved a single collection per image, whereas the test trials were collected 35 times per image. All images were displayed with fixation in a 3T scanner (TR, 3s; voxel size, 3 × 3 × 3 mm). In accordance with previous studies [21], the fMRI data collected for each test image were averaged to improve the signal-to-noise ratio (SNR). Additionally, this fMRI dataset provides masks for various visual cortex regions, including V1, V2, V3, V4, LOC, FFA, and PPA. For more details about the Horikawa2017 dataset, please refer to [30].

2) Evaluation Indicators: Considering the notable complexity differences in stimulus images between the Horikawa2017 and vanGerven2010 datasets, distinct evaluation criteria were employed for assessing the reconstruction quality of these two datasets. For vanGerven2010, we utilize the Pearson Correlation Coefficient (PCC) and the Structural Similarity Index (SSIM) [42] as evaluation indicators to facilitate comparison with prior studies. Given two images X and Y, the expression of PCC is:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X\sigma_Y} \tag{11}$$

where σ_X, σ_Y, and $cov(X,Y)$ are the standard deviation and covariance of X, Y, respectively. This metric can be used to assess the linear relationship between the reconstructed and original image. SSIM is a measure that quantifies human visual features, measuring the similarity of local structures between the reconstructed image and the original image. Its expression is:

$$SSIM = \frac{(2\mu_X\mu_Y + c_1)(2\sigma_{XY} + c_2)}{\left(\mu_X^2 + \mu_Y^2 + c_1\right)\left(\sigma_X^2 + \sigma_Y^2 + c_2\right)} \tag{12}$$

where μ_X, μ_Y and σ_X^2, σ_Y^2 represent mean and variance of X, Y, respectively. σ_{XY} denotes covariance, c_1 and c_2 are constants.

For the reconstruction of Horikawa2017 natural images, in order to objectively evaluate the reconstruction quality of HS-GAN, qualitative and quantitative comparisons are performed in this paper. For qualitative comparison, images reconstructed by different methods are shown directly. For quantitative comparison, we used six metrics: histogram similarity (HS) [43], mutual information (MI) [44], SSIM identification accuracy (SSIM-Acc), perceptual similarity identification accuracy (Perceptual-Acc), AlexNet(2) and AlexNet(5) identification accuracy. For histogram similarity, the formula is:

$$D(S,M) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{|s_i - m_i|}{\max(s_i, m_i)}\right) \tag{13}$$

where $S = \{s_1, s_2, \cdots, s_n\}$ and $M = \{m_1, m_2, \cdots, m_n\}$ denote histograms and n is the dimension of the histogram. MI is calculated using the following formula:

$$MI(A,B) = H(A) + H(B) - H(A,B) \tag{14}$$

where $H(A)$ and H(B) represent the information entropy of images A and B, respectively, and $H(A, B)$ is the joint entropy of A and B. The equations are as follows:

$$H(A) = -\sum_{i=0}^{N-1} p_i \log p_i \tag{15}$$

$$H(B) = -\sum_{i=0}^{N-1} p_i \log p_i \tag{16}$$

$$H(A, B) = -\sum_{ab} p_{AB}(a, b) \log p_{AB}(a, b) \tag{17}$$

For the identification accuracy metric, the recovered image is assessed using two candidate images: the actual image and a randomly selected one from the test set (excluding the actual image). If the reconstructed image is more similar to the actual image than the randomly selected one, the reconstruction is deemed successful. The formula is:

$$\text{Acc} = \frac{N_{\text{correct}}}{N_{\text{compare}}} \tag{18}$$

For the 50 images in the test set, a total of 2450 comparisons are made.

SSIM-Acc and Perceptual-Acc use Structural Similarity Index (SSIM) and LPIPS as similarity metrics, respectively, where the calculation of perceptual similarity is described in equation (3). AlexNet(2) and AlexNet(5) refer to the computation of PCC similarity using image features extracted from the second and fifth layers of AlexNet.

3) Implementation Details: Our proposed method was implemented using PyTorch, and model training was performed on an NVIDIA 3090 GPU. Self-supervision learning of the reconstruction network was conducted using 40,000 randomly selected images from the ILSVRC2012 dataset. The dimensions of hierarchical latent vectors and semantic feature vectors were set to 1024 and 4096, respectively. The dimension of the hidden layers of the neural decoder is 2048.

Training settings. During self-supervised training, the input images are resized to 128 × 128, the generator and discriminator are trained for 400 epochs using the Adam optimizer with an initial learning rate of 2×10^{-4}, and the cosine annealing learning rate tuning strategy is invoked. For training stability, the discriminator uses the Patch-GAN design with a patch size of 16. For neural decoder training, the initial learning rate is 3×10^{-4}, the weight decay is set to 1×10^{-2}, 240 epochs are trained using the Adam optimizer and a learning rate scheduler is employed. The batch size for all training sessions is 64. The loss curves regarding the generator and the discriminator during the training period are displayed in Fig. 6. It can be observed that the generator loss smoothly converges around 200 epochs, but we continue to train up to 400 epochs to obtain a more robust generator. Various hyperparameters of formula (5) in the image reconstruction network loss term were fine-tuned during the training process, that is, $\lambda_1 = 1.0$, $\lambda_2 = 0.01$, see the appendix.

B. Image Reconstruction Performance

1) Natural Image Reconstruction of Horikawa2017: We assess our approach on the Horikawa2017 dataset, and partial reconstruction examples are displayed in Fig. 7. From Fig. 7, our model captures the crucial characteristics of the object in the stimulus image, such as shape, contour, *etc.*, and performs well on all subjects. More reconstruction examples can be found in the appendix.

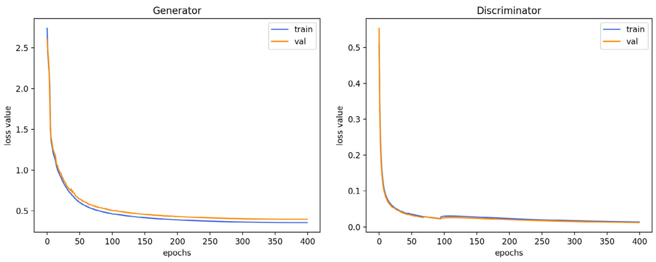

Fig. 6. The loss curves of the generator and discriminator during training process.

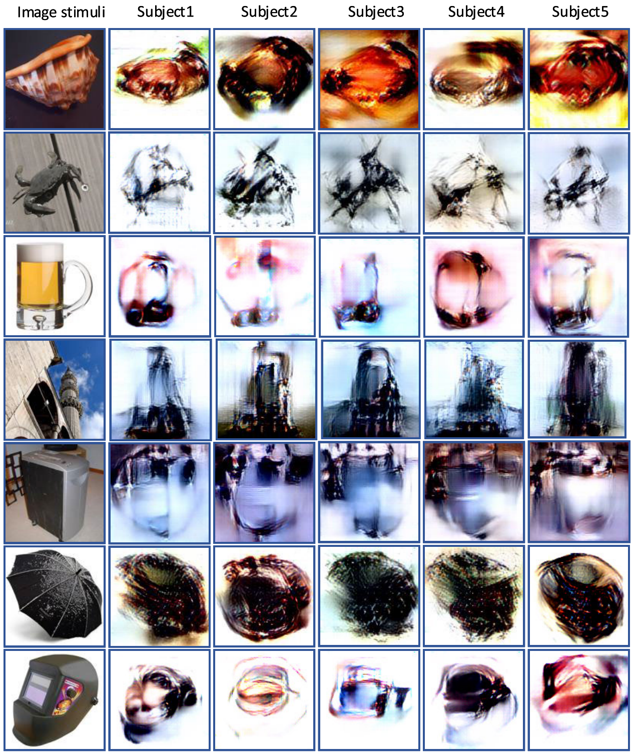

Fig. 7. Reconstruction examples for all subjects.

We also compared the image reconstruction results qualitatively and quantitatively with other state-of-the-art methods, including Shen et al. [20], Shen et al. [21], Beliy et al. [12], Fang et al. [31], Kai et al. [13], Ozcelik et al. [17] and Chen et al. [25]. Note that the focus of Ozcelik et al. and Chen et al. is different from our approach. They utilize the pre-trained generative model (GAN or LDM) on large-scale dataset to synthesize original image from a noise vector using fMRI as the conditional guide. However, for a broad comparison, we also provide their results. For qualitative comparison, we directly use the recovered images provided by the aforementioned authors in their respective papers. Fig. 8 showcases partially reconstructed images, all obtained from the fMRI data of Subject 3. To enhance the SNR, all fMRI voxels from the

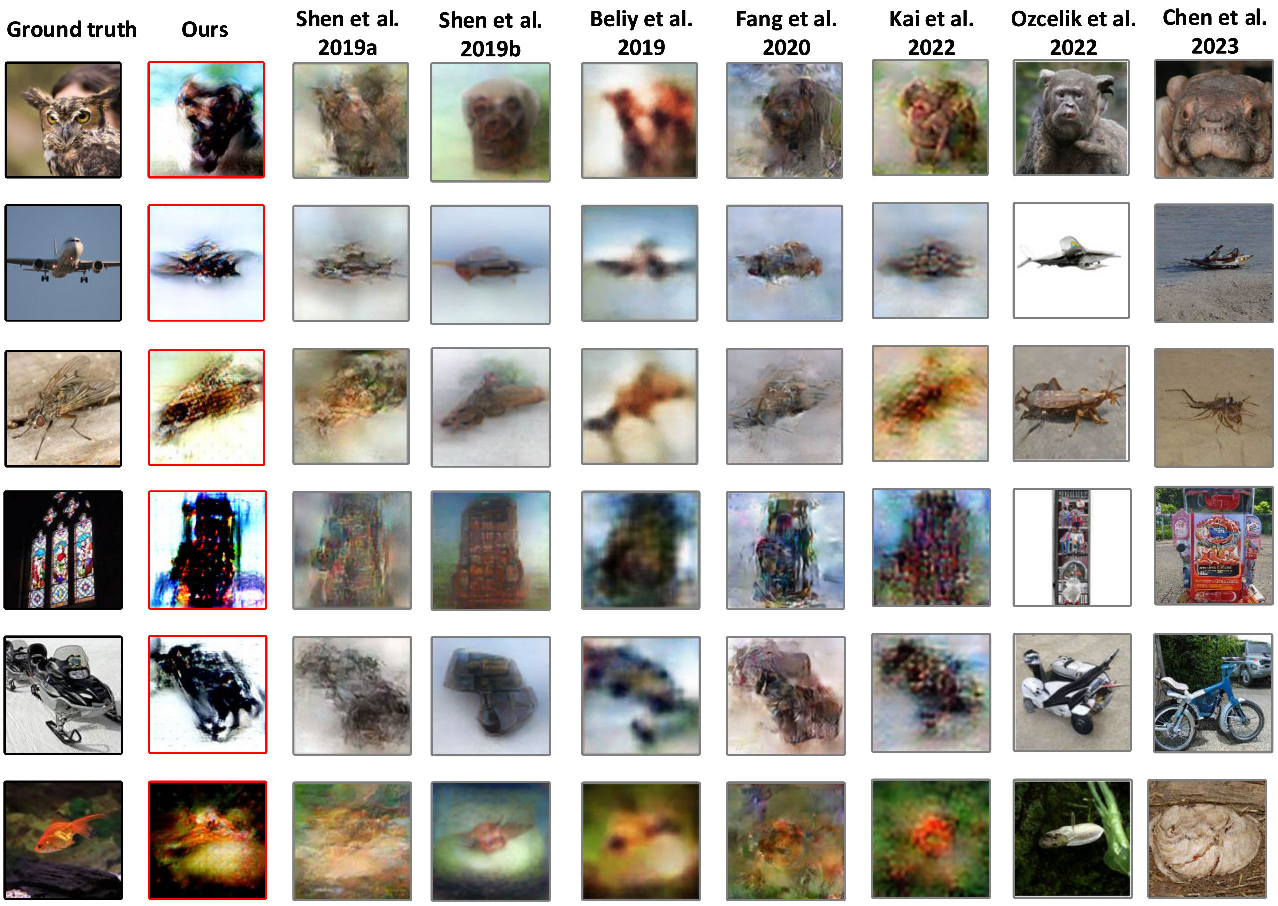

Fig. 8. Qualitative comparison of different methods to reconstruct natural images on the Horikawa2017 dataset.

TABLE I
QUANTITATIVE COMPARISON OF RECONSTRUCTION RESULTS OBTAINED USING fMRI OF SUBJECT 3

Methods	HS	MI	SSIM-Acc	Perceptual-Acc	AlexNet(2)	AlexNet(5)
Shen *et al.* 2019a [20]	0.392±0.091	0.649±0.133	63.22%	82.37%	87.59%	80.73%
Shen *et al.* 2019b [21]	0.401±0.079	**0.738±0.147**	63.06%	82.94%	84.16%	86.20%
Beliy *et al.* 2019 [12]	0.432±0.083	0.554±0.067	67.67%	70.98%	77.02%	68.98%
Fang *et al.* 2020 [31]	—	—	68.40%	84.50%	—	—
Kai *et al.*. 2022 [13]	—	—	71.60%	78.50%	—	—
Ozcelik *et al.* 2022 [17]	0.432±0.094	0.613±0.154	65.35%	89.631%	87.55%	94.10%
Chen *et al.* 2023 [25]	0.395±0.102	0.578±0.106	54.20%	79.79%	73.63%	88.37%
VD-VAE [45]	0.431±0.0895	0.662±0.150	69.88%	84.86%	86.45%	90.37%
HS-GAN	**0.447±0.095**	0.683±0.136	**78.90%**	**95.38%**	**96.24%**	**94.82%**
HS-GAN without self-training	0.435±0.010	0.662±0.149	74.31%	90.42%	91.86%	88.57%

test image were normalized and subsequently averaged. As demonstrated in Fig. 8, the images reconstructed by our method exhibit rich colors, clear contours, and better preservation of underlying details such as shape and texture from the original images. Consequently, our reconstructions appear more natural, clear, and recognizable, representing a significant advancement over previous methods. Compared with previous methods focusing on pixel reconstruction ([12], [13], [20], [21], [31]), HS-GAN achieves further improvement in reconstruction quality. However, fMRI data often suffer from spatial redundancy, noise, and sample sparsity, resulting in poor representation of fMRI signals and potential overfitting of noise distribution. The above challenges make it difficult for our decoder to accurately predict the corresponding image features from the fMRI voxels, resulting in the existing reconstruction still hardly to replicate the original stimulus (exhibiting blurry and unclear), and thus the realism of the reconstruction still needs to be improved. Moreover, compared with methods emphasizing semantic similarity ([17], [25]), our method retains more low-level

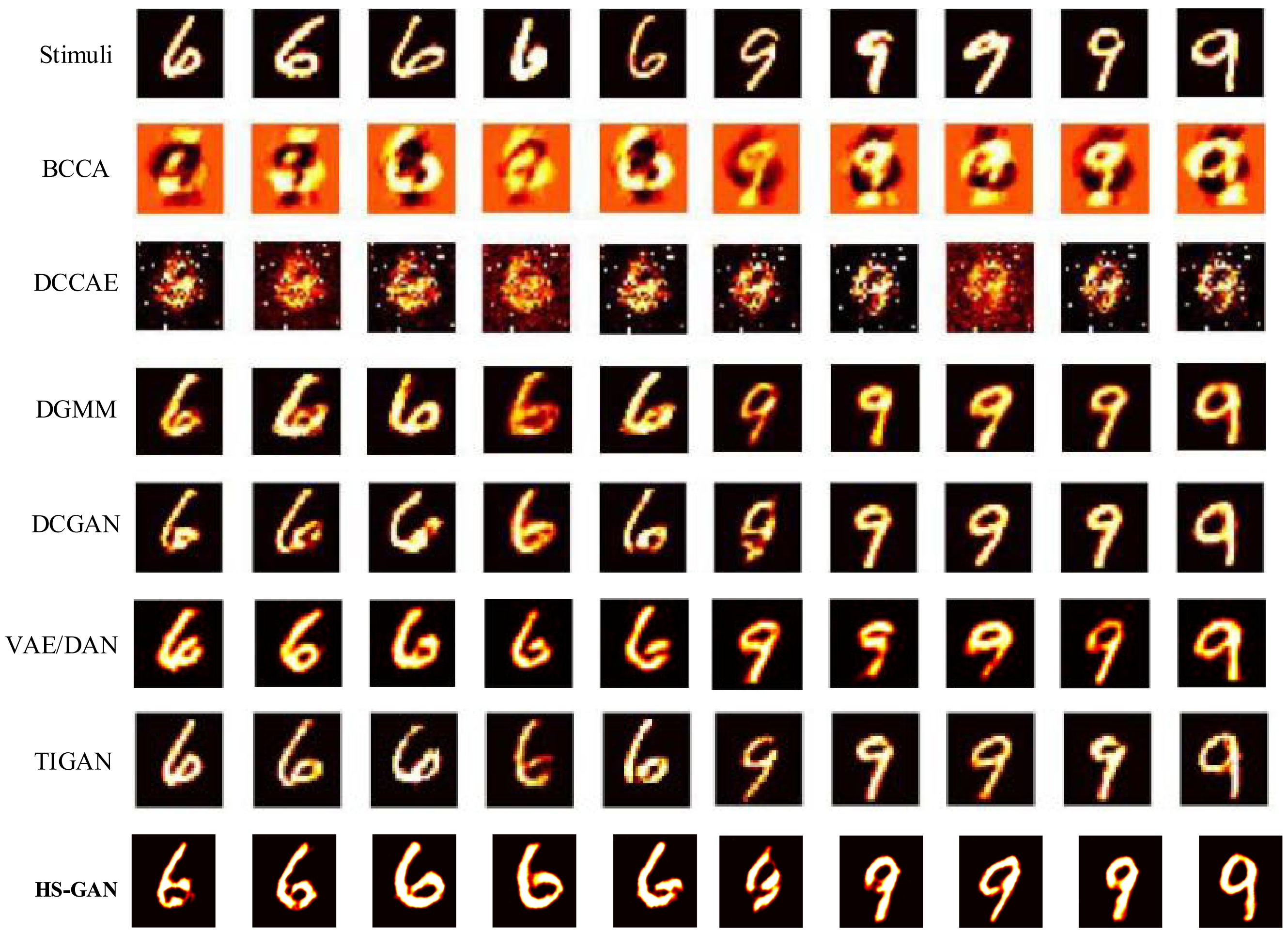

Fig. 9. Visual comparison of handwritten numeral reconstruction results.

TABLE II

Quantitative Comparison of Reconstruction Quality for vanGerven2010 Dataset, Where the Best Results Are Highlighted in Bold. (↑: The Higher the Value, the Better the Reconstruction Performance of the Method)

Methods	PCC ↑	SSIM ↑
BCCA [49]	0.411±0.157	0.192±0.035
DCCAE [50]	0.548±0.044	0.358±0.097
DGMM [47]	0.803±0.063	0.645±0.054
DCGAN [51]	0.531±0.049	0.529±0.043
DVAE/GAN [48]	**0.837±0.014**	0.714±0.014
TIGAN [46]	0.812±0.059	0.729±0.021
HS-GAN	0.796±0.051	**0.783±0.038**

visual features of the original image, providing more realistic and reliable reconstruction. Although semantic-focused approaches can produce relatively high-quality images, it is hardly to ensure that the recovered images are consistent with the semantic information of fMRI, as shown in Fig. 8.

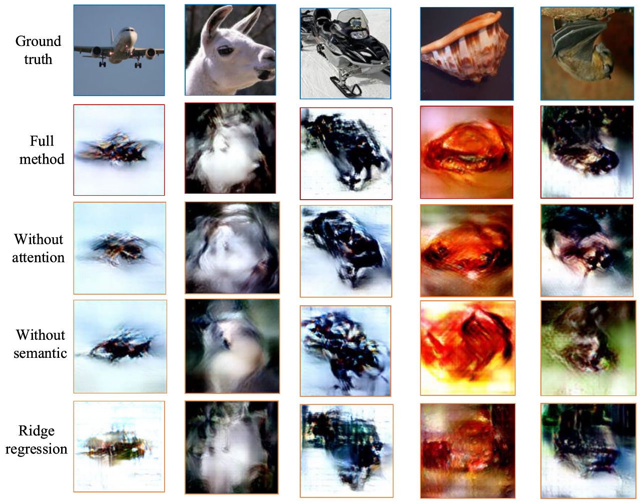

Fig. 10. Qualitative comparative results of ablation experiments with different model components.

To provide an objective evaluation of the reconstruction performance of our proposed method, we quantitatively compared the results with the aforementioned methods using six metrics mentioned above. Since not all methods enable the calculation of the above metrics (depending on the content provided by the author), corresponding metrics for the different methods are reported. Notably, since all reconstructions provided in the papers are for Subject 3, we uniformly used the results

TABLE III
QUANTITATIVE COMPARISON RESULTS OF ABLATION EXPERIMENTS WITH DIFFERENT MODEL COMPONENTS

Models	HS	MI	SSIM-Acc	Perceptual-Acc	AlexNet(2)	AlexNet(5)
Without semantic encoder	0.430±0.099	0.659±0.141	71.59%	90.71%	89.87%	86.65%
Without attention module	0.438±0.010	0.668±0.143	76.18%	92.37%	94.78%	93.29%
Ridge regression decoder	0.433±0.091	0.647±0.132	73.43%	90.78%	91.6%	91.10%
Full method	**0.447±0.095**	**0.683±0.136**	**78.90%**	**95.38%**	**96.24%**	**94.82%**

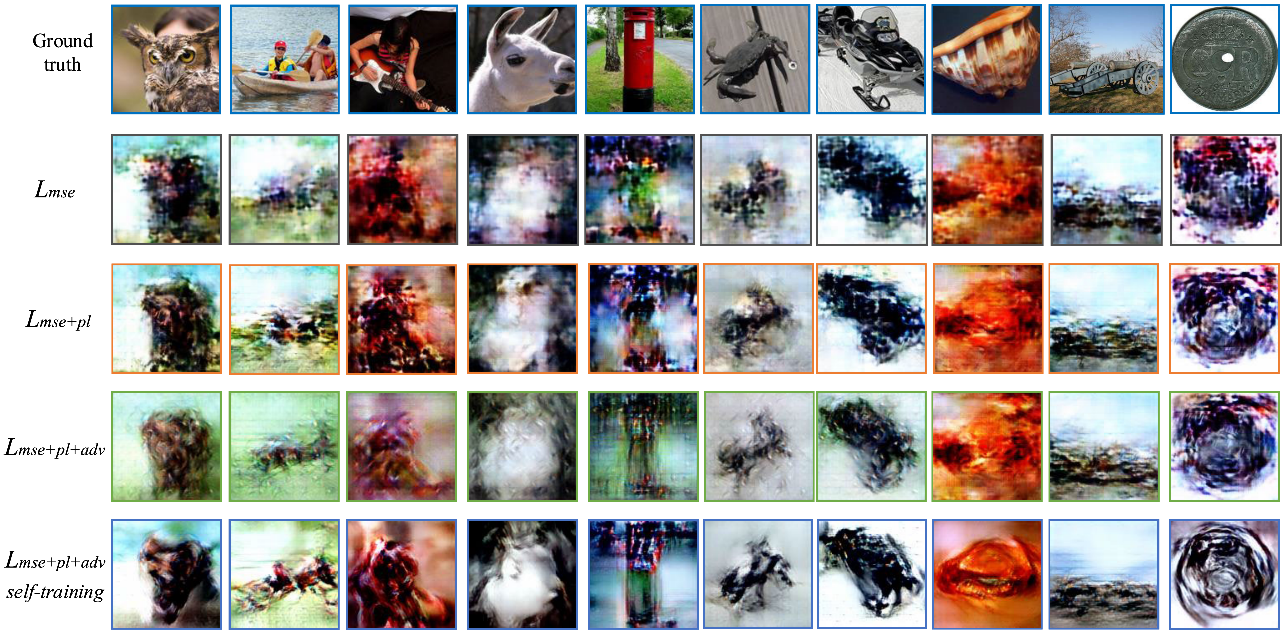

Fig. 11. Qualitative comparison of the generator using different loss functions.

TABLE IV
QUANTITATIVE EVALUATION OF DIFFERENT LOSS FUNCTIONS FOR THE GENERATOR

Loss	HS	MI	SSIM-Acc	Perceptual-Acc	AlexNet(2)	AlexNet(5)
L_{mse}	0.435±0.099	0.664±0.159	65.77%	78.76%	81.84%	75.84%
$L_{mse} + L_{pl}$	0.432±0.104	0.664±0.151	71.72%	88.53%	90.08%	84.12%
$L_{mse} + L_{pl} + L_{adv}$	0.435±0.010	0.662±0.149	74.31%	90.42%	91.86%	88.57%
$L_{mse} + L_{pl} + L_{adv}$+self−training	**0.447±0.095**	**0.683±0.136**	**78.90%**	**95.38%**	**96.24%**	**94.82%**

TABLE V
QUANTITATIVE EVALUATION OF DIFFERENT LOSS FUNCTIONS FOR THE DECODER

Loss	HS	MI	SSIM-Acc	Perceptual-Acc	AlexNet(2)	AlexNet(5)
L_{feat}	0.436±0.103	0.669±0.148	72.48%	92.10%	93.59%	92.94%
L_{gen}	0.394±0.095	0.621±0.130	72.01%	80.53%	87.84%	83.35%
$L_{feat} + L_{gen}$	**0.447±0.095**	**0.683±0.136**	**78.90%**	**95.38%**	**96.24%**	**94.82%**

of Subject 3 for the indicator calculations. As shown in Table I, HS-GAN reconstructed images obtained the highest histogram similarity of 0.447, and the mutual information similarity for the images is second only to the method of Shen et al. [21], which indicates that our proposed method better preserves the low-level visual features of the original images and achieves a more reliable reconstruction. Furthermore, HS-GAN obtained the highest SSIM-Acc and Perceptual-Acc (78.90% and 95.38%, respectively), indicating that our reconstructed images are more consistent with human visual perception. For the metrics computed in the AlexNet feature space, HS-GAN also achieves the best performance (AlexNet(2) 96.24%, AlexNet(5) 94.82%), demonstrating that the reconstructed images also retain the high-level features of the original images. We also applied a hierarchical variational autoencoder VDVAE [45] (Very Deep VAE) for visual reconstruction, and the quantitative comparison results demonstrate that the reconstruction of HS-GAN is superior to that of VDVAE, which further proves the superiority of our model. A description of the visual reconstruction utilizing VDVAE is provided in the Appendix, and some of the reconstructed images are presented. Overall, HS-GAN achieves the best performance, with advantages in all quantitative metrics.

It should be noted that although the reconstructed images of Ozcelik et al. and Chen et al. appear visually more plausible, their reconstructions tend to differ significantly from the original images, resulting in lack of reliability. As a result, the performance in the assessment of pixel and perceptual similarity metrics is not impressive. However, for the visual reconstruction task, consistency of reconstruction is more important than diversity. Particularly for applications in real-world scenarios, such as the diagnosis of neurological diseases. In addition, we additionally provide the quantitative evaluation when the model is not trained with additional data, and it can be found that our method also achieves the best performance on several metrics (HS: 0.435, SSIM-Acc: 74.31%, Perceptual-Acc: 90.42%, AlexNet (2): 91.86%) compared to other approaches. This demonstrates the effectiveness of our model design.

2) Grayscale Digital Image Reconstruction of vanGerven2010: In order to assess the generalization ability of our model beyond natural images, we conducted experiments on the reconstruction of handwritten digital images from the vanGerven2010 dataset. This task presents a challenge as our image feature encoder and generator were initially trained on natural images, without additional training on handwritten characters. Specifically, we fixed the parameters of the image feature encoder and generator, and then inputted the numeric characters scaled to 128 × 128 pixels into the encoder to extract visual features. Subsequently, a neural decoder was employed to map fMRI signals to the latent vectors. Finally, the predicted latent vectors were fed into the generator to reconstruct the corresponding handwritten digits. The reconstruction results are depicted in Fig.9, where it is evident that our model successfully reconstructs the digits 6 and 9.

Table II presents the quantitative comparison results for the vanGerven2010 dataset. Notably, our method achieves the highest Structural Similarity Index (SSIM) of 0.783, an improvement of 7.4% compared to TIGAN [46]. In visual comparison, HS-GAN reconstructed images have clearer contours. This is mainly attributed to: (1) Our model utilizes diverse visual features at different levels to reconstruct the stimulus image and introduce reconstruction loss in the neural decoder. (2) The specially designed generator effectively transmits more low-level details from the original image into the reconstruction space. Although DGMM [47], TIGAN [46], and DVAE/GAN [48] achieve higher Pearson Correlation Coefficient (PCC) values, they also exhibit the issue of blurred reconstruction. These comparative results demonstrate the robust versatility of our model, which is not simply limited to template matching, making it suitable for reconstructing images from other domains as well. Moreover, the adaptability of the model pre-trained on complex images to perform well in simpler image reconstruction tasks is evident from our findings.

C. Ablation Studies of Different Components

Our proposed HS-GAN incorporates several crucial components, including the semantic encoder, self-attention module, and neural decoder. In this section, we conduct ablation experiments to examine the effects of these components on model performance. The specific experimental results are presented in Fig. 10 and Table III.

As shown in Fig. 10, incorporating semantic features in the generative model allows the reconstructed images to have more accurate shapes, textures, and colors. For example, the shell in the fourth column, the reconstructed image after adding semantic features is visually more similar to the original image. The attention module allows the generator to reconstruct the original image with more precise details, such as airplanes and bats, the reconstructed outline is more similar to the real image after adding the attention module. The neural decoder proposed in this paper obtains more natural reconstructions compared to ridge regression, which is commonly used in previous methods, and achieves better performance in quantitative evaluation. This can be interpreted in two aspects: (1) Ridge regression can only capture linear relationships between fMRI patterns and DNN features, which may have complicated nonlinear relationships. (2) Ridge regression ignores the correlation between DNN feature units, while our decoding model is able to capture this correlation.

We present the results of the quantitative evaluation in Table III, where the ridge regression decoder means that using ridge regression to learn the mapping from fMRI to latent features. From Table III, it can be seen that using the complete model achieves the best reconstruction quality, and different model components contribute to improving network performance. Especially, the introduction of category information significantly improves the reconstruction quality, which proves the effectiveness of our method.

D. Impact of Different Loss Functions

1) Generator Loss Functions: To evaluate the effectiveness of introducing perceptual loss and adversarial loss during generator training, we trained our model using three loss functions: L_{img}, $L_{img} + L_{pl}$, and $L_{img} + L_{pl} + L_{adv}$. The reconstructed images and quantitative evaluation results are presented in Fig. 11 and Table IV, respectively. Our findings indicate that:

1) Using only image loss results in fuzzy and difficult-to-recognize images, with the lowest recognition accuracy. This is attributed to the MSE loss function causing the reconstructed images to lose precise details from the original images.
2) The introduction of perceptual loss leads to clearer images with distinct outlines. This is because that perceptual loss places more emphasis on perceptually

TABLE VI
QUANTITATIVE EVALUATION OF SELF-SUPERVISED TRAINING STRATEGY

Training Strategy	HS	MI	SSIM-Acc	Perceptual-Acc	AlexNet(2)	AlexNet(5)
without self-training	0.435±0.010	0.662±0.149	74.31%	90.42%	91.86%	88.57%
with self-training	**0.447±0.095**	**0.683±0.136**	78.90%	**95.38%**	**96.24%**	94.82%
self-training on larger dataset	0.442±0.096	0.675±0.157	**79.30%**	95.30%	96.07%	**95.20%**

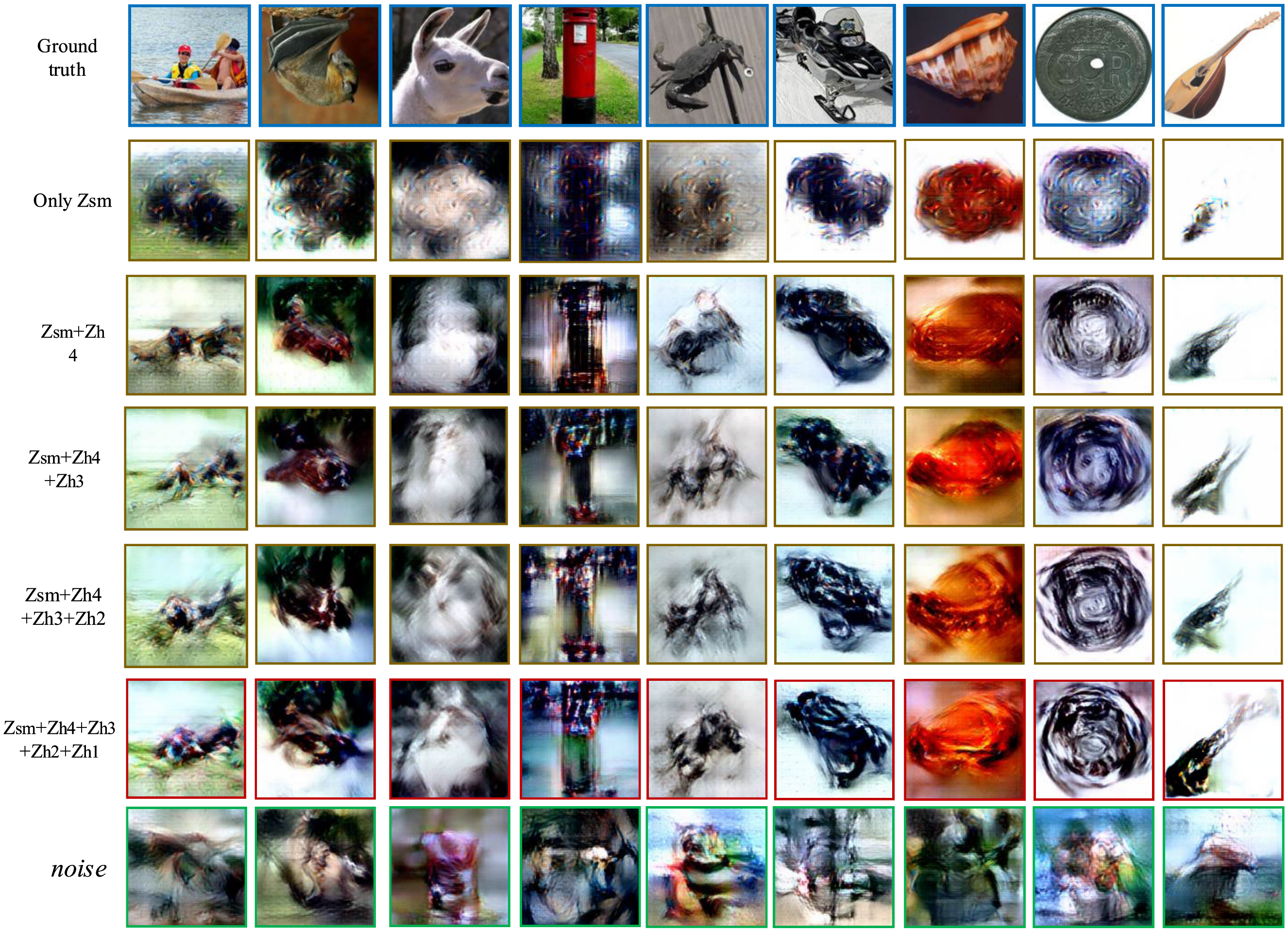

Fig. 12. Qualitative comparisons of reconstruction results by fusing different levels of features.

TABLE VII
QUANTITATIVE EVALUATION INCORPORATING DIFFERENT LEVELS OF FEATURES

Loss	HS	MI	SSIM-Acc	Perceptual-Acc	AlexNet(2)	AlexNet(5)
only z_{sm}	0.428±0.104	0.601±0.161	68.33%	82.28%	78.33%	83.02%
$z_{sm} + z_{h4}$	0.436±0.096	0.679±0.158	71.55%	92.86%	91.47%	90.26%
$z_{sm} + z_{h4} + z_{h3}$	0.440±0.097	0.673±0.148	73.67%	93.14%	93.67%	91.47%
$z_{sm} + z_{h4} + z_{h3} + z_{h2}$	**0.459±0.093**	**0.685±0.143**	75.63%	94.04%	93.63%	93.91%
$z_{sm} + z_{h4} + z_{h3} + z_{h2} + z_{h1}$	0.447±0.095	0.683±0.136	**78.90%**	**95.38%**	**96.24%**	**94.82%**
noise	0.422±0.088	0.611±0.075	50.45%	53.51%	52.37%	50.08%

important features (e.g., edges and textures) and is less sensitive to subtle changes in the image. In addition, the recognition accuracy of SSIM, Perceptual, AlexNet(2) and AlexNet(5) was also significantly improved, which indicates that the reconstructed image recovers most of the visual features of the original image, demonstrating the effectiveness of introducing perceptual loss.

3) Furthermore, adding adversarial loss further improves the quality of the reconstructed images by enforcing

TABLE VIII
LPIPS COMPARISON WITH VARIOUS λ_1 AND λ_2, THE LOWER THE VALUE, THE BETTER

λ_1 \ λ_2	10^{-3}	10^{-2}	0.1	1.0	10
10^{-3}	0.413	0.537	0.462	0.726	0.781
10^{-2}	0.343	0.377	0.368	0.564	0.614
0.1	0.320	0.329	0.310	0.413	0.579
1.0	0.258	**0.226**	0.269	0.274	0.372
10	0.311	0.317	0.374	0.364	0.472

the generator to produce more natural-looking images. Additionally, by introducing natural image prior information through self-supervised learning, the images reconstructed by the generator become more natural and recognizable, achieving the highest recognition accuracy.

4) Using pixel-level similarity evaluation indicators, blurred images also have high similarity, which is inconsistent with human visual perception. For example, although the images reconstructed using the MSE loss were blurry, their HS and MI metrics also received high evaluations compared with other comparative experiments. Therefore, we prefer to use the similarity assessment in the image feature space to measure the reconstruction performance.

2) Neural Decoder Loss Functions: To demonstrate the effectiveness of introducing reconstruction loss in the training process of the neural decoder, ablation experiments with different loss functions of the neural decoder are performed in this section.

The results of quantitative comparisons are presented in Table V, it can be observed that the quality of the reconstructions can be improved by introducing reconstruction loss, and the best performance is obtained in all evaluation indicators. This is because our ultimate target is to reconstruct realistic and reliable stimulus images, and the introduction of the reconstruction loss allows the latent features predicted by the decoder to be more suitable for generating natural images. However, when only reconstruction loss is employed, it is difficult to ensure accurate alignment of the decoder latent space with the image encoding space, which leads to low-quality reconstruction.

E. Effectiveness of Self-Supervised Training Strategy

In this section, we performed ablation experiments to verify the effectiveness of using the self-supervised training strategy. The results regarding the quantitative assessment are displayed in Table VI.

It can be observed that the use of self-supervised training strategy significantly improves the quality of model reconstruction, and achieves better performance on all quantitative metrics. This indicates that jointly training the image encoder and generator on additional image data to introduce the prior information of natural images, which can make the images reconstructed by the generator more natural. In addition, to explore the impact of the additional dataset on the reconstruction performance of HS-GAN, we use a larger number of images (100,000) for self-supervised training. The results in Table VI show that the reconstruction performance of the model does not change significantly when using more image data (close to the performance when using 40,000 images). This proves that our model is not data-hungry. Considering the computational burden of a larger dataset, we use 40,000 images for self-supervised training of the model.

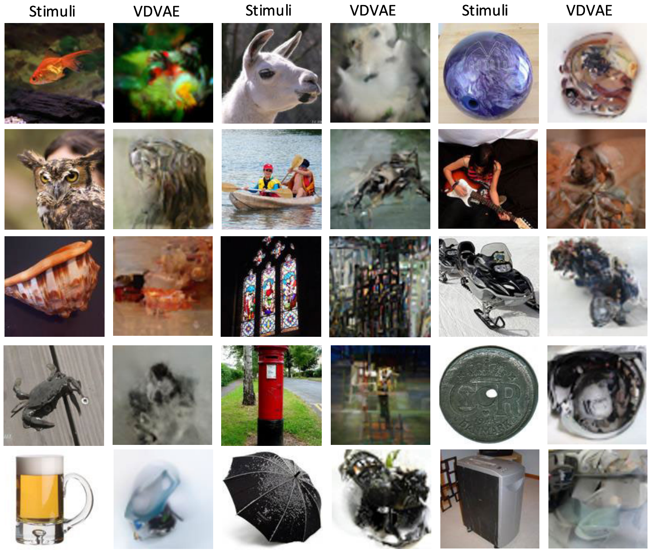

Fig. 13. Examples of VDVAE reconstruction.

F. Effectiveness of Hierarchical Features

We utilize hierarchical and semantic features of images for visual reconstruction, and to demonstrate the effectiveness of merging different levels of image features, we perform comparative experiments. Qualitative and quantitative comparisons of the reconstruction results are shown in Fig. 12 and Table VII, respectively.

From Fig. 12, it can be observed that when only semantic feature is used for reconstruction, the image quality is not satisfactory (the generated image is blurry and hard to identify) and the quantitative evaluation shows the lowest performance. We believe that this is because the semantic feature loses most of the low-level features in the original image, making it difficult for generator to recover the precise details. When the visual feature z_{h4} of the image is fused, more low-level features can be transmitted into the reconstruction space, so the shape, contour and color of the reconstructed image are more precise, and the model performance is significantly improved.

As shown in Table VII, as more levels of image features are introduced into the generator, the reconstruction quality is further improved. Especially in visual comparison, the reconstructed image using the complete method maintains maximum consistency with the original image in terms of low-level features (shape, color, etc.). In quantitative comparison, *noise* refers to the quantitative evaluation result of the reconstructed image obtained by feeding the random noise from a standard gaussian distribution into the decoder.

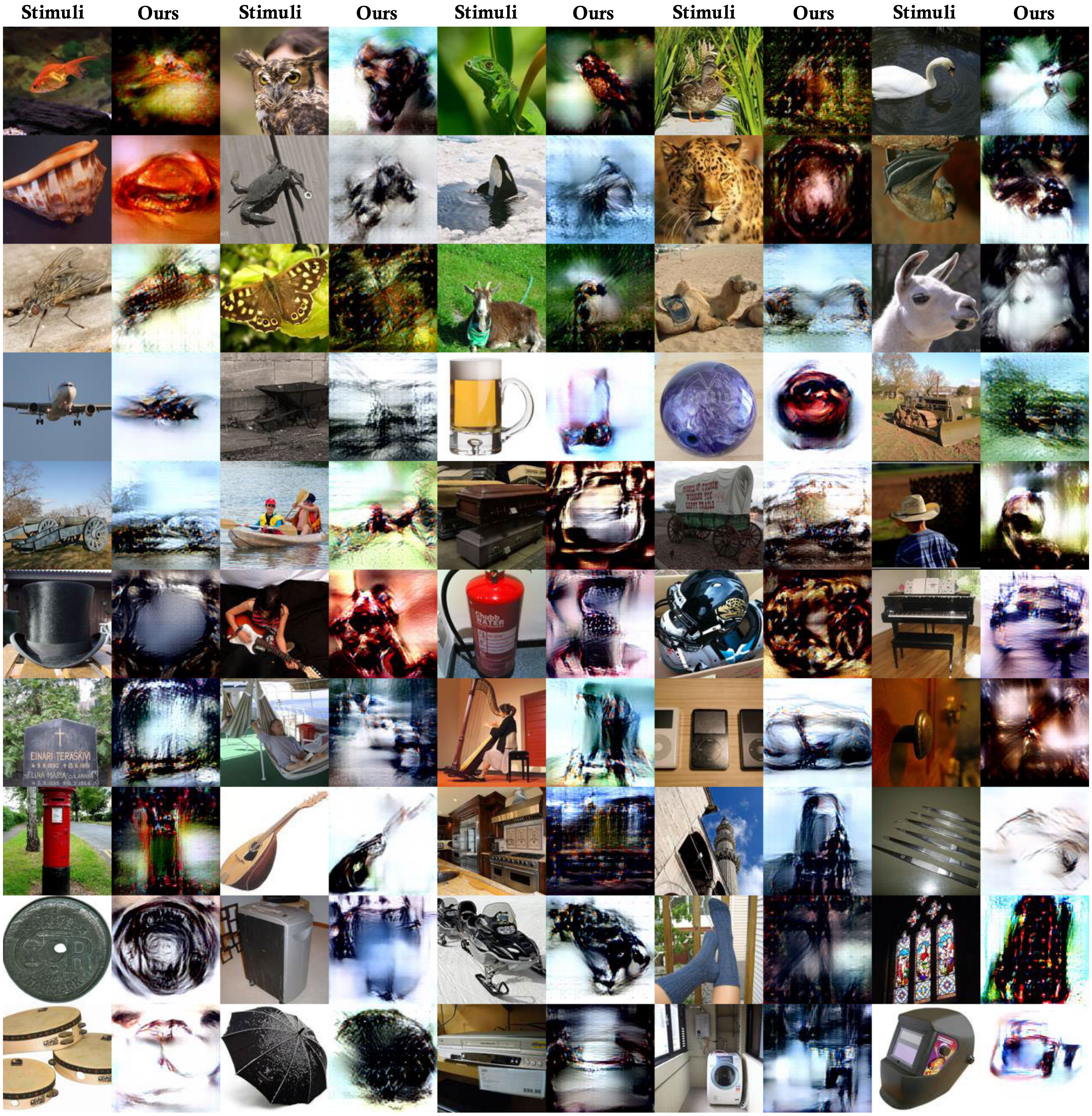

Fig. 14. Full samples for Subject 3 in Horikawa2017 test set.

Overall, our complete method achieves the best performance (SSIM-Acc: 78.90%, Perceptual-Acc: 95.38%, AlexNet(2): 96.24%, AlexNet(5): 94.82%) and the recognition accuracy far exceeds that of noise-based reconstruction. This demonstrates that HS-GAN learns the complex mapping of fMRI signals to visual features of stimulus images, and the reconstruction is consistent with human visual perception.

V. Conclusion and Discussion

Conclusion: In this research, we introduced a novel approach, the semantics-guided hierarchical feature encoding Generative Adversarial Network (GAN), to address the challenge of reconstructing visual images from fMRI recordings. Our method draws inspiration from the hierarchical encoding observed in the visual cortex and the homology of information processing between the brain and deep neural networks. Our proposed framework consists of an image feature encoder, which extracts hierarchical and semantic features from input images and encodes them as latent vectors. Subsequently, a neural decoder with residual connections is trained to learn the representation from the fMRI signal to the image feature space. Finally, the predicted hierarchical and semantic features are combined to reconstruct the image through the generator. The validation of our approach was conducted using two publicly available datasets, and we compared its performance with other advanced methods. By leveraging information from different visual cortex regions, our method achieved significantly improved results, yielding reconstructed visual images that are more natural and recognizable compared to previous approaches.

Discussion: This research opens up promising avenues for further advancements in the field of decoding visual information from brain activity. Although our approach achieves competitive results on realistic and semantic consistency of reconstructed images, there are still some limitations. First, due to the high cost of collecting fMRI data, it is difficult to obtain a large number of paired samples, which makes it difficult for the decoder to accurately predict the corresponding image features from the fMRI voxels. In future work, building deep learning models that can effectively understand fMRI patterns will better facilitate downstream tasks. Second, our model employs a two-stage training strategy, using latent feature vectors as the medium for fMRI voxel to image transformation, reducing the dependence on paired samples. However, the two-stage approach somewhat leads to information loss in fMRI, thus realizing an end-to-end decoding model from fMRI to image still needs attention. In our experiments, we observed some variation in reconstruction quality across subjects, although this is common in other decoding efforts. In the future, we should investigate versatile cross-subject models to efficiently project fMRI representations from different subjects into the same embedding space. In addition, while existing methods have significantly improved the quality of reconstructed images from fMRI, what is the upper limit? The exploration of this question in future studies is expected to provide new insights in the field of visual decoding.

Appendix

Here, we describe how to use the VDVAE as a generator for visual reconstruction. The VDVAE is a hierarchical variational autoencoder model that consists of 75 layers and is pre-trained on the ImageNet dataset. Specifically, we first trained a neural decoder D_ψ to learn the mapping of fMRI voxels to the embedding space of the VDVAE encoder. Here we employed the embedding vectors of the first 31 layers of VDVAE and combine them into a 91168-dimensional feature vector z. Subsequently, we feed the feature latent vector $\hat{z}$ predicted by the decoder into the image decoder of the VDVAE to obtain the corresponding reconstruction. Partial reconstructed images are shown in Fig. 13.

In addition, We present all reconstruction examples (50 categories) of subject3 in Horikawa 2017 dataset, see Fig 14. Through the reconstruction results, we found that our method performed well on images centered on a single object (*e.g.*, a shell), but not satisfactorily for images with complex backgrounds (row 7, column 4). This may be attributed to the interference of image background on the subject's attention, resulting in a lower signal-to-noise ratio of the recorded fMRI signal. To explore the impact of λ_1 and λ_2 on the image reconstruction performance, we computed the LPIPS for models with different parameters $\lambda_1 = [0.001, 0.01, \ldots, 10]$ and $\lambda_2 = [0.001, 0.01, \ldots, 10]$, and the results on validation set are shown in Table VIII. It can be seen that the best results are obtained at $\lambda_1 = 1.0$ and $\lambda_2 = 0.01$.

References

[1] D. Marr and L. Vaina, "Representation and recognition of the movements of shapes," *Proc. Roy. Soc. London. Ser. B. Biol. Sci.*, vol. 214, no. 1197, p. 501–524, 1982.

[2] J. Belliveau et al., "Functional mapping of the human visual cortex by magnetic resonance imaging," *Science*, vol. 254, no. 5032, pp. 716–719, 1991.

[3] Z. Rakhimberdina, Q. Jodelet, X. Liu, and T. Murata, "Natural image reconstruction from fMRI using deep learning: A survey," *Frontiers Neurosci.*, vol. 15, Dec. 2021, Art. no. 795488.

[4] R. A. Poldrack and M. J. Farah, "Progress and challenges in probing the human brain," *Nature*, vol. 526, no. 7573, pp. 371–379, Oct. 2015.

[5] Y. Miyawaki et al., "Visual image reconstruction from human brain activity using a combination of multiscale local image decoders," *Neuron*, vol. 60, no. 5, pp. 915–929, Dec. 2008.

[6] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant, "Bayesian reconstruction of natural images from human brain activity," *Neuron*, vol. 63, no. 6, pp. 902–915, Sep. 2009.

[7] B. Thirion et al., "Inverse retinotopy: Inferring the visual content of images from brain activation patterns," *NeuroImage*, vol. 33, no. 4, pp. 1104–1116, Dec. 2006.

[8] E. Yargholi and G.-A. Hossein-Zadeh, "Reconstruction of digit images from human brain fMRI activity through connectivity informed Bayesian networks," *J. Neurosci. Methods*, vol. 257, pp. 159–167, Jan. 2016.

[9] S. Schoenmakers, M. Barth, T. Heskes, and M. van Gerven, "Linear reconstruction of perceived images from human brain activity," *NeuroImage*, vol. 83, pp. 951–961, Dec. 2013.

[10] T. Horikawa and Y. Kamitani, "Hierarchical neural representation of dreamed objects revealed by brain decoding with deep neural network features," *Frontiers Comput. Neurosci.*, vol. 11, p. 4, Jan. 2017.

[11] H. Wen, J. Shi, Y. Zhang, K.-H. Lu, J. Cao, and Z. Liu, "Neural encoding and decoding with deep learning for dynamic natural vision," *Cerebral Cortex*, vol. 28, no. 12, pp. 4136–4160, Dec. 2018.

[12] R. Beliy, G. Gaziv, A. Hoogi, F. Strappini, T. Golan, and M. Irani, "From voxels to pixels and back: Self-supervision in natural-image reconstruction from FMRI," in *Proc. NIPS*. Red Hook, NY, USA: Curran Associates, 2019, pp. 1–11.

[13] K. Chen, Y. Ma, M. Sheng, and N. Zheng, "Foreground-attention in neural decoding: Guiding loop-Enc-Dec to reconstruct visual stimulus images from fMRI," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Padua, Italy, Jul. 2022, pp. 1–8, doi: 10.1109/IJCNN55064.2022.9892276.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[15] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. NIPS*, Montreal, QC, Canada, 2014, pp. 2672–2680.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "Highresolution image synthesis with latent diffusion models," 2022, *arXiv:2112.10752*.

[17] F. Ozcelik, B. Choksi, M. Mozafari, L. Reddy, and R. VanRullen, "Reconstruction of perceived images from fMRI patterns and semantic brain exploration using instance-conditioned GANs," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Padua, Italy, Jul. 2022, pp. 1–8.

[18] R. VanRullen and L. Reddy, "Reconstructing faces from fMRI patterns using deep generative neural networks," *Commun. Biol.*, vol. 2, no. 1, pp. 193–203, May 2019, doi: 10.1038/s42003-019-0438-y.

[19] T. Dado et al., "Hyperrealistic neural decoding: Reconstructing faces from fMRI activations via the GAN latent space," *Sci. Rep.*, vol. 12, no. 1, pp. 141–150, Jan. 2022, doi: 10.1038/s41598-021-03938-w.

[20] G. Shen, K. Dwivedi, K. Majima, T. Horikawa, and Y. Kamitani, "End-to-end deep image reconstruction from human brain activity," *Frontiers Comput. Neurosci.*, vol. 13, p. 10, Apr. 2019, doi: 10.3389/fncom.2019.00021.

[21] G. Shen, T. Horikawa, K. Majima, and Y. Kamitani, "Deep image reconstruction from human brain activity," *PLOS Comput. Biol.*, vol. 15, no. 1, Jan. 2019, Art. no. e1006633.

[22] E. J. Allen et al., "A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence," *Nature Neurosci.*, vol. 25, no. 1, pp. 116–126, Jan. 2022, doi: 10.1038/s41593-021-00962-x.

[23] S. Lin, T. Sprague, and A. K. Singh, "Mind reader: Reconstructing complex images from brain activities," 2022, *arXiv:2210.01769*.

[24] Y. Takagi and S. Nishimoto, "High-resolution image reconstruction with latent diffusion models from human brain activity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, p. 14453, doi: 10.1109/cvpr52729.2023.01389.

[25] Z. Chen, J. Qing, T. Xiang, W. Lin Yue, and J. Helen Zhou, “Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding,” 2022, *arXiv:2211.06956*.

[26] F. Ozcelik and R. VanRullen, “Natural scene reconstruction from fMRI signals using generative latent diffusion,” *Sci. Rep.*, vol. 13, no. 1, p. 15666, Sep. 2023.

[27] M. Ferrante, T. Boccato, and N. Toschi, “Semantic brain decoding: From fMRI to conceptually similar image reconstruction of visual stimuli,” 2022, *arXiv:2212.06726*.

[28] P. S. Scotti et al., “Reconstructing the mind’s eye: FMRI-to-image with contrastive learning and diffusion priors,” 2023, *arXiv:2305.18274*.

[29] Y. Liu, Y. Ma, W. Zhou, G. Zhu, and N. Zheng, “BrainCLIP: Bridging brain and visual-linguistic representation via CLIP for generic natural visual stimulus decoding,” 2023, *arXiv:2302.12971*.

[30] T. Horikawa and Y. Kamitani, “Generic decoding of seen and imagined objects using hierarchical visual features,” *Nature Commun.*, vol. 8, no. 1, pp. 1–15, May 2017.

[31] T. Fang, Y. Qi, and G. Pan, “Reconstructing perceptive images from brain activity by shape-semantic GAN,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Vancouver, BC, Canada, 2020, pp. 13038–13048.

[32] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[33] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, 2009, pp. 248–255.

[35] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” 2015, *arXiv:1505.04597*.

[36] A. Vaswani et al., “Attention is all you need,” in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA, 2017, pp. 6000–6010.

[37] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake, UT, USA, Jun. 2018, pp. 586–595, doi: 10.1109/CVPR.2018.00068.

[38] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” 2016, *arXiv:1603.08155*.

[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[40] Z. Liu, C. Rios, N. Zhang, L. Yang, W. Chen, and B. He, “Linear and nonlinear relationships between visual stimuli, EEG and BOLD fMRI signals,” *NeuroImage*, vol. 50, no. 3, pp. 1054–1066, 2010.

[41] M. A. J. van Gerven, F. P. de Lange, and T. Heskes, “Neural decoding with hierarchical generative models,” *Neural Comput.*, vol. 22, no. 12, pp. 3127–3142, Dec. 2010.

[42] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[43] Y. Ma, X. Gu, and Y. Wang, “Histogram similarity measure using variable bin size distance,” *Comput. Vis. Image Understand.*, vol. 114, no. 8, pp. 981–989, Aug. 2010.

[44] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, “Mutual-information-based registration of medical images: A survey,” *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.

[45] R. Child, “Very deep VAEs generalize autoregressive models and can outperform them on images,” 2020, *arXiv:2011.10650*.

[46] S. Huang, L. Sun, M. Yousefnezhad, M. Wang, and D. Zhang, “Temporal information-guided generative adversarial networks for stimuli image reconstruction from human brain activities,” *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 3, pp. 1104–1118, Sep. 2022.

[47] C. Du, C. Du, L. Huang, and H. He, “Reconstructing perceived images from human brain activities with Bayesian deep multiview learning,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 8, pp. 2310–2323, Aug. 2019.

[48] Z. Ren et al., “Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning,” *NeuroImage*, vol. 228, Mar. 2021, Art. no. 117602.

[49] Y. Fujiwara, Y. Miyawaki, and Y. Kamitani, “Modular encoding and decoding models derived from Bayesian canonical correlation analysis,” *Neural Comput.*, vol. 25, no. 4, pp. 979–1005, 2013.

[50] W. Wang, R. Arora, K. Livescu, and J. Bilmes, “On deep multiview representation learning,” in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, Lille, France, 2015, pp. 1083–1092.

[51] K. Seeliger, U. Güçlü, L. Ambrogioni, Y. Güçlütürk, and M. A. J. van Gerven, “Generative adversarial networks for reconstructing natural images from brain activity,” *NeuroImage*, vol. 181, pp. 775–785, Nov. 2018.