# Temporal Information Guided Generative Adversarial Networks for Stimuli Image Reconstruction from Human Brain Activities

Shuo Huang, Liang Sun, Muhammad Yousefnezhad, Meiling Wang, and Daoqiang Zhang[†]

***Abstract*—Understanding how the human brain work has attracted increasing attention in both fields of neuroscience and machine learning. Previous studies use autoencoder and generative adversarial networks (GAN) to improve the quality of stimuli image reconstruction from functional Magnetic Resonance Imaging (fMRI) data. However, these methods mainly focus on acquiring relevant features between two different modalities of data, i.e., stimuli images and fMRI, while ignoring the temporal information of fMRI data, thus leading to sub-optimal performance. To address this issue, in this paper, we propose a temporal information guided GAN (*TIGAN*) to reconstruct visual stimuli from human brain activities. Specifically, the proposed method consists of three key components, including 1) an image encoder for mapping the stimuli images into latent space, 2) a Long Short-Term Memory (LSTM) generator for fMRI feature mapping, which is used to capture temporal information in fMRI data, and 3) a discriminator for image reconstruction, which is used to make the reconstructed image more similar to the original image. In addition, to better measure the relationship of two different modalities of data (i.e., fMRI and natural images), we leverage a pairwise ranking loss to rank the stimuli images and fMRI to ensure strongly associated pairs at the top and weakly related ones at the bottom. Experimental results on real-world datasets suggest that the proposed TIGAN achieves better performance in comparison with several state-of-the-art image reconstruction approaches.**

***Index Terms*—Stimuli image reconstruction, functional Magnetic Resonance Imaging, long-short term memory, generative adversarial networks**

## I. Introduction

How to understand human brain has been one of the most significant problems in the field of neuroscience in the past for a long time [1]–[4]. To this end, the topic called human brain encoding and decoding is proposed, where the encoding part embeds information into neural activities, while the decoding part extracts information from neural activities [5]–[7]. Functional Magnetic Resonance Imaging (fMRI) is one of the most popular tools for studying the human brain, using blood oxygen level dependence (BOLD) signals as a proxy for neural activity visualization. The main idea of the human brain encoding and decoding is to learn cognitive states by measuring neural activities [8]–[10].

Compared with cognitive state classification tasks, reconstruction of visual images can provide more details to understand human minds, even though there are still great challenges in obtaining the details of the stimuli images. Many studies are developed to explore the stimuli image reconstruction. As an early exploratory study, Thirion et al. [11] use rotating Gabors to reconstruct dot patterns from stimuli and imagery. They infer the visual content of real or imaginary scenes from the brain activity patterns that they elicit via well-known retinotopy of the visual cortex. Moreover, in [12], Miyawaki et al. firstly ask subjects to watch flashing checkerboard images as visual stimuli and record the evoked BOLD signal responses of these stimuli in the early visual cortex (V1/V2/V3). Then, they build a multi-scale local image decoder model for visual stimuli reconstruction.

In recent years, the development of deep neural networks (DNNs) technology revolutionizes many fields, e.g., image classification [13], [14], speech recognition [15], [16] and medical image segmentation [17], [18]. Further, several DNN-based methods have been proposed for decoding the cognitive states in human brains. For instance, some studies use the outputs of DNN to reveal the neural activities in the human visual cortex [19]–[22]. However, there are still some challenges for stimuli image reconstruction from human brain activity with fMRI data. In particular, 1) fMRI data is usually high-dimensional with a lot of complex noises, which interfere with the mining of real brain activity and influence the reconstruction results; 2) the pairwise samples are treated as time point samples, which ignores the temporal information of the visual task; 3) the limited mapping between the stimuli images and the evoked brain activity patterns, which fails to assess the correlation between two cross-modal data accurately.

To address these issues, in this paper, we propose a novel visual stimuli reconstruction method called temporal information guided GAN (TIGAN) to reconstruct stimuli images from human brain activities. In order to effectively utilize temporal information provided by fMRI data, we use the LSTM network to process fMRI data to obtain the context correlation information. Specifically, there are three key components in our method. The first part is the stimuli image autoencoder, which is used to map the stimuli images to a latent space through a deep neural network. The second part is an LSTM network,

S. Huang, L. Sun, M. Yousefnezhad, M. Wang and D. Zhang are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing 211106, China.

[†]Corresponding author: D. Zhang (dqzhang@nuaa.edu.cn).

This work was supported by the National Natural Science Foundation of China (Nos. 61876082, 61861130366, 61732006, and 62006115), the National Key R&D Program of China Grant (Nos. 2018YFC2001600, 2018YFC2001602), the China Postdoctoral Science Foundation (No. 2021T140323), the Jiangsu Postdoctoral Science Foundation (No. 2021K419C), and also by the CAAI-Huawei MindSpore Open Fund.

used for fMRI feature mapping to extract temporal information from fMRI. The third part is the discriminator for stimuli image generation, which generates the images as similar as the original inputs. Furthermore, we employ the pairwise ranking loss [23] to encourage the similarity of ground truth image-fMRI pairs to be greater than that of all other negative ones.

The major contributions of this paper are listed as follows:

- We propose a novel method to reconstruct the stimuli images from the evoked fMRI data. A TIGAN method is proposed to capture the temporal information in fMRI data via the LSTM network and complete the task of stimuli image reconstruction through GAN architecture.
- We introduce a pairwise ranking loss to measure the relationship between the stimuli images and fMRI data. This loss function ranks the stimuli images and fMRI that ensure strongly associated (corresponding) is at the top and weakly correlated at the bottom.
- We perform our method on two datasets for reconstructing natural images and handwritten digits. The experimental results show that our method achieves the best visual stimuli reconstruction from brain activity patterns compared with the state-of-the-art methods.

## II. Related Works

### A. Cross-Modal Reconstruction

As the coming of the era of big data, different modalities of data such as texts, images (i.e., natural images, medical images, satellite images, etc), and videos are growing at an unprecedented rate [24]–[26]. Such multi-modal data exhibit heterogeneous properties, making it difficult for users to search for information of interest effectively.

In recent years, a large number of studies have focused on bridging the heterogeneity between different modalities of data [23], [24], [27]. Some studies based on traditional machine learning methods are proposed for the reconstruction of cross-modal data [28], [29]. These methods aim to obtain a better reconstruction effect by fitting different modalities of data from the point of view of data distribution. However, there are still some challenges to estimate the distribution of one modality of data by the other due to the heterogeneity of different data. For stimuli image reconstruction tasks, the data distributions of fMRI scans and natural images are greatly different, so it is difficult to obtain fine reconstruction results.

In the last decade, deep learning is attracting more and more attention due to its powerful performance. At the same time, some cross-modal data reconstruction methods based on deep learning have been proposed, and better results have been obtained [30]–[32]. For example, in [33], the authors propose a cross-modal feature embedding framework, CNN and Skip-Gram are used in the framework to extract the features from different modalities of data. Further, the extracted feature representations are associated with a structured objective in which the distance between the matched pair of two different modalities of data is smaller than that between the mismatched pair. Similarly, a framework that uses a Gated Recurrent Unit (GRU) as a decoder is proposed in [34]. In this work, two different modalities of data, i.e., images and sentences, are mapped onto a common space to measure the modal difference. However, these efforts ignore the structural information in the visual scene, making the results less satisfactory. As one of the most popular deep learning methods, the generative adversarial networks (GANs) make it easier and more powerful for models to learn to distinguish between different feature representations. Some models based on GAN architecture are proposed to do cross-modal reconstruction. For example, a framework of cross-modal generative adversarial networks (CM-GANs) is proposed in [35]. For bridging the heterogeneity gap, CM-GANs learn a common feature representation of two modalities of data. Furthermore, in [36], the authors use the most difficult negative samples to replace the sum violations in the negative samples. In this way, they successfully overcome the limitations of [34].

### B. Visual Image Reconstruction

*1) Bayesian-based Linear Reconstruction:* Inspired by [12], several reconstruction models based on Bayesian framework are proposed to explore the correlations among fMRI voxels that can naturally reflect the characteristics of corresponding visual stimuli. For example, in [37], the authors introduce a Bayesian framework by using the structural and semantic features of encoding brain activity to accurately reflect the spatial structure and semantic categories of the objects contained in the observed natural image. However, it is always time-consuming to acquire the fMRI data. Therefore, it is difficult to reconstruct continuous data. To address this issue, a Bayesian decoding framework is proposed in [38] to reconstruct movies from the evoked BOLD signals. They propose a motion-energy encoding model that largely overcomes the limitation of tardiness of BOLD signals measured via fMRI. However, these methods neglect to mine the relationship between the images and the evoked fMRI data. Fujiwara et al. [39] develop a Bayesian Canonical Correlation Analysis (BCCA) model to automatically learn image bases, each module is modeled by a latent variable that associates with a set of pixels in a visual image. CCA is used to construct an invertible mapping based on the Bayesian model. Zhan et al. [40] propose a reconstruction method based on support vector machine (SVM) and Bayesian classifier followed by independent component analysis (ICA) to improve the efficiency of feature extraction and reconstruction performance. Du et al. [41] use Bayesian inference to derive missing latent variables, and effectively reconstruct handwritten digits with $10\times10$ binary images. Their joint generative model of external stimuli and brain activities can not only extract non-linear features in the visual image, but also capture the correlation among voxel activities recorded by fMRI.

*2) Non-linear Reconstruction:* In the last decade, an exhilarating achievement has been made in the field of stimuli image reconstruction based on deep learning methods. Several methods based on VAE are proposed to fit the distribution of stimuli images in the mapping space so that the reconstructed image is as analogous as possible to the original one [42]. For example, in [20], the authors propose a deep generative multiview model (DGMM) for reconstructing the perceive images from brain

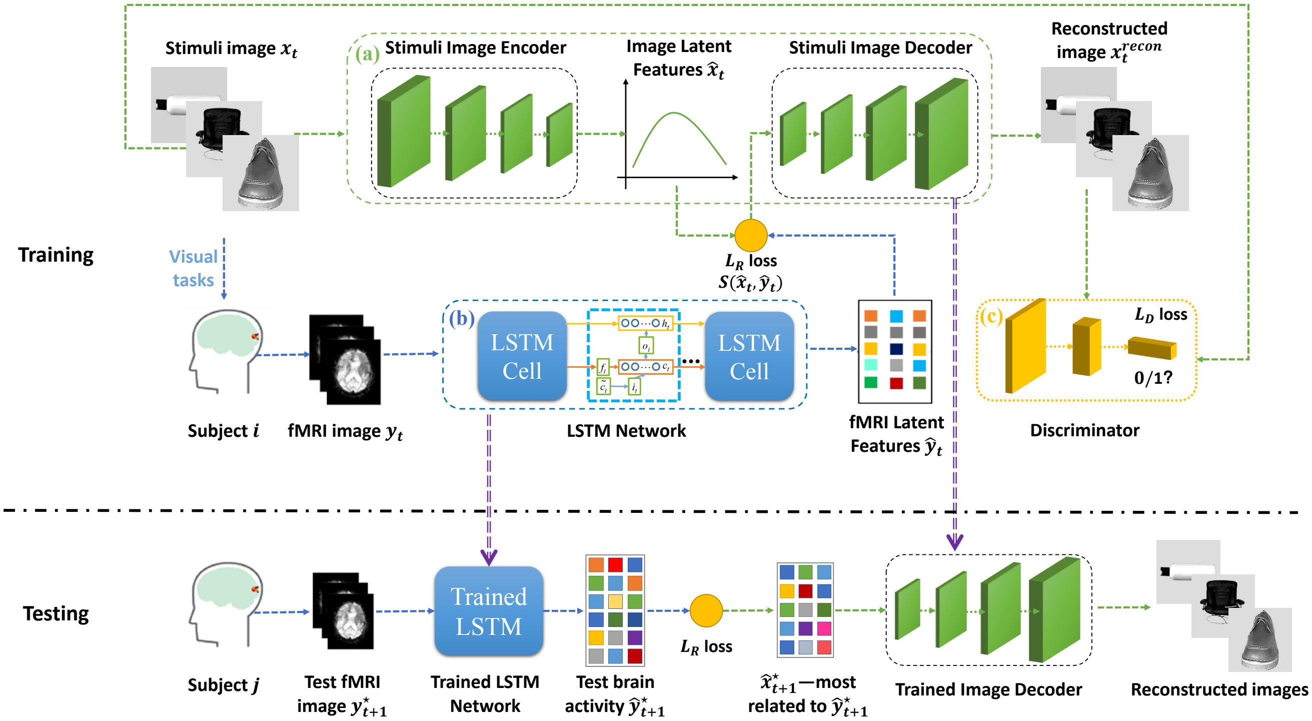
Stimuli image $x_t$
(a) Stimuli Image Encoder
Image Latent Features $\hat{x}_t$
Stimuli Image Decoder
Reconstructed image $x_t^{recon}$
Training
Visual tasks
$L_R$ loss
$S(\hat{x}_t, \hat{y}_t)$
(b)
LSTM Cell
LSTM Cell
(c)
$L_D$ loss
0/1?
Subject $i$
fMRI image $y_t$
LSTM Network
fMRI Latent Features $\hat{y}_t$
Discriminator
Testing
Trained LSTM
$L_R$ loss
Subject $j$
Test fMRI image $y_{t+1}^{\star}$
Trained LSTM Network
Test brain activity $\hat{y}_{t+1}^{\star}$
$\hat{x}_{t+1}^{\star}$—most related to $\hat{y}_{t+1}^{\star}$
Trained Image Decoder
Reconstructed images

Fig. 1. The schematic diagram of the proposed temporal information guided GAN method. **Training:** Three subnetworks are included in the proposed model, i.e., (a) an image autoencoder for mapping the stimuli images into latent space, which is in the dotted green area; (b) a LSTM generator for learning the temporal information of the fMRI, which is in the dotted blue area; and (c) a discriminator for image reconstruction, which is in the dotted yellow area. **Testing:** By using LSTM, the testing brain activities are encoded to $\hat{y}_{t+1}^{\star}$. Via the trained ranking loss, we can obtain $\hat{x}_{t+1}^{\star}$ which is most related to $\hat{y}_{t+1}^{\star}$. Given $\hat{x}_{t+1}^{\star}$, we can reconstruct the visual image through the trained stimuli image decoder.

fMRI activities. The DGMM can be viewed as a nonlinear extension of the linear BCCA. More recently, in [21], they train the DNN model with fMRI data and the corresponding stimuli images to build an end-to-end reconstruction model. The results show that the end-to-end model can learn a direct mapping between brain activities and visual stimulus.

Meanwhile, some GAN-based visual stimuli reconstruction models have been proposed and greatly improve the precision of the reconstruction results [19], [43], [44]. For instance, in [43], they train generative adversarial networks to learn a generative model of images that is conditioned on measurements of brain activity. Furthermore, in [44], the authors expand on the idea of using adversarial training for reconstruction but explore the capabilities of reconstructing arbitrary natural images via GANs. They train a deep convolutional generative adversarial network (DCGAN) separately on large image data sets and let it learn the latent space in an unsupervised manner. Recently, with the rapid development of GAN technology, more and more GAN-based methods are proposed [45]–[47]. Among them, Du et al. [45] use a hierarchically structured framework for neural decoding. And multi-task transfer learning of DNN representations and a matrix-variate Gaussian prior are used in their framework.

The reconstruction models based on the Bayesian framework aim to find the relationship between the visual stimuli and the corresponding fMRI signals and establish a linear mapping between them to achieve the task of image reconstruction. However, the linear mapping usually cannot truly reflect the relationship between the two cross-modal data, and the reconstruction results obtained are often coarse-grained, resulting in difficulty to describe the details of the images. Models based on deep networks can implement nonlinear transformations, greatly improve the accuracy of image reconstruction, and describe images in fine granularity.

However, most of the existing deep learning methods based on pairwise samples neglect the temporal information contained in the fMRI data. Hence, in this paper, we will propose a novel method that not only obtains high-accuracy image reconstruction results through the GAN architecture but also takes the temporal information into account by using the LSTM network to describe more complementary temporal information for the reconstruction task. We also introduce a pairwise ranking loss to measure the relationship between the stimuli images and fMRI data, which ensures the correspondence between the two cross-modal data.

## III. Proposed Method

### A. Notations

Let $N$ be the number of images which we used in the visual tasks, and let $D$ denotes the dimensions of stimuli images. We let $X = \{x_{pq}\} \in \mathbb{R}^{N \times D}, p = 1 : N, q = 1 : D$ denotes the stimuli images. At the same time, the preprocessed fMRI scans for $S$ subjects is denoted by $Y = \{y_{mn}\} \in \mathbb{R}^{T_f \times V}, m = 1 : T_f, n = 1 : V$, where $T_f$ is the number of time points in units of Repetition Time (TR), $V$ is the number of voxels, and $y_{mn}$ denotes the functional activity for the subject in the $m$-th time point and the $n$-th voxel. As proposed method is a cross-modal data reconstruction task, the samples are pairwise, which is saying that the number of the samples is $T$, and $T = N = T_f$. Here, for convenience, we let $(x_t, y_t)$ be a pairwise sample at time point $t, t = 1, 2, \ldots, T$.

### B. Temporal Information Guided GAN

We develop a TIGAN method for modeling the relationship between the stimuli images and the evoked brain activities. The proposed method generates two different modalities of data onto a common latent space by two specific generative networks and reconstructs the stimuli images via a discriminative network. The schematic diagram of proposed TIGAN is shown in Fig.1. There are three subnetworks in the proposed model, i.e., 1) an image autoencoder for mapping the stimuli images into latent space, 2) a LSTM generator for learning the temporal information in the fMRI data, and 3) a discriminator for image reconstruction.

*1) Stimuli Image Autoencoder:* Due to only a small number of samples can be used to train the proposed network for stimuli image reconstruction, we refer the pretrain strategy in [44] to pretrain an autoencoder to improve the model performance. Then, the pretrained encoder network is employed to map the stimuli images onto a latent representation space. Herein, the image encoder network maps the features of visual stimuli images onto the latent space $z^i$, where the latent feature $\hat{x_t} = E_\theta(x_t)$. Here $E(\cdot)$ is an encoder function, $\theta$ is the parameters in the encoder. While the decoder network reconstructs the image $x_t^{recon} = G_\phi(\hat{x_t})$ by using the nonlinear function $G(\cdot)$, where $\phi$ is the parameters of the decoder network. The loss function of the autoencoder can be defined as

$$\min_{\theta,\phi} \frac{1}{T} \sum_{t=1}^{T} \| x_t - G_\phi(E_\theta(x_t)) \|_F^2 . \tag{1}$$

*2) LSTM Network for fMRI Feature Mapping:* LSTM network consists of repeated cells that receive input from the previous cell as well as the data input $y_t$ for the current timestep $t$. Each LSTM cell contains a cell state $c_t$ and a hidden state $h_t$ , which are modulated by four neural network layers that control the flow of information into and out of cell memory. The equations governing the LSTM are defined as

$$\begin{aligned}
i_t &= \sigma(W_i y_t + U_i h_{t-1} + b_i),\\
f_t &= \sigma(W_f y_t + U_f h_{t-1} + b_f),\\
\widetilde{c_t} &= tanh(W_c y_t + U_c h_{t-1} + b_c),\\
c_t &= i_t * \widetilde{c_t} + f_t * c_{t-1},\\
o_t &= \sigma(W_o y_t + U_o h_{t-1} + b_o),\\
h_t &= o_t * tanh(c_t).
\end{aligned} \tag{2}$$

The fMRI generator produces an output image $\hat{y_t}$ given the corresponding brain activities in sequential order $y_t, t = 1, 2, \ldots, T$. Here, the generated image $\hat{y_t}$ is as analogous as possible to the reconstructed image in the next step $y_{\hat{t+1}}$. Therefore, the generator should be a sequential LSTM model, which produces the sequentially next image, $\hat{y_t} = \mathcal{L}(y_1, y_2, \cdots, y_t)$, $t = 1, 2, \ldots, T$. The LSTM network maps the fMRI signals into the fMRI latent space $z^f$, where the latent feature $\hat{y_t} = \mathcal{L}(y_t)$, $t = 1, 2, \ldots, T$. Here, $\mathcal{L}(\cdot)$ defines the LSTM network mapping.

*3) Discriminator for Stimuli Image Generation:* As a reconstruction method, the goal of TIGAN is to make the reconstructed image as similar to the original as possible. So the discriminator in the model takes a reconstructed or original image as input, then a binary decision is made to decide whether the input is real or fake, which will result in the output 1 or 0, respectively. Herein, we let the original image $x_t$ as the real sample and the generated image $x_t^{recon}$ as the fake one at the same time. The proposed TIGAN method is essentially a sequence generation model which receives the context images as the preceding time-steps and the real image in the generator as the last time-step. The final hidden state is mapped onto a sigmoid predicting whether it is a real or fake image.

We involve two loss components to compute the loss between the generated image $x_t^{recon}$ and the original image $x_t$ on the basis of features from the trained deep neural networks. The first component is feature reconstruction loss $\mathcal{L}_f$, which determines whether features are activated above a threshold at all. The feature reconstruction loss is obtained via mean absolute error (MAE), which is calculated between the generated image $x_t^{recon}$ and the original image $x_t$. The feature reconstruction loss $\mathcal{L}_f$ can be determined as

$$\min \sum_{t=1}^{T} \sum_{q=1}^{D} |(x_t)^{(q)} - (x_t^{recon})^{(q)}|, \tag{3}$$

where $D$ denotes the dimensions of stimuli images and $q$ means the $q$-th pixel of the image $(q = 1, 2, \ldots, D)$.

The second component of the losses is the discriminator loss $\mathcal{L}_d$. The discriminator discriminates the real sample. Here, to make the discriminative result close to 1, we let the generated image $x_t^{recon}$ close to the real image $x_t$ to fool the discriminator. The discriminator loss $\mathcal{L}_d$ can be defined as

$$\begin{aligned}
\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log(D(x))] + \\
E_{z\sim p_z(z)}[\log(D(1 - D(G(z))))].
\end{aligned} \tag{4}$$

The hybrid loss function $\mathcal{L}_D$ combine the two loss components as

$$\mathcal{L}_D = \frac{1}{T}(\mathcal{L}_f + \mathcal{L}_d). \tag{5}$$

### C. Ranking Loss for Cross-Modal Data Fusion

One of the most significant challenges in the field of stimuli image reconstruction is how to model the relationship between the stimuli images and the evoked fMRI scans. Inspired by [23], we develop the pairwise ranking loss from the image-textual reveral to visual stimuli reconstruction field for measuring the relationship between two different modalities of data, i.e., images and brain activities. The schematic diagram of pairwise ranking loss in visual stimuli reconstruction is shown in Fig.2. Here, we denote $(\hat{x}_t, \hat{y}_t)$ as the pairwise features at time point $t$, which generated from stimuli image encoder

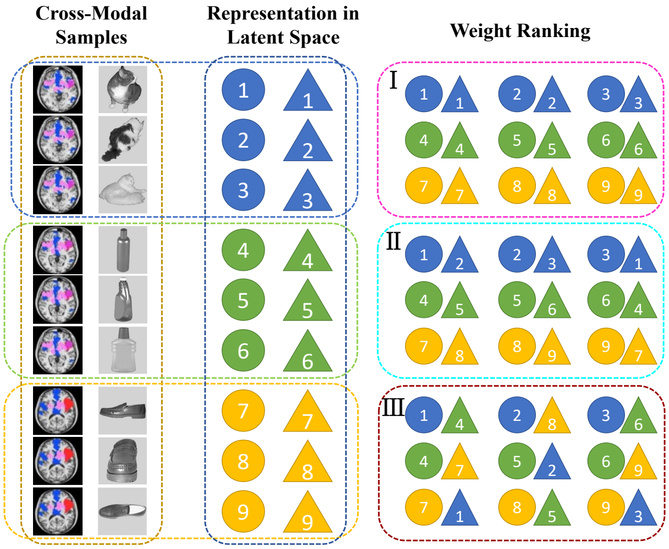

Fig. 2. The schematic diagram of pairwise ranking loss in visual stimuli reconstruction. Circles represent the fMRI data and the triangle represent the stimuli images. Different colors means different categories in the dataset. Here, blue="cats", green="bottles", yellow="shoes".

and LSTM network, respectively. We further denote the non-corresponding samples by using $\hat{x}'_t$ and $\hat{y}'_t$, where $\hat{x}'_t$ goes over stimuli images independent of $\hat{y}_t$, and $\hat{y}'_t$ goes over brain activities not evoked by $\hat{x}_t$. The objective function ensures that the groundtruth image-fMRI pairs at the top and weakly related ones at the bottom. More specifically, for the samples which are assortative, they will have the maximum value of weight. For the pair which is the same category but not the *correct one*, the weight will be medium, for the pair in which the image belongs to the misclassification category, we assume that this pair is an unrelated sample, and give this sample the lowest weight. Therefore, we optimize the ranking loss

$$\mathcal{L}_R = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_{Rank}(\hat{x}_t, \hat{y}_t), \tag{6}$$

where ranking loss $\mathcal{L}_{Rank}$ of the single pairwise sample $(\hat{x}_t, \hat{y}_t)$ is defined as

$$\begin{aligned}\mathcal{L}_{Rank} = &\sum_{\hat{x}'_t}[\alpha - s(\hat{x}_t, \hat{y}_t) + s(\hat{x}'_t, \hat{y}_t)]_+ + \\ &\sum_{\hat{y}'_t}[\alpha - s(\hat{x}_t, \hat{y}_t) + s(\hat{x}_t, \hat{y}'_t)]_+,\end{aligned} \tag{7}$$

where $\alpha$ is a margin, $s(\hat{x}_t, \hat{y}_t) = - \parallel (\max(0, \hat{x}_t - \hat{y}_t)) \parallel^2$ is the order-violation penalty used to measure the similarity between the stimuli images and the evoked brain activities. Futher, $[x]_+$ represents $\max(x, 0)$.

The complete loss function is then given as

$$\mathcal{L}_{loss} = \lambda_D \mathcal{L}_D + \lambda_R \mathcal{L}_R, \tag{8}$$

where $\lambda_D, \lambda_R$ are hyper-parameters to balance the effects of the two loss functions. We randomly choose all the parameters from $\{0.05, 0.1, 0.5, 1, 5, 10\}$.

## D. Implementation Details

*1) Stimuli Image Encoder:* In this work, we refer to [44], which based the model on a publicly available framework and implementation[1]. The encoder network consists of one linear and four convolutional layers, each followed by batch normalization and ReLU activation functions. The linear layer takes $z$ and maps it to the first deconvolutional layer that expects 512 feature channels. The generator then maps to 256, 128, 64 and 1 feature channels across the convolutional layers. For the first two convolutional layers, kernel sizes are 2×2 and stride is 1. For the last two convolutional layers, kernel sizes are 4×4 and stride is 2.

*2) LSTM Network for fMRI Feature Mapping:* First, the LSTM network mapped the fMRI activity patterns to the first fully connected layer. Then, the first fully connected output as input to a single-layer LSTM module. Finally, the second fully connected layer mapped the output of the LSTM module to latent features. The two fully connected layers in the LSTM network use the ReLU activation function. The Adam optimization algorithm [48] was used to optimize the LSTM network. When there is only a one-time point in our multi-time point data fusion, we set the time steps of LSTM to 9, which equals the number of time-points in one run experiment of a label.

*3) Discriminator for Stimuli Image Generation:* The decoder network consists of 4 deconvolutional layers, followed by batch normalization and ReLU activations. Except in the initial layer (which had 3×3 kernels) all layers use kernel sizes of 4×4 and a stride of 2. The layers map from 1 to 32, then 64, 128 up to 256 feature channels, and are followed by a linear layer mapping all final activations to a single value reflecting the discriminator decision.

In the training stage, we input the visual stimuli images $x_t, t = 1, 2, \ldots, T$ to the image encoder and the fMRI data $y_t$ to the LSTM generator. Then the image generator maps the images into the image latent space $z^i$ and the output of the LSTM generator is the fMRI latent space $z^f$. We combine the two cross-modal latent space into a common space via the ranking loss $\mathcal{L}_R$, to strengthen the relationship between two cross-modal data. The trained image latent features $\hat{x}_t$ is sent to the image decoder as the input and the output is the generated image $x_{recon}$.

In the test stage, we only use the fMRI data in the test set as the input. By using LSTM, the testing brain activities $y^{\star}_{t+1}$ are encoded to $\hat{y}^{\star}_{t+1}$. Via the trained ranking loss, the image features $\hat{x}^{\star}_{t+1}$ which is most related to $\hat{y}^{\star}_{t+1}$ could be determined based on the fMRI features. Then the image features are utilized as the input to the trained image decoder. Given $\hat{x}^{\star}_{t+1}$, we can reconstruct the visual image through the trained stimuli image decoder. Herein, we use Adam [48] as the optimizer with the learning rate of 0.0005. The batch size of the DS105 is set as 64, and 16 for handwritten digits dataset.

[1] http://github.com/musyoku/improved-gan

TABLE I
PROPERTIES OF THE DATASETS USED IN THE EXPERIMENTS

| Datasets | Instances | Categories | Pixels | Voxels | Training |
|---|---|---|---|---|---|
| DS105 | 8712 | 7 | 100×100 | 2294 | 3465 |
| Handwritten Digits | 100 | 2 | 28×28 | 3092 | 90 |

## IV. EXPERIMENTAL RESULTS

### A. Datasets

In this paper, we employ two publicly available datasets to validate the proposed method, including, a) Open NEURO[2] dataset, and b) Handwritten digits dataset. More details can be found as follows.

*a) Open NEURO dataset:* We utilize a publicly available datasets shared by Open NEURO for running empirical studies. In this paper, we select the dataset numbered DS105 [1]. This task is a one-time retest task. DS105 consists of eight categories stimuli images, which are face, house, cat, bottle, scissors, shoes, chair, and the meaningless pattern. The images are all grayscale with the resolutions of 400×400. In order to reduce the feature dimension and improve the computational efficiency of the model, we down-sampling the stimuli image into 100×100 by means of descending sampling. And also, the edge gradation that does not have practical meaning in the grayscale image is converted to 0. The dataset was preprocessed by the software easy fMRI[3], i.e., slice timing, smoothing, normalization, and anatomical alignment.

In DS105 [1], 6 subjects were stimulated with grayscale images in 8 categories, and each subject underwent 12 runs of experiments. Among them, Subject #5 miss one run of data record, with only 11 runs of data. After the preprocess, each subject has 1452 time points with 2294 voxel-level features at each time point. Herein, we exclude the meaningless label in the dataset to verify the reconstruction effect of the real stimuli images, and the samples of seven categories are reserved. Two brain regions of ventral temporal cortex were specialized for representing specific categories, which are fusiform face area (FFA) and the parahippocampal place area (PPA), respectively.

According to the training and testing strategies of machine learning, the data of the test samples cannot appear in any form in the training process. Therefore, we believe that if we use the same subject neural responses for training and testing phase, new time points that are unseen during training are still a potential risk of data leakage in the testing phase. In this paper, we use a leave-one-out cross validation strategy to adjust the parameters and evaluate the effectiveness of the method we propose. In each phase, data from five subjects are used for training, while data from one subject is used during the test stage. In the training stage, we use data that 5(subjects) × 11(runs) × 9(images) × 7(categories) = 3465 samples. And in the test phase, we use one subject data in one run. We then repeat the experiment six times and calculate the average results as the final results.

*b) Handwritten Digits dataset [49]:* This dataset contains 100 gray-scale images of handwritten digits (50 of digital "6" and the equal numbers of digital "9"). The image resolution is 28×28. The evoked fMRI data contain voxels from the V1, V2, and V3 areas. Similar to [20], 10-fold cross validation is performed (i.e., each category contained 45 training data and 5 testing data per experiment). In each fold, the training set consists of 90 time points of image-fMRI samples and the test set consists of 10 time points of image-fMRI samples. Table I shows the attributes and information for the datasets used in this paper.

### B. Compared Methods

The proposed method is compared with four well-known methods, including

*a) Bayesian canonical correlation analysis (BCCA) [39]:* BCCA is a multi-view generative model used for brain activity pattern analysis. However, as a kind of linear generative model, its linear architecture and spherical covariance assumption may influence the generation results.

*b) Deep canonically correlated autoencoder (DCCAE) [50]:* DCCAE, a DNN-based model combining CCA and autoencoder-based terms, consists of two basic autoencoders used for learning the deep representations from multi-view data. However, DCCAE ignores the interview reconstruction errors of multi-modal data.

*c) Deep generative multiview model (DGMM) [20]:* DGMM is a deep generative multi-view learning model for reconstructing the stimuli images from brain activities. It can be viewed as a nonlinear extension of the BCCA. However, DGMM does not take the temporal information of fMRI data into account.

*d) Deep convolutional generative adversarial network (DCGAN) [44]:* DCGAN uses generative adversarial networks for arbitrary image generation from stimuli images (handwritten characters or natural grayscale images). However, just like other methods, the temporal information of fMRI data is not taken into consideration in DCGAN.

### C. Evaluation Metrics

In our experiment, three evaluation metrics are used to measure the reconstruction performances of different methods, which are Euclidean distance (Euc_dis), Pearson's correlation coefficient (PCC) and mean squared error (MSE), respectively. Here, Euc_dis measures the distance between the original and reconstructed image. The smaller this value is, the closer the reconstructed image is to the original image in feature space. And the PCC value shows the correlation between the original and reconstructed images. The last but not the least, the reconstruction accuracy is measured by MSE, which calculates the pixel-level error between the original image and the reconstructed image. The smaller the error, the more similar the reconstructed image is to the real image.

[2] http://openneuro.org

[3] https://easyfmri.learningbymachine.com/

TABLE II

Quantitative performances of compared methods on the $DS$105 dataset. (↑: The higher the value is, the better performance the method get. ↓: The lower the value is, the better performance the method get.)

| Model | Euc_dis↓ | p-value | PCC↑ | p-value | MSE↓ | p-value |
|---|---|---|---|---|---|---|
| BCCA | 0.787±0.093 | 1.3839$e$-14 | 0.561±0.079 | 9.8467$e$-11 | 0.208±0.062 | 8.2238$e$-9 |
| DCCAE | 0.751±0.096 | 1.6552$e$-10 | 0.584±0.103 | 8.5767$e$-10 | 0.171±0.104 | 2.0229$e$-7 |
| DGMM | 0.652±0.082 | 2.4369$e$-7 | 0.636±0.096 | 2.8228$e$-7 | 0.124±0.069 | 3.4167$e$-4 |
| DCGAN | 0.641±0.089 | 7.9318$e$-5 | 0.651±0.096 | 8.0966$e$-6 | 0.116±0.074 | 0.0055 |
| TIGAN(Proposed) | **0.609±0.061** | —— | **0.689±0.063** | —— | **0.091±0.051** | —— |

TABLE III

Quantitative performances of compared methods on the handwritten digits dataset.

| Model | Euc_dis↓ | p-value | PCC↑ | p-value | MSE↓ | p-value |
|---|---|---|---|---|---|---|
| BCCA | 0.679±0.155 | 1.1709$e$-10 | 0.423±0.139 | 1.6853$e$-22 | 0.119±0.023 | 1.8554$e$-20 |
| DCCAE | 0.631±0.064 | 9.5486$e$-9 | 0.529±0.047 | 5.0496$e$-20 | 0.077±0.018 | 3.3630$e$-11 |
| DGMM | 0.585±0.061 | 0.0025 | 0.801±0.061 | 0.0291 | 0.037±0.019 | 0.004 |
| DCGAN | 0.581±0.055 | 0.0238 | 0.799±0.057 | 0.0163 | 0.038±0.022 | 0.0096 |
| TIGAN(Proposed) | **0.568±0.037** | —— | **0.812±0.059** | —— | **0.033±0.015** | —— |

### *D. Experimental Results*

*1) Quantitative Analysis:* Performances of compared methods on two datasets are reported in Table II and III. Table II shows the experimental results of dataset DS105, several observations can be drawn as follows. First, the proposed TIGAN obtains a considerably better performance compared with the other methods. Second, by comparing deep learning methods (i.e., DCCAE, DGMM, DCGAN and the proposed TIGAN method) with BCCA, a linear model for stimuli reconstruction, we can see that our method is always out-perform BCCA. These results show that our reconstruction method with deep network is better than linear model by extracting nonlinear features from visual images and fitting images. Third, compared with DCCAE, the proposed method shows significantly better performance. The possible reason for the improvements is that the temporal information can be provided by LSTM in the proposed method. Fourth, the performance of TIGAN is more moderate than DGMM on both of the two datasets. This may be caused by the performance gap between the deep network in DGMM and the generative discriminant model in TIGAN. Finally, compared with DCGAN, the ranking loss in our method plays an important role in mining the correlation between the stimuli images and the brain activity patterns.

For the handwritten digit dataset, the results are shown in Table III. The quantitative results on the three evaluation metrics are also at the best level. For the three compared methods of BCCA, DCCAE, and DGMM, we refer to the experimental settings in [20] and also refer to their experimental results on MSE. And for Euc_dis and PCC here, we obtained similar results to that on the DS105 dataset. The reason is as analyzed above. Compared with DCGAN, our method also takes better results because of the use of the LSTM network and the cross-modal ranking loss. In addition, p-values are also displayed in the tables to verify the significance of our experimental results.

*2) Qualitative Analysis:* The reconstructed results on two different datasets are shown in Fig.3-5, respectively. In each figure, the top row shows the presented visual images, while the following rows show the reconstructed results of all compared methods.

The reconstructing results of DS105 (category = ”bottles”) are shown in Fig.3. As illustrated, our method produces better reconstruction results than the compared methods. Fig.3 also indicates that the effect of our method is obviously better than other methods on the reconstruction of natural images. In particular, BCCA and DCCAE cannot provide acceptable performance in characterizing detailed contours, which may be related to their mapping capabilities. DGMM and DCGAN are better than the first two methods, but they are not as good as our method when describing image details, such as color.

The reconstructing results on handwritten digits dataset are shown in Fig.4. As is shown in Fig.4, the reconstructed digits are very similar to the original images. Compared with our method, the performances of BCCA and DCCAE are not acceptable. The complex noises often influence their reconstruction results and the results also lack of the basic features in the original images. Furthermore, the reconstruction results of DGMM and DCGAN are coarse too. Although their results are better than those of BCCA and DCCAE, some detailed information are lost in DGMM and DCGAN compared with our TIGAN method, because they did not take the temporal information into account.

Fig. shows the reconstruction results of proposed TIGAN method on seven different categories in DS105 dataset. As is shown in the figure, bottles and shoes get the best results as the images in both categories are simple and clear. Then, scissors and houses get the sub-optimal performance. The reason may be that although the images in these two categories have clear edges, there is a lot of detailed information inside, which is more difficult to capture than the first two categories. The other three categories of images do not perform as well as the previous results due to their complexity.

By comprehensively comparing the experimental results

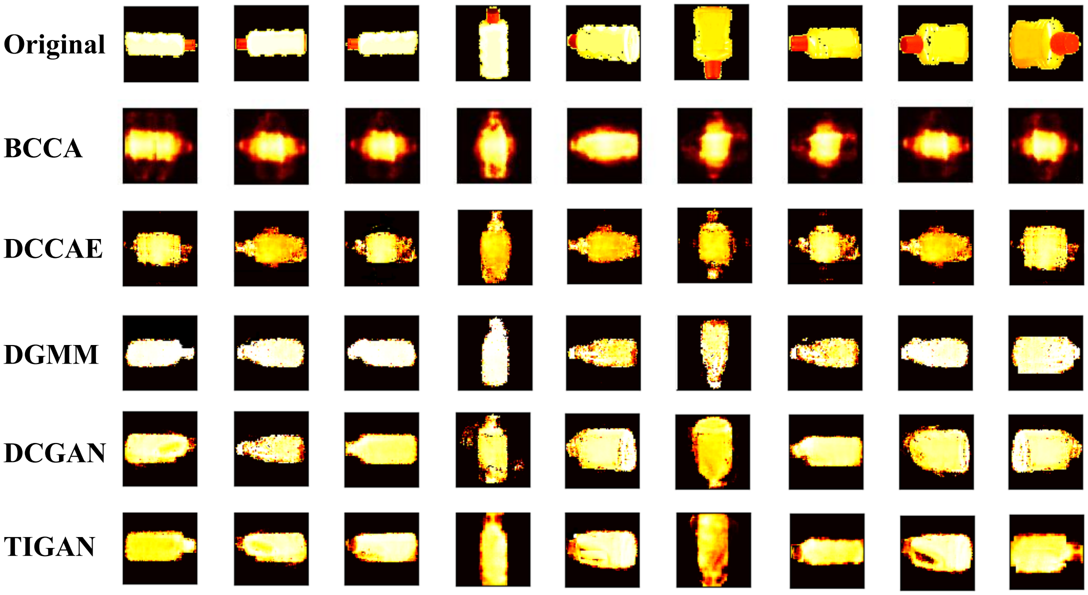

Fig. 3. Qualitative performances of compared methods on $DS105$ (category=bottles) dataset.

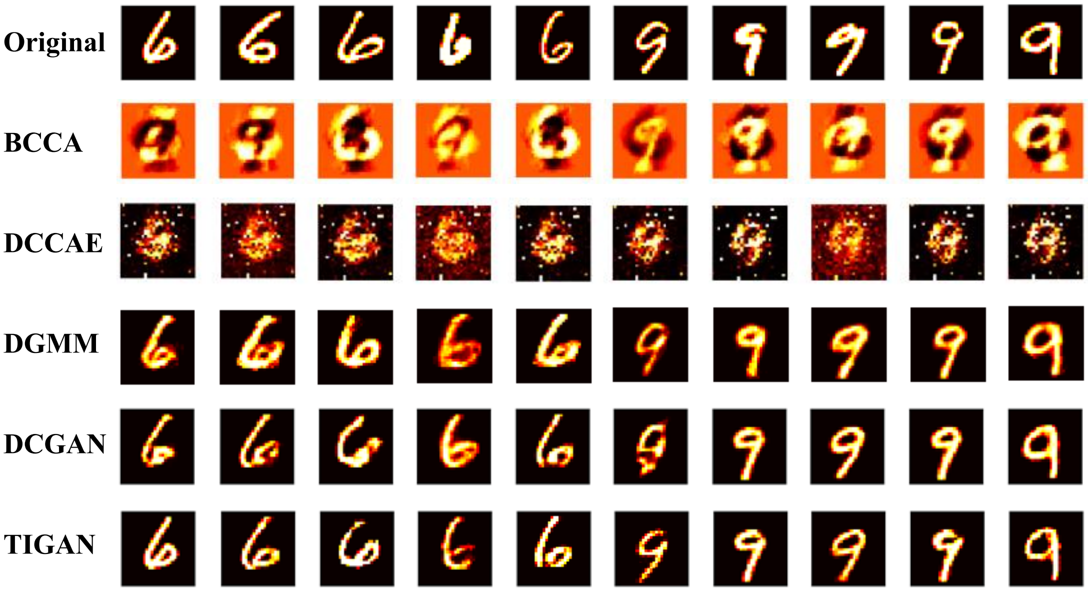

Fig. 4. Qualitative performances of compared methods on the handwritten digits dataset.

on the two datasets DS105 and handwritten digits, we can see that whether through quantitative analysis or qualitative analysis, the handwritten digits dataset has achieved better results than DS105. The reason may be mainly due to the complexity of the images. The images in DS105 are grayscale natural images, while the handwritten digits dataset is some handwritten digits, which is relatively simple. In addition, the images of the handwritten digits dataset have black bottoms and white numbers, which have clear edges and easy to be discriminated. However, grayscale images are more difficult to distinguish on the edges, and the pixel values of the target and background parts are closer.

## V. Discussion

In this section, firstly, we perform ablation study to evaluate the effectiveness of each component (i.e., LSTM network, ranking loss and discriminator loss respectively) in our method. Secondly, the influence of different error representations (MAE and MSE) is evaluated. Thirdly, we present the results of visual stimuli reconstruction on data from different subjects. Then, we evaluate the effects of regularization parameters in our model. In addition, the influence of different number of LSTM layers and the correlation of cross-modal pairwise samples are also taken into consideration. Finally, we present the limitations of this work as well as the possible future research direction.

TABLE IV
PERFORMANCES OF DIFFERENT CATEGORIES ON THE $DS105$ DATASET.

| Category | Evaluation | BCCA | DCCAE | DGMM | DCGAN | TIGAN |
|---|---|---|---|---|---|---|
| bottles | Euc_dis ↓ | 0.742±0.085 | 0.724±0.098 | 0.635±0.083 | 0.617±0.085 | **0.581±0.069** |
| | PCC ↑ | 0.591±0.081 | 0.615±0.097 | 0.668±0.102 | 0.675±0.101 | **0.719±0.062** |
| | MSE ↓ | 0.179±0.065 | 0.144±0.102 | 0.104±0.074 | **0.079±0.071** | 0.081±0.049 |
| shoes | Euc_dis ↓ | 0.762±0.102 | 0.733±0.106 | 0.627±0.081 | 0.623±0.094 | **0.591±0.062** |
| | PCC ↑ | 0.601±0.083 | 0.605±0.107 | 0.653±0.092 | 0.683±0.091 | **0.704±0.065** |
| | MSE ↓ | 0.186±0.068 | 0.151±0.104 | 0.109±0.069 | 0.112±0.075 | **0.079±0.053** |
| scissors | Euc_dis ↓ | 0.778±0.083 | 0.743±0.091 | 0.645±0.078 | 0.635±0.093 | **0.605±0.055** |
| | PCC ↑ | 0.575±0.077 | 0.581±0.111 | 0.639±0.097 | **0.701±0.097** | 0.698±0.072 |
| | MSE ↓ | 0.206±0.063 | 0.173±0.105 | 0.126±0.075 | 0.121±0.073 | **0.089±0.055** |
| houses | Euc_dis ↓ | 0.801±0.099 | 0.757±0.094 | 0.647±0.083 | 0.645±0.088 | **0.611±0.069** |
| | PCC ↑ | 0.559±0.071 | 0.578±0.106 | 0.631±0.104 | 0.632±0.094 | **0.688±0.062** |
| | MSE ↓ | 0.202±0.062 | 0.179±0.101 | 0.129±0.066 | 0.126±0.077 | **0.091±0.047** |
| faces | Euc_dis ↓ | 0.795±0.089 | 0.753±0.099 | 0.652±0.086 | 0.637±0.087 | **0.618±0.051** |
| | PCC ↑ | 0.547±0.089 | 0.587±0.103 | 0.635±0.092 | 0.630±0.097 | **0.681±0.055** |
| | MSE ↓ | 0.211±0.061 | 0.172±0.113 | 0.125±0.065 | 0.138±0.079 | **0.088±0.052** |
| chairs | Euc_dis ↓ | 0.806±0.103 | 0.767±0.095 | 0.684±0.082 | 0.664±0.088 | **0.621±0.062** |
| | PCC ↑ | 0.521±0.078 | 0.565±0.095 | 0.617±0.094 | 0.621±0.099 | **0.672±0.066** |
| | MSE ↓ | 0.232±0.056 | 0.193±0.106 | 0.134±0.064 | **0.098±0.069** | 0.102±0.048 |
| cats | Euc_dis ↓ | 0.824±0.091 | 0.779±0.089 | 0.675±0.080 | 0.667±0.089 | **0.635±0.059** |
| | PCC ↑ | 0.535±0.074 | 0.559±0.101 | 0.611±0.093 | 0.617±0.095 | **0.662±0.061** |
| | MSE ↓ | 0.239±0.061 | 0.188±0.098 | 0.142±0.071 | 0.136±0.075 | **0.109±0.052** |

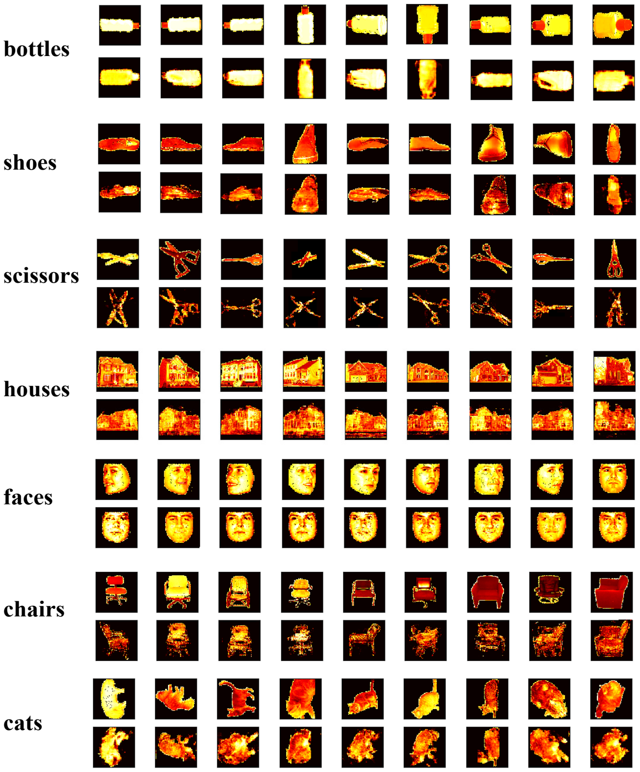

Fig. 5. Qualitative performances of TIGAN on all the seven categories in $DS105$ dataset.

## *A. Ablation Study*

As mentioned above, there are three key components in the proposed TIGAN method, i.e., 1) LSTM network used for mining the temporal information in fMRI data; 2) the ranking loss used for measuring the relationship between the stimuli images and fMRI data; and 3) the discriminator loss used for making the reconstructed images more similar to the original ones. In order to evaluate the contribution of different components to model performance, we conduct ablation study with detailed results to evaluate the effectiveness of each component in our method. Fig.6 shows the reconstruction performance (MSE) via different number of pixel features of images in DS105.

*1) With only one component:* Here, we use only one component to reconstruct the stimuli images and measure each component's contribution to the model. As can be seen in Fig.6(a), only LSTM means that we don't use ranking loss to emphasize the relationship between fMRI and images and there is also no discriminator loss to make the reconstructed image more similar to the original one. Only ranking loss refers that there are no LSTM and GAN architecture in the method. Only discriminator loss means that only DCGAN is used for the image reconstruction.

*2) With 2/3 of components:* Here, two of the three components are included to evaluate the impact on the model when a component is absent. Fig.6(b) shows the reconstruction performance. The LSTM block in our TIGAN method can balance the contributions of temporal information for the stimuli image reconstruction task. To study the influence of the LSTM block used in TIGAN method, we use a multi-layer perceptron (MLP) to replace the LSTM block in our method. Hence, the new method without the LSTM block. Meanwhile, in our proposed hybrid loss function, the discriminator loss is used to make the reconstructed image more similar to the original image, and the ranking loss is used to measure the

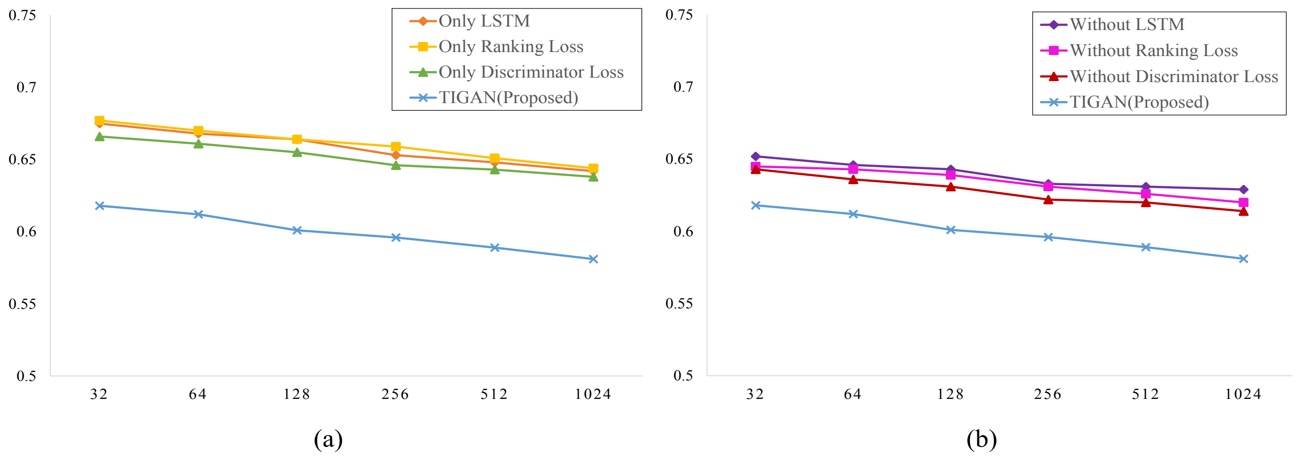

Fig. 6. Ablation study results on the $DS105$ (category = bottles) dataset. (a) Different performance that compared TIGAN with only one component. (b) Different performance that compared TIGAN with two components in the method.

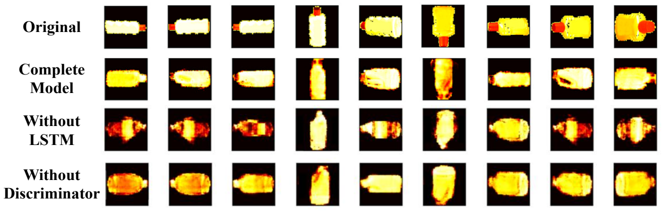

Fig. 7. Reconstruction results with/without LSTM or the discriminator on the $DS105$ (category = bottles) dataset.

relationship between two modalities of data (i.e., fMRI and images). To study the influence of each term in the hybrid loss, we train our proposed methods without the discriminator loss and ranking loss, respectively.

We can mainly observe two potential trends and conclusions through the figure. First, compared the three main component in our method (i.e., LSTM network, ranking loss and discriminator loss), the LSTM gets better MSE performance. Second, the proposed method which combine all the three components gets the best result. This indicates that every component of the method contributes to improving the reconstruction results.

Besides the quantitative analysis above, we also give the visualization results with or without each component in our model. And the visualization results are shown in Fig.7. The first row is the original input images. The second row is the reconstructed results of TIGAN. The third row is the reconstructed results without LSTM and the last row is the results without discriminator ($\lambda_D = 0$). As shown in Fig.7, we can easily find that in the case of network missing, although the reconstruction results can still be obtained, the accuracy is not as good as that of the complete model. In the visualization results, in the absence of LSTM, the reconstruction of the object's edge shape is unstable, which may be due to the lack of the structural semantics contained in the temporal information. In the absence of the discriminator, the reconstruction results are not precise enough, which may be due to the lack of the game process of the discriminator in the image reconstruction.

In addition, since both the proposed TIGAN and the compared method DCGAN [44] use GAN architecture as the generative model, one of the main differences between the two methods is the utilization of ranking loss. Thus, we conduct experiment performed on DS105 dataset (category=bottles) to add the ranking loss to DCGAN for comparing the performance of the two methods. The experimental results are reported in Table V.

TABLE V
THE EXPERIMENTAL RESULTS OF DIFFERENT COMPARE METHODS WITH OR WITHOUT THE RANKING LOSS. (METHOD-$L_R$ MEANS THAT $L_R$ LOSS IS USED IN THE MODEL.)

| Methods | Euc_dis↓ | PCC↑ | MSE↓ |
|---|---|---|---|
| DCGAN | 0.617±0.082 | 0.675±0.101 | 0.079±0.071 |
| DCGAN-$L_R$ | 0.608±0.063 | 0.686±0.082 | **0.074±0.077** |
| TIGAN | 0.588±0.099 | 0.709±0.057 | 0.088±0.091 |
| TIGAN-$L_R$ | **0.581±0.069** | **0.719±0.062** | 0.081±0.049 |

As we can see from the table, the ranking loss leads to the improvement of the methods' performance. Note that after adding the ranking loss to the compare methods, our proposed TIGAN method still achieved good results. Compared with DCGAN, after adding the ranking loss, the most obvious difference between DCGAN and TIGAN is the LSTM module which is used for mining the temporal information. And the leverage of the fMRI temporal information led the proposed TIGAN method to obtain better performance.

### *B. Influence of MAE and MSE as Different Error Representations*

In our proposed method, mean absolute error (MAE) was used for calculating the pixel reconstruction error. And MSE was used as an evaluation metric to measure the model performance. For comparing the influence of these two different error representations, in this section, we conduct an experiment, and the samples with categories of bottles and

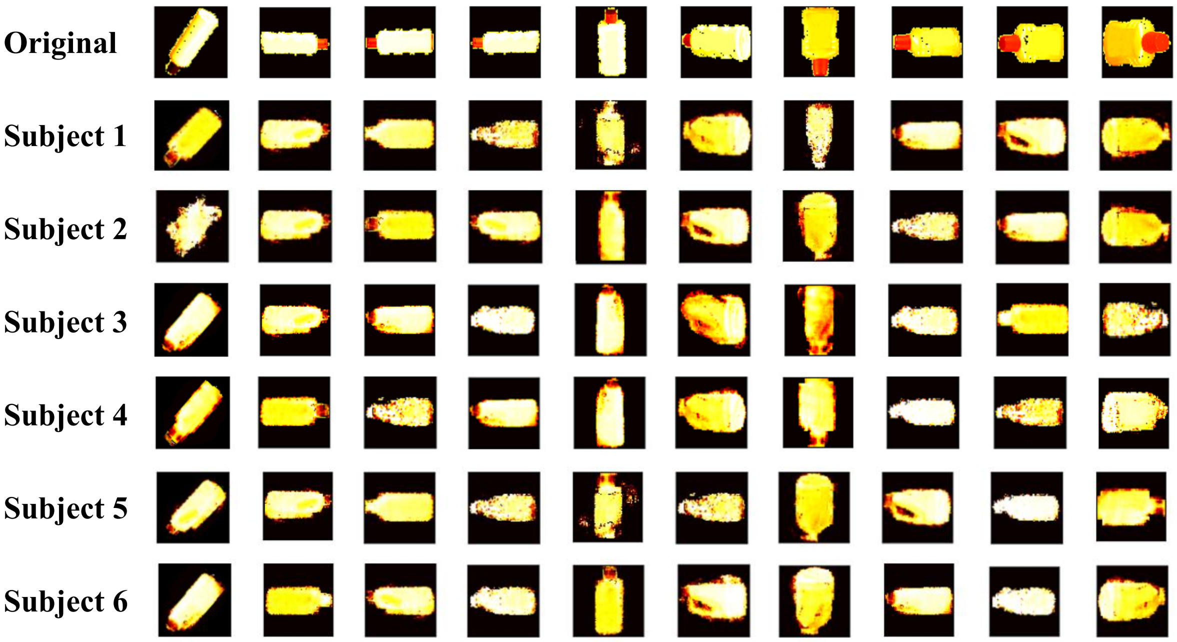

Fig. 8. Cross-subject reconstruction results for all six subjects in DS105 (category=bottles).

shoes in the dataset DS105 are used. In the experiment, MAE loss was replaced by MSE loss to measure the difference in pixel level. The reconstruction results are shown in Table VI. As can be seen in the table, with the MAE loss, Euc_dis and PCC get better performance. With the MSE loss, the same evaluation metric MSE gets better performance. The reason may be that the MSE results are trained to decline.

TABLE VI
THE EXPERIMENT RESULTS OF USING MAE OR MSE AS THE FEATURE RECONSTRUCTION LOSS.

| Categories | Methods | Euc_dis↓ | PCC↑ | MSE↓ |
|---|---|---|---|---|
| bottles | TIGAN-MSE | 0.593±0.082 | 0.696±0.104 | **0.076±0.056** |
|  | TIGAN-MAE | **0.581±0.069** | **0.719±0.062** | 0.081±0.049 |
| shoes | TIGAN-MSE | 0.606±0.077 | 0.693±0.042 | **0.077±0.088** |
|  | TIGAN-MAE | **0.591±0.062** | **0.704±0.065** | 0.079±0.053 |

The mean square error (MSE) is the most commonly used regression loss function, which is calculated by the square sum of the distance between the predicted value and the ground truth value. Mean absolute error (MAE) is another loss function used in regression models. MAE is the sum of the absolute values of the difference between the target and predicted values. It measures only the mean modulus length of the predicted value error, not the direction. MSE is easy to calculate, but MAE is more robust to outliers.

## C. Individual Differences Between Subjects

One of the main challenges of task-based fMRI research is the use of multi-subject datasets. On one hand, multi-subject analysis is critical to understanding the universality and validity of results generated across subjects. On the other hand, analyzing multi-subject fMRI data requires accurate functional and anatomical alignment between the brain activities of different subjects to improve the performance of the final results [51]–[53].

By using the proposed method, the reconstructed natural images for all six subjects in DS105 (category=bottles) are shown in Fig.8, where the top row is the presented visual stimuli, and the following rows are the reconstructed images obtained from six individual subjects in the dataset. As is shown in Fig.8, although the details of the reconstructed images varied slightly, all the subjects achieve acceptable reconstruction results. Further, we calculate the correlation of the reconstruction results of different subjects, all the correlations between subjects are greater than 0.7.

TABLE VII
THE EXPERIMENT RESULTS ON DS105 (CATEGORY=BOTTLES) VIA INDIVIDUAL-SUBJECT AND MULTI-SUBJECT

|  | Euc_dis↓ | PCC↑ | MSE↓ |
|---|---|---|---|
| Individual S1(S1) | 0.583±0.075 | 0.728±0.091 | 0.069±0.071 |
| Individual S1(S2) | 0.625±0.086 | 0.694±0.102 | 0.083±0.076 |
| Multi-Subject(S1) | 0.593±0.047 | 0.711±0.055 | 0.078±0.049 |
| Multi-Subject(S2) | 0.586±0.063 | 0.707±0.085 | 0.073±0.052 |

In order to clarify the individual differences between subjects, we also conduct an experiment performed on the DS105 dataset (category=bottles). To compare the difference between the individual subject and multi-subject, we use **Subject #1** as an example. As the data from Subject #1 have 12 runs of visual task experiment, we use the leave-one-run-out strategy for cross-validation. The results are listed in Table VII. In the table, (S#) (#=1,2) means that the #-th subject is the test subject. Compared with multi-subject data, when the experiment is performed on individual Subject #1, better performance has

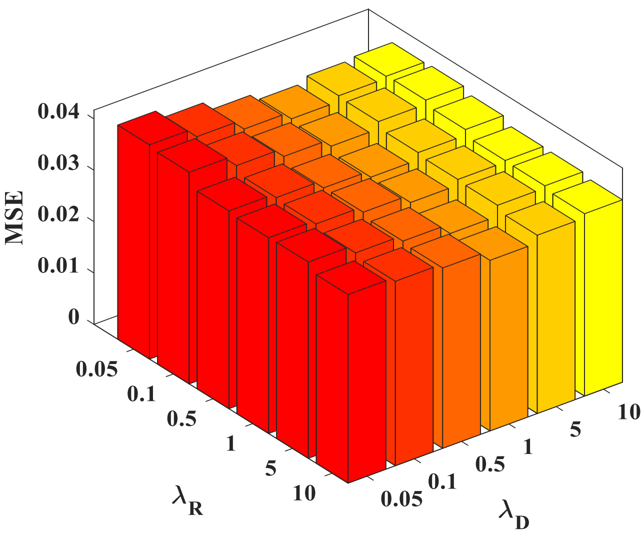

Fig. 9. Reconstruction results (MSE) of the Handwritten Digits dataset vis different values of $\lambda_R$ and $\lambda_D$.

been obtained because of the isotropic distribution of data. However, when the new Subject #2 is added, the utilization of multi-subject data can maintain more stable results than individual data.

### D. Effects of Regularization Parameters

In our proposed method, the optimized objective function is mainly composed of two parts, pairwise ranking loss $L_R$ and discriminator loss $L_D$. In order to balance the two loss parts and control their influence in the model, we set two hyperparameters $\lambda_R$ and $\lambda_D$. We conduct image reconstruction experiments on the Handwritten Digit Dataset with different regularization parameters chosen from $\{0.05, 0.1, 0.5, 1, 5, 10\}$, and the results are displayed in Fig. 9.

As can be seen from Fig. 9, we can observe that with different parameters, our TIGAN method can obtain relatively stable reconstruction results. And the best regularization parameter can be chosen from $\lambda_D = 1$ and $\lambda_R = \{1, 5, 10\}$, where the proposed TIGAN achieves better results.

### E. Influence of Different Number of LSTM Layers

In this section, we conduct image reconstruction experiments on the category of bottles in DS105 to discuss the decoding effect of different number of LSTM layers.

Under the premise of keeping the other structures of the model unchanged, we set the number of layers of the LSTM module in our proposed TIGAN as 1, 2 and 3. The models with different numbers of LSTM layers were tested in stimuli image reconstruction experiments. Same as the experiments above, three evaluation metrics are used to measure the reconstruction performances of different methods, which are Euclidean distance (Euc_dis), Pearson's correlation coefficient (PCC) and mean squared error (MSE), respectively. And the results are listed in Table VIII.

It can be seen from Table VIII that with the increase in the number of layers of LSTM, the performance of PCC showed a trend of decline and the Euc_dis and MSE showed a trend of rising. When the number of layers of LSTM was increased to three, the performance of the model decreased significantly, and over-fitting was likely to occur.

TABLE VIII
THE EXPERIMENT RESULTS OF USING DIFFERENT NUMBERS OF LSTM LAYERS. (# MEANS THE NUMBER OF LSTM LAYERS)

| # | Euc_dis↓ | PCC↑ | MSE↓ |
|---|---|---|---|
| 1 | **0.581±0.069** | **0.719±0.062** | **0.081±0.049** |
| 2 | 0.589±0.053 | 0.712±0.081 | 0.087±0.061 |
| 3 | 0.606±0.077 | 0.701±0.042 | 0.098±0.059 |

### F. Correlation of Cross-modal Pairwise Samples

One of the most significant challenges in the field of stimuli image reconstruction is how to model the relationship between two modalities of data, stimuli images and the evoked brain activities, respectively. In this section, we will discuss the effectiveness of the ranking loss used in our model. Three sets of experiments are included and the description are as follows:

**Circumstance A**: training the model by "bottles", testing by "cats".

**Circumstance B**: training the model by "bottles", testing by "bottles" images but the activities corresponding to seeing "cats".

**Circumstance C**: training the model by "bottles" and "cats", testing by "bottles" images but the activities corresponding to seeing "cats".

The reconstruction results under the three circumstances are shown in Fig. 10. As can be seen in the figure, when we train the model by "bottles" and test by "cats", "cats" seem to be a new label under this circumstance. But we can also obtain the reconstruction results of "cats". Under circumstance B, the reconstruction gets the worst performance. The reason may be that all the image samples are "bottles", the new-added "cats" brain activity cannot learn the related image features. At the last, when we add the "cats" samples to the training set, TIGAN can reconstruct the true images even uses the random data-label pairs.

In addition, we also did quantitative analysis by computing the ranking loss to show the relevance of pairwise sample $(\hat{x}_t, \hat{y}_t)$. The results are listed in Table IX. As is shown in the

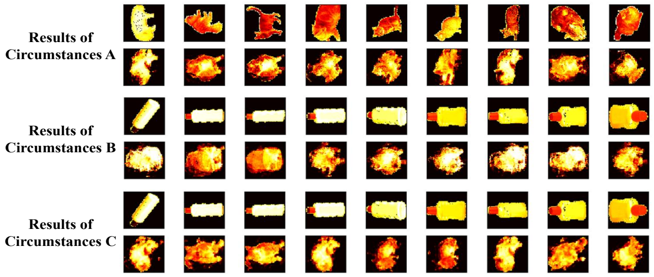

Fig. 10. Experimental results by randomizing the data-label pair.

table, the positive (ground truth) pairs obtained the minimizing ranking loss, which means that they are the most relevant pairs. For circumstance A, even the training set and the test set have different labels, but the value of the ranking loss is also small because the test set samples are data-label pairs of “cats”. For circumstances B and C, the value of ranking loss gets high because of the randomizing pair-label pairs, and circumstances B is higher because of the disappearance of “cats” in the training phase.

TABLE IX
THE RESULTS OF RANKING LOSS UNDER DIFFERENT CIRCUMSTANCES.

| | Ranking Loss |
|---|---|
| Positive Pair | 0.00789 |
| Circumstances A | 0.00969 |
| Circumstances B | 0.36221 |
| Circumstances C | 0.25604 |

### G. Limitations and Future Work

We believe that there are still several limitations in the current work. First, the sample size of the task-based fMRI datasets used in this paper is still small due to the difficulty of data collection. To solve the problem, some studies are proposed based on domain adaptation [54] and transfer learning [55] algorithms. In order to make the algorithms available to large-scale and multi-site fMRI datasets, this is an important issue. Second, the proposed method consists of three subnetworks, which will increase the memory burden for visual reconstruction. Hence, model compression is an important research direction for practical applications. Third, as an end-to-end machine learning method, deep neural networks often have difficulty observing internal mechanisms. In the future, we plan to add the human visual imaging mechanism as prior knowledge to the visual image reconstruction task, to realistically simulate the human brain activities when processing visual signals. Fourth, the proposed method in this paper does not make good use of the structural information of whole-brain structure data. In future studies, we plan to develop information-based models based on understanding the intrinsic information of whole-brain structure data to smooth the data information of small areas. It makes the information valid area in the whole brain data clearer and provides better input information for visual image reconstruction. Finally, as research of neural science, it is also important to use machine learning methods to explore some questions related to biological information. The proposed TIGAN method is a fMRI→image path currently. In the future work, we will also build the image→fMRI path in our model to discover the brain activation maps for specific stimuli.

## VI. CONCLUSION

In this paper, we present a temporal information guided GAN (TIGAN) method for stimuli image reconstruction from human brain activities. Three key components are consisted in the proposed TIGAN method, including a stimuli image encoder, an LSTM generator and a discriminator for image reconstruction. The proposed TIGAN is not only a generative model to model the relationship between the stimuli image and the evoked brain activities, but also takes the temporal information of fMRI data into account. Furthermore, the pairwise ranking loss is introduced to measure the relationship between the stimuli images and the evoked fMRI scans, which ensures that the strongly associated pairs are at the top and the weakly related ones are at the bottom. Experiments on both the DS105 and the handwritten digits datasets suggest that our reconstruction model can also achieve better performance in comparison with state-of-the-art reconstruction methods.

## REFERENCES

[1] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, “Distributed and overlapping representations of faces and objects in ventral temporal cortex,” *Science*, vol. 293, no. 5539, pp. 2425–2430, 2001.

[2] K. Smith, “Brain decoding: reading minds,” *Nature*, vol. 502, no. 7472, pp. 428–430, 2013.

[3] M. Yousefnezhad, A. Selvitella, L. Han, and D. Zhang, “Supervised hyperalignment for multi-subject fMRI data alignment,” *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

[4] E. Goddard and K. T. Mullen, “fMRI representational similarity analysis reveals graded preferences for chromatic and achromatic stimulus contrast across human visual cortex,” *NeuroImage*, vol. 215, p. 116780, 2020.

[5] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant, “Identifying natural images from human brain activity,” *Nature*, vol. 452, no. 7185, pp. 352–355, 2008.

[6] J. V. Haxby, A. C. Connolly, and J. S. Guntupalli, “Decoding neural representational spaces using multivariate pattern analysis,” *Annual review of neuroscience*, vol. 37, no. 1, pp. 435–456, 2014.

[7] H. Wu, N. Zheng, and B. Chen, “Feature-specific denoising of neural activity for natural image identification,” *IEEE Transactions on Cognitive and Developmental Systems*, 2021.

[8] R. J. Brachman and J. G. Schmolze, “An overview of the KL-ONE knowledge representation system,” *Cognitive Science*, vol. 9, no. 2, pp. 171–216, April–June 1985.

[9] M. A. Van Gerven, B. Cseke, F. P. De Lange, and T. Heskes, “Efficient bayesian multivariate fmri analysis using a sparsifying spatio-temporal prior,” *NeuroImage*, vol. 50, no. 1, pp. 150–161, 2010.

[10] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, “How does the brain solve visual object recognition?” *Neuron*, vol. 73, no. 3, pp. 415–434, 2012.

[11] B. Thirion, E. Duchesnay, E. Hubbard, J. Dubois, J.-B. Poline, D. Lebihan, and S. Dehaene, “Inverse retinotopy: inferring the visual content of images from brain activation patterns,” *Neuroimage*, vol. 33, no. 4, pp. 1104–1116, 2006.

[12] Y. Miyawaki, H. Uchida, O. Yamashita, M.-a. Sato, Y. Morito, H. C. Tanabe, N. Sadato, and Y. Kamitani, “Visual image reconstruction from human brain activity using a combination of multiscale local image decoders,” *Neuron*, vol. 60, no. 5, pp. 915–929, 2008.

[13] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015.

[14] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural networks*, vol. 61, pp. 85–117, 2015.

[15] F. Tao and C. Busso, “Gating neural network for large vocabulary audiovisual speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1290–1302, 2018.

[16] S. Ghorbani, S. Khorram, and J. H. Hansen, “Domain expansion in dnn-based acoustic models for robust speech recognition,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 107–113.

[17] L. Sun, W. Shao, D. Zhang, and M. Liu, “Anatomical attention guided deep networks for ROI segmentation of brain MR images,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 2000–2012, 2020.

[18] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, “CE-Net: Context encoder network for 2D medical image segmentation,” *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.

[19] Y. Güçlütürk, U. Güçlü, K. Seeliger, S. Bosch, R. van Lier, and M. A. van Gerven, "Reconstructing perceived faces from brain activations with deep adversarial neural decoding," in *Advances in Neural Information Processing Systems*, 2017, pp. 4246–4257.

[20] C. Du, C. Du, L. Huang, and H. He, "Reconstructing perceived images from human brain activities with bayesian deep multiview learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 8, pp. 2310–2323, 2019.

[21] G. Shen, K. Dwivedi, K. Majima, T. Horikawa, and Y. Kamitani, "End-to-end deep image reconstruction from human brain activity," *Frontiers in computational neuroscience*, vol. 13, p. 21, 2019.

[22] H. Wang, L. Huang, C. Du, D. Li, B. Wang, and H. He, "Neural encoding for human visual cortex with deep neural networks learning "what" and "where"," *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

[23] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7181–7189.

[24] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, "From deterministic to generative: Multimodal stochastic rnns for video captioning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 10, pp. 3047–3058, 2018.

[25] X. Jiang, F. Wu, X. Li, Z. Zhao, W. Lu, S. Tang, and Y. Zhuang, "Deep compositional cross-modal learning to rank via local-global alignment," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 69–78.

[26] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 299–307.

[27] M. Wang, W. Shao, X. Hao, L. Shen, and D. Zhang, "Identify consistent cross-modality imaging genetic patterns via discriminant sparse canonical correlation analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, early access, October, 2019, doi:10.1109/TCBB.2019.2944825.

[28] I. Bloch, C. Pellot, F. Sureda, and A. Herment, "3D reconstruction of blood vessels by multi-modality data fusion using fuzzy and markovian modelling," in *International Conference on Computer Vision, Virtual Reality, and Robotics in Medicine*. Springer Berlin Heidelberg, 1995, pp. 392–398.

[29] N. Pugeault, F. Worgotter, and N. Kruger, "Multi-modal scene reconstruction using perceptual grouping constraints," in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*. IEEE, 2006, p. 195.

[30] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2008–2020, 2015.

[31] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1363–1377, 2016.

[32] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, "Cross-modal retrieval with cnn visual features: A new baseline," *IEEE Transactions on Cybernetics*, vol. 47, no. 2, pp. 449–460, 2017.

[33] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 2121–2129. [Online]. Available: http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model.pdf

[34] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[35] Y. Peng and J. Qi, "CM-GANs: Cross-modal generative adversarial networks for common representation learning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1, pp. 1–24, 2019.

[36] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612*, 2017.

[37] 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, 2009.

[38] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, "Reconstructing visual experiences from brain activity evoked by natural movies," *Current Biology*, vol. 21, no. 19, pp. 1641–1646, 2011.

[39] Y. Fujiwara, Y. Miyawaki, and Y. Kamitani, "Modular encoding and decoding models derived from bayesian canonical correlation analysis," *Neural computation*, vol. 25, no. 4, pp. 979–1005, 2013.

[40] Y. Zhan, J. Zhang, S. Song, and L. Yao, "Visual image reconstruction from fmri activation using multi-scale support vector machine decoders," in *International Conference on Human-Computer Interaction*. Springer, 2013, pp. 491–497.

[41] C. Du, C. Du, and H. He, "Sharing deep generative representation for perceived image reconstruction from human brain activity," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 1049–1056.

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

[43] G. St-Yves and T. Naselaris, "Generative adversarial networks conditioned on brain activity reconstruct seen images," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2018, pp. 1054–1061.

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

[45] C. Du, C. Du, L. Huang, H. Wang, and H. He, "Structured neural decoding with multitask transfer learning of deep neural network representations," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[46] W. Huang, H. Yan, C. Wang, X. Yang, J. Li, Z. Zuo, J. Zhang, and H. Chen, "Deep natural image reconstruction from human brain activity based on conditional progressively growing generative adversarial networks," *Neuroscience Bulletin*, vol. 37, no. 3, pp. 369–379, 2021.

[47] T. Fang, Y. Qi, and G. Pan, "Reconstructing perceptive images from brain activity by shape-semantic GAN," *arXiv preprint arXiv:2101.12083*, 2021.

[48] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[49] M. A. van Gerven, F. P. de Lange, and T. Heskes, "Neural decoding with hierarchical generative models," *Neural computation*, vol. 22, no. 12, pp. 3127–3142, 2010.

[50] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International Conference on Machine Learning*, 2015, pp. 1083–1092.

[51] J. V. Haxby, J. S. Guntupalli, A. C. Connolly, Y. O. Halchenko, and P. J. Ramadge, "A common, high-dimensional model of the representational space in human ventral temporal cortex," *Neuron*, vol. 72, no. 2, pp. 404–416, 2011.

[52] A. Lorbert and P. J. Ramadge, "Kernel hyperalignment," *Advances in Neural Information Processing Systems*, pp. 1790–1798, 2012.

[53] P. H. Chen, J. Chen, Y. Yeshurun-Dishon, U. Hasson, and P. J. Ramadge, "A reduced-dimension fmri shared response model," in *Advances in Neural Information Processing Systems*, 2015, pp. 460–468.

[54] M. Wang, D. Zhang, J. Huang, P.-T. Yap, D. Shen, and M. Liu, "Identifying autism spectrum disorder with multi-site fmri via low-rank domain adaptation," *IEEE transactions on medical imaging*, vol. 39, no. 3, pp. 644–655, 2019.

[55] H. Zhang, P.-H. Chen, and P. Ramadge, "Transfer learning on fmri datasets," in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 595–603.