Functional Alignment-Auxiliary Generative Adversarial Network-based Visual Stimuli Reconstruction via Multi-subject fMRI

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

***Abstract*—Functional Magnetic Resonance Imaging (fMRI) provides more precise spatial and temporal information to reconstruct stimulus images than other technologies that can be used to measure the human brain's neural responses. The fMRI scans, however, generally show heterogeneity among different subjects. The majority of the existing methods aim primarily at mining correlations between stimuli and evoked brain activity, disregarding the heterogeneity among subjects. Therefore, this heterogeneity will impair the reliability and applicability of multi-subject decoding results, leading to sub-optimal results. The present paper proposes the functional alignment-auxiliary generative adversarial network (FAA-GAN) as a novel multi-subject approach for visual image reconstruction that employs functional alignment to alleviate the heterogeneity between subjects. Our proposed FAA-GAN includes three key components: 1) a generative adversarial network (GAN) module for reconstructing visual stimuli, which consists of a visual image encoder as the generator that uses a nonlinear network to convert stimuli images into an implicit representation and a discriminator that generates the images comparable to the original images in detail; 2) a multi-subject functional alignment module, which is used to precisely align the individual fMRI response space of each subject in a common space to reduce the heterogeneity among different subjects; and 3) a cross-modal hashing retrieval module used for similarity retrieval of two modalities of data, i.e., the visual images and the evoked brain responses. Experiments on real-world datasets show that our FAA-GAN method does better than other state-of-the-art deep learning-based reconstruction methods with fMRI.**

***Index Terms*—Functional Magnetic Resonance Imaging, visual image reconstruction, multi-subject analysis, functional alignment, generative adversarial network**

I. Introduction

HUMAN brain comprehension has been one of the challenges that needs urgent attention for a long time [1]–[4]. Human brain mapping and decoding is an interdisciplinary field of study that explores how the brain performs different cognitive functions [5]–[7]. The key idea behind brain decoding is to identify cognitive states through the measurement of neural activity [8]–[10]. Functional Magnetic Resonance Imaging (fMRI) is an imaging technique that has made tremendous advancements in the area of human brain research. Brain decoding is made possible by fMRI's ultrahigh spatiotemporal resolution, which can provide more accurate information for neural activity analysis by using blood oxygen level dependent (BOLD) signals as a surrogate for neural activity visualization [6], [11]–[13].

Despite the enormous advancement of machine learning technologies, which has enabled excellent decoding results, detailed information about visual stimuli is still ignored due to classification task limitations. Researchers use visual image reconstruction to better understand human minds and display more details. Thirion et al. [14] published a preliminary work reconstructing dot patterns from both seen and imagined images using rotating Gabors. The visual cortex's retinotopy was used to deduce the visual information of real or imagined scenes from the neural responses they elicited. Individuals' BOLD signals in the early visual cortex were recorded when they were presented with the stimuli of flashing checkerboard images. A multi-scale local image decoder was developed to reconstruct perception. Inspired by the pioneers, scientists did a series of studies [15]–[19] and made a great contribution to this field. With the rapid development of deep learning, a number of deep neural network (DNN)-based methods have been proposed for reading people's minds. Some studies have used DNN outputs to shed light on how the human visual cortex actually works [20]–[23]. However, no matter conventional machine learning or deep learning methods, most of them neglect the heterogeneous patterns in multi-subject datasets, which is one of the difficulties that need to be overcome in the study of human cognitive analysis [11], [24]–[26].

Aligning fMRI data from various subjects is necessary for multi-subject cognitive analysis tasks. This is done to overcome the heterogeneity that exists across the subjects [11], [13], [25], [27]. This alignment problem can be viewed as multi-view representation learning from a machine learning perspective [25], [27]. Indeed, functional alignment is based on the idea that there exists shared information among subjects, and neural data alignment means pulling out this shared information. On the other hand, the goal of functional alignment is to perfectly align the response space across subjects.

Herein, we propose a novel stimuli reconstruction method, functional alignment-auxiliary generative adversarial network (FAA-GAN), to reconstruct visual stimuli from multi-subject brain response patterns. To address the heterogeneity in the

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. M. Yousefnezhad is with Department of Computing Science, University of Alberta, T6G 2R3, Canada.

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

This work was supported by the National Natural Science Foundation of China (Nos. 62136004, 61732006, 62006115, and 62106104), the National Key R&D Program of China Grant (Nos. 2018YFC2001600, 2018YFC2001602), the China Postdoctoral Science Foundation (No. 2022T150320) and the Chinese Association for Artificial Intelligence (CAAI)-Huawei MindSpore Open Fund.

fMRI data from various subjects, we add a functional alignment module in our reconstruction model to align the response patterns in a common latent space. The proposed FAA-GAN is comprised of three main components. Firstly, for visual stimulus reconstruction, a GAN-based module is covered. For the generator, we pretrain an image autoencoder, whose encoder maps the stimulus images to a latent space and the decoder can be viewed as the generator that reconstructs the visual images. Simultaneously, reconstructed images are made to look as close as possible to the originals using the discriminator in our GAN module. The second part is the multi-subject functional alignment module, which is utilized to align each subject's neural response in the common space accurately. The third part is the cross-modal hashing retrieval module used for similarity retrieval across the stimuli and the brain response patterns.

The major contributions of this paper are listed as follows:

- We propose a novel visual stimuli reconstruction method that reconstructs visual images from different subjects' brain activities. Our method not only focuses on cross-modal data reconstruction but also considers the heterogeneity across subjects. We add functional alignment as a module of the method, where the parameters are learned and updated during the model training.
- We introduce a cross-modal retrieval module to estimate the correlation between visual stimuli and resulting responses. Given a response pattern, the most associated (corresponding) features of the image could be retrieved.
- We evaluate the proposed FAA-GAN method via two different datasets, one for natural image reconstruction and one for character reconstruction. When compared to other state-of-the-art deep learning-based approaches, FAA-GAN delivers the best reconstruction performance, as shown by experimental results.

II. Related Work

A. Visual Image Reconstruction

Compared with brain decoding, i.e., the classification tasks, reconstructing stimulus images remains an enormous challenge. Researchers have explored visual stimulus reconstruction, which may be separated into conventional machine learning and deep learning methods. For conventional methods, inspired by [28], Bayesian reconstruction models were developed to study fMRI voxel correlations that naturally reflect visual stimuli. Naselaris et al. [15] established a Bayesian framework that exploits the structural and semantic characteristics of encoding the activities of the human brain to more accurately depict the spatial structure and semantic categories contained during image observation. Nishimoto et al. [16] suggested a Bayesian decoding approach to speed up fMRI data acquisition. They devised a motion-energy encoding paradigm to overcome fMRI BOLD signal latency. However, these methods don't look into how the stimuli and the neural responses are linked. To automatically acquire knowledge of image databases, according to the Bayesian canonical correlation analysis (BCCA) model proposed by Fujiwara et al. [17], each module is expressed by a latent variable that is linked to a distinct subset of pixels. On the basis of 2D images, by constructing contrast-decoding, disparity-decoding, and contrast-disparity decoding models, Zheng et al. [29] reconstruct 3D contrast images from fMRI data in visual regions.

In the previous decade, deep learning-based stimuli image reconstruction achieved exciting progress. For example, in [21], [30], Du et al. proposed the deep generative multi-view model (DGMM), which is a nonlinear version of BCCA. Shen et al. used fMRI and visual images to train a DNN model in [22]. One of the main contributions of this research is the development of an end-to-end model that can directly link neural responses with external visual stimuli. Besides, Du et al. [31] utilized a hierarchically structured framework with multi-task transfer learning of DNN representations and the matrix-variable Gaussian prior for neural decoding.

Meanwhile, several GAN-based methods have been proposed, significantly improving the accuracy of reconstruction performance [20], [32], [33]. For example, a GAN-based framework was proposed in [33] that learns a generative model of visual stimuli that is conditioned by observations of neural responses. Additionally, Seeliger et al. [33] proposed a deep convolutional generative adversarial network (DCGAN) method employing adversarial training to reconstruct arbitrary natural images. DCGAN can be trained independently on sizable image datasets and learn the hidden space unsupervised. Recently, more and more GAN-based methods have been proposed as a result of the technology's rapid development [4], [34], [35]. However, most of the existing methods use multi-subject fMRI data for visual stimulus reconstruction while neglecting the heterogeneity across different subjects, which will affect the stability and robustness of the model's performance. In this paper, we propose a new method that not only reconstructs visual stimuli with high accuracy using a GAN-based framework, but also takes into account the differences between subjects by mapping the fMRI responses of different people onto a common space.

B. Multi-Subject Functional Alignment

As an extremely important study of functional alignment in the initial period, Haxby et al. [11] developed hyperalignment (HA), a technique that standardizes the high-dimensional environment in which brain responses from different people occur. Following Haxby, a slew of studies have offered numerous improved techniques for improving hyperalignment accuracy. Xu et al. [24] presented regularized hyperalignment (RHA) to discover optimal regularization parameters. They found that the regularization parameters control how each normalized dataset's singular vectors are weighted, and altering them improves classification accuracy. The non-linear transformation in the embedded kernel space can be accomplished with the help of the kernel hyperalignment (KHA) technique described in [12]. When the number of individuals increases, the challenge of HA shifts, and KHA is able to address both the voxel and feature expansion difficulties simultaneously. Chen et al. proposed a series of important methods for functional alignment. First, they proposed the singular value decomposition hyperalignment (SVDHA) [27] and joint SVD was

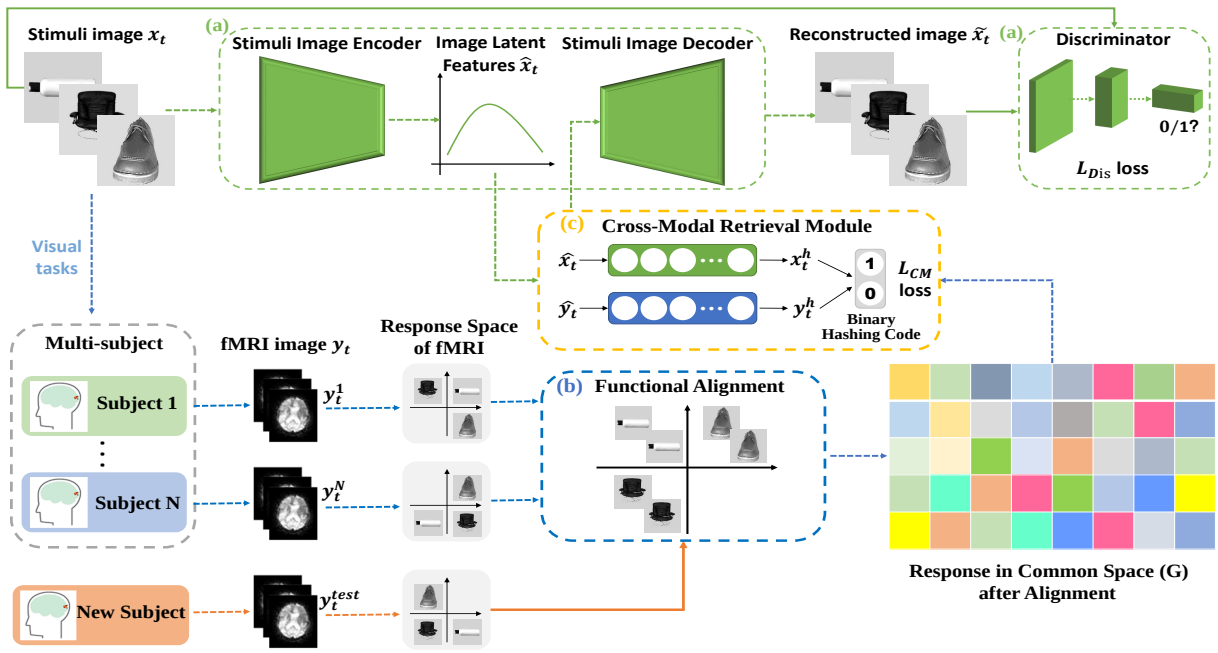

Fig. 1. The schematic diagram of the proposed functional alignment-auxiliary GAN method. Three main components are included in the proposed model, i.e., (a) the generative adversarial networks (GANs) module for image reconstruction, which is in the dotted green area; (b) the multi-subject functional alignment module for aligning the fMRI response onto a common space, which is in the dotted blue area; and (c) the cross-modal retrieval module for similarity retrieving across different modalities of data, which is in the dotted yellow area.

used for decomposing the response matrix. The subjects were then aligned in a new feature space with fewer dimensions via the HA method, which speeds up calculations without sacrificing accuracy in classification. Additionally, they developed a shared response model (SRM) [36], a unique transformation for the probability principal component analysis (PCA). Specifically, the loading matrix is subject to orthogonal constraints. With SRM, the size of the shared feature space was also reduced. In addition, a convolutional autoencoder (CAE) method was developed as a nonlinear technique for use with whole-brain fMRI data [37]. As a parametric kernel model, a method named deep hyperalignment (DHA) was developed by Yousefnezhad et al. [25]. Nonlinear issues were resolved via DHA, while the classification prediction performance was also improved. Further, they also suggested supervised approaches, local discriminant hyperalignment (LDHA) [13] and supervised hyperalignment (SHA) [26]. It is possible for stimuli from the same category to be more correlated with one another and for stimuli from other categories to be less associated with one another in the shared spaces, thanks to the use of the supervised technique.

III. Proposed Method

A. Notations

Let $\boldsymbol{X} = \{x_{td}\} \in \mathbb{R}^{T\times D}, t = 1:T, d = 1:D$ be the images presented to the subjects in the visual tasks. T refers to the quantity of images and D is the dimensions of each image. Here, we assume that the images used in a same visual task have the same size. In our multi-subject fMRI data, we let S be the number of subjects. For the i-th subject of S subjects, the brain neural responses are denoted by $\boldsymbol{Y}^{(i)} = \{y_{tv}^{(i)}\} \in \mathbb{R}^{T\times V}, i = 1:S, t = 1:T, v = 1:V$, where T is the number of time points in units of Repetition Time (TR), V is the number of voxels, and $y_{tv}^{(i)}$ denotes the brain neural responses of the i-th subject in the t-th time point and the v-th voxel. Furthermore, we use $\boldsymbol{Y} = \{\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}, \ldots, \boldsymbol{Y}^{(S)}\}$ to represent the neural responses in all the subjects' brains in the dataset. In our multi-subject visual image reconstruction task, there are two assumptions, i.e., 1) the images and evoked brain activities are pairwise samples, which asserts that there are the same number of images and neural responses, so we use the same T in our paper; 2) it is assumed that the training set's stimuli are already aligned in time, i.e., all the subjects watch the same stimuli image at any t-th time point.

B. Functional Alignment-Auxiliary Generative Adversarial Network (FAA-GAN)

We develop a multi-subject visual image reconstruction method, which aims to generate seen images through different subjects' brain neural responses. In our proposed method, two distinct modules for visual images and multi-subject response patterns map two different modalities of data into a common latent space, respectively, and the GAN architecture reconstructs the visual images. Fig. 1 is a simplified representation of our FAA-GAN method. Specifically, the proposed model consists of three main modules, i.e.,

1) the GAN-based module for image reconstruction, which consists of a pre-trained auto-encoder and a discriminator. The encoder maps the input image features $\boldsymbol{X}$ into latent image features $\hat{x}_t$ on a latent space z, and the decoder can be viewed as the generator that reconstructs the latent features $\hat{x}_t$ to the reconstructed images $\tilde{x}_t$. The discriminator is used to make

the reconstructed images $\widetilde{x_t}$ similar to the original inputs $\boldsymbol{X}$;

2) the multi-subject functional alignment module, which is used to align each subject's neural response in the common space accurately. For functional alignment, the response neural of each subject $\boldsymbol{Y}^{(i)}, i = 1, 2, \ldots, S$ is input to the network $f_\ell(\boldsymbol{Y}^{(\ell)}; \theta^{(\ell)})$, then rotated to the common space G via the transformation matrix $\boldsymbol{R}^{(\ell)}$, the aligned neural features $\hat{\boldsymbol{Y}^{(\ell)}} = f_{(\ell)}(\boldsymbol{Y}^{(\ell)}; \theta^{(\ell)})\boldsymbol{R}^{(\ell)}$;

3) the cross-modal hashing retrieval module is used for similarity retrieving across two modalities of data, i.e., neural responses and images. In this module, the latent features $(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{y}}_t)$ are encoded into the hash code $(\boldsymbol{x}_t^h, \boldsymbol{y}_t^h)$. We learn the relationship between $\boldsymbol{x}_t^h$ and $\boldsymbol{y}_t^h$ and retrieving the most related hash code from one modal to the other.

During the model training, the visual images and the fMRI data from different training subjects are input into the image encoder and different networks for functional alignment, respectively. Then the image encoder maps the images into image latent space z and the output of functional alignment is the common response space G, and the transformation matrix. We combine the two cross-modal latent spaces into a common space via the cross-modal retrieval module to learn the relationship between the stimuli and the brain response patterns. The trained image latent features are sent to the image generator as input, and the output is the reconstructed image.

During the testing, we only use the activity patterns of the test subject as the input. Specifically, the common space is learned in the training phase. Then, for the unseen subject in testing and practical applications, we will find a transformation to map the unseen subject into the learned common space. Via the trained cross-modal retrieval module, the most relevant image features will be retrieved based on the response patterns. Next, the trained image generator takes the image latent features as input, and reconstruct the visual stimuli. The functional alignment provides more robust and generalized results in the field of multi-subject analysis, which could be adapted to more application scenarios.

1) Visual Image Generative Module: Due to the visual task limitation, only a small number of related samples could be used for network training. To pretrain an autoencoder and increase its performance, we refer to [33] and our previous work [35]. After autoencoder pretraining, the image encoder network maps the seen images to the latent space z, and the latent features $\hat{x}_t = E_\theta(x_t)$, where $E(\cdot)$ is an encoder function, and θ is its parameter. For the decoder network, the reconstructed image $\widetilde{x_t} = D_\phi(\hat{x}_t) = D_\phi(E_\theta(x_t))$ generated via the nonlinear decode function $D(\cdot)$. Here, ϕ is the decoder parameters. The autoencoder's loss function is described as

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

2) Multi-Subject Functional Alignment Module: In order to overcome the problem of response heterogeneity among different subjects [11], [26], [27], in this paper, we add a module for the functional alignment of multi-subject neural responses, which is used for multi-subject fMRI analysis in our visual image reconstruction method. One of the most popular functional alignment technologies is hyperalignment proposed in [11]. A fundamental assumption of HA is that different subjects' brain activity patterns are noisy 'rotations' of a common space [11]. Let $\boldsymbol{Y}^{(i)} \in \mathbb{R}^{T\times V}, t = 1:T, v = 1:V, i = 1:S$ denote the i-th subject's response matrix, $\boldsymbol{R}^{(i)} \in \mathbb{R}^{V\times V}$ is the corresponding transformation matrix, HA can be viewed as a CCA-based minimization problem [6], [24], [25], [27]:

$$\begin{aligned} \min_{\boldsymbol{R}^{(i)},\boldsymbol{R}^{(j)}} \sum_{i=1}^{S}\sum_{j=i+1}^{S} \left\| \boldsymbol{Y}^{(i)}\boldsymbol{R}^{(i)} - \boldsymbol{Y}^{(j)}\boldsymbol{R}^{(j)} \right\|_F^2 \\ s.t. (\mathbf{Y}^{(\ell)}\mathbf{R}^{(\ell)})^\top \boldsymbol{Y}^{(\ell)}\boldsymbol{R}^{(\ell)} = \boldsymbol{I}, \ell = 1:S. \end{aligned} \tag{2}$$

To build an end-to-end deep learning-based model, we use a nonlinear neural network in [25] and improvement has been made on this work. As aforementioned, for the i-th subject of the S subjects in the dataset, the response matrix $\boldsymbol{Y}^{(i)}$ is denoted by $\boldsymbol{Y}^{(i)} = \{y_{tv}^{(i)}\} \in \mathbb{R}^{T\times V}, t = 1:T, v = 1:V, i = 1:S$. Then, the alignment function can be defined as follows:

$$\begin{aligned} \min_{\substack{\theta^{(i)},\boldsymbol{R}^{(i)} \\ \theta^{(j)},\boldsymbol{R}^{(j)}}} \sum_{i=1}^{S}\sum_{j=1}^{S} \left\| f_i(\boldsymbol{Y}^{(i)};\theta^{(i)})\boldsymbol{R}^{(i)} - f_j(\boldsymbol{Y}^{(j)};\theta^{(j)})\boldsymbol{R}^{(j)} \right\|_F^2 \\ s.t. (\mathbf{R}^{(\ell)})^\top ((f_\ell(\boldsymbol{Y}^{(\ell)};\theta^{(\ell)})^\top f_\ell(\boldsymbol{Y}^{(\ell)};\theta^{(\ell)}) + \epsilon\boldsymbol{I})\boldsymbol{R}^{(\ell)} = \boldsymbol{I}, \\ \ell = 1:S. \end{aligned} \tag{3}$$

Here, $\boldsymbol{\theta} = \{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(S)}\}$ represents all the parameters of each subject's alignment network. The transformation matrix $\boldsymbol{R} = \{\boldsymbol{R}^{(1)}, \boldsymbol{R}^{(2)}, \ldots, \boldsymbol{R}^{(S)}\}$ transfers the individual neural responses of each subject into the common space, and the regularized parameter ϵ is a very small constant, such as 10^{-8}. In formulation (3), every i-th and j-th subjects need to be calculated, creating a huge amount of computation. At the same time, during the test, the testing subject need to be calculated with all the training subject one-by-one. To solve this problem, a reformulated objective function of equation (3) also be proposed [12], [25], [38]:

$$\min_{\boldsymbol{G},\theta^{(i)},\boldsymbol{R}^{(i)}} \sum_{i=1}^{S} \left\| \boldsymbol{G} - f_i(\boldsymbol{Y}^{(i)};\theta^{(i)})\boldsymbol{R}^{(i)} \right\|_F^2, \tag{4}$$

where $\boldsymbol{G} = \frac{1}{S}\sum_{j=1}^{S} f_j(\boldsymbol{Y}^{(j)};\theta^{(j)})\boldsymbol{R}^{(j)}$.

During training, we build S MLP networks $f_\ell(\boldsymbol{Y}^{(\ell)};\theta^{(\ell)})$, where S equals the number of subjects for each dataset, which means that each subject has an individual network. After the network mapping, the generated features are rotated to the common space $\boldsymbol{G}$ via the transformation matrix $\boldsymbol{R}^{(\ell)}$ based on (4), and $\theta^{(\ell)}$, $\boldsymbol{R}^{(\ell)}$, and $\boldsymbol{G}$ are learned.

In the testing phase, the response matrix of the new given subject $\boldsymbol{Y}^{(test)}$ will be mapped into the learned common space $\boldsymbol{G}$ based on (4). This is just needed as the first step in the testing phase because the common space $\boldsymbol{G}$ is calculated for this phase based on the training samples. As the second step during testing, the parameters $\theta^{(test)}$ must be updated. Here, we refer to the back-propagation algorithm [39] and optimize strategy in [24], [25].

Finally, the neural activity patterns after mapping:

$$\hat{\boldsymbol{Y}} = f_{(test)}(\boldsymbol{Y}^{(test)};\theta^{(test)})\boldsymbol{R}^{(test)}. \tag{5}$$

3) Cross-Modal Hashing Retrieval Module: In the stimuli image reconstruction task, a challenge that needs to be addressed is how to mine the correlation of cross-modal data, which means the correlation between the seen images and the neural responses of subjects. In the field of computer vision (CV), an increasingly important and powerful solution is cross-modal hashing retrieval [40]–[43]. Among them, Cao et al. [40] proposed a cross-modal hashing approach called Cross-Modal Hamming Hashing (CMHH). The compact and highly concentrated hash codes provided by CMHH assist the available and efficient retrieval of Hamming space. In this paper, we add CMHH to our task of reconstructing visual stimuli in order to find out how the images relate to the neural responses they cause.

Through the stimuli image generator and multi-subject functional alignment module above, the latent feature representation of images and the response patterns are generated respectively. For the stimuli images, we add a hash layer of H hidden units. The output of the stimuli image generator is converted into continuous code $\hat{\boldsymbol{x}}_t \in \mathbb{R}^H$ by the image hash layer, and this process is repeated for each image $\boldsymbol{x}_t$. The hash code of image $\boldsymbol{x}_t$ is obtained through sign threshold $\boldsymbol{x}_t^h = sgn(\hat{\boldsymbol{x}}_t)$. For neural responses, we also adopt a hash layer of H hidden units after the multi-subject functional alignment module. We put the aligned neural response features in the common space $\boldsymbol{G}$ as the input to the hash layer. Then, the hash code $\boldsymbol{y}_t^h$ for each response pattern $\boldsymbol{y}_t$ can be obtained via sign threshold $\boldsymbol{y}_t^h = sgn(\hat{\boldsymbol{y}}_t)$. Quality hash codes used for effective retrieval are guaranteed by preserving the similarity between paired training samples $\{(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{y}}_t, s_{ij}) : s_{ij}\} \in \mathcal{S}$ and by controlling quantization error [40]. Given a set of pairwise training samples with labeled similarity as $\{(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{y}}_t, s_{ij}) : s_{ij}\} \in \mathcal{S}$, the hash codes denoted by $\boldsymbol{X}^h = [\boldsymbol{x}_1^h, \boldsymbol{x}_2^h, \ldots, \boldsymbol{x}_t^h]$ and $\boldsymbol{Y}^h = [\boldsymbol{y}_1^h, \boldsymbol{y}_2^h, \ldots, \boldsymbol{x}_t^h]$, where $t = 1, 2, \ldots, T$ for T pairwise training samples. The learning and optimization strategies of the cross-modal retrieval module follow [40] and the cross-modal focal loss L_{CM} can be derived as

$$\begin{aligned}\min \sum_{s_{ij}\in\mathcal{S}} [s_{ij}(1 - exp(-\gamma dis(x,y)))^{\sigma}\gamma dis(x,y) - (1- \\ s_{ij})(exp(-\gamma dis(x,y)))^{\sigma} log(exp(1-\gamma dis(x,y)))],\end{aligned} \tag{6}$$

$$dis(x,y) = \left\|\boldsymbol{x}_i^h - \boldsymbol{y}_j^h\right\|_2^2, \tag{7}$$

where γ is the scaling parameter that controls the precision-recall trade-off, $\sigma \geq 0$ is a hyper-parameter that control the relative weight of mismatched sample pairs, and $dis(x, y)$ refers to the continuous codes' Euclidean distance or the hash codes' Hamming distance.

4) Reconstructed Image Discrimination: Our goal when utilizing a GAN-based technique to reconstruct an image is to come up with a version that is visually indistinguishable from the original. The discriminator in GANs takes an input that is either an original or a generated image. It then uses a binary decision to decide if the input is real or not. This gives an output of 1 or 0. In our task, the real sample is the original visual image x_t, and the fake sample is the reconstructed image $\widetilde{x}_t$. The hidden state is then given to a sigmoid function, which is asked to predict if the image is real.

We compute the reconstruction loss $\mathcal{L}_{RE}$ between the reconstructed image $\widetilde{x}_t$ and the seen image x_t using two loss components computed by trained deep neural networks. The pixel-level loss, $\mathcal{L}_{pix}$, is the first loss component, which influences whether features are activated after crossing a threshold. The difference between the original image and the reconstructed image is measured by the mean squared error (MSE). The pixel-level loss $\mathcal{L}_{pix}$ can be determined as

$$\min \sum_{t=1}^{T}\sum_{p=1}^{P}(x_t^p - \widetilde{x}_t^p)^2, \tag{8}$$

where $t = 1, 2, \ldots, T$ means the t-th image in all the T images, and $p = 1, 2, \ldots, P$ means the p-th pixel of the image with P pixels.

The second loss component is the discrimination loss $\mathcal{L}_{dis}$. In our method, we trick the discriminator by making the reconstructed image look as much like the real input image as possible. This brings the final result of the discrimination process closer to 1 to fool the discriminator. The definition of $\mathcal{L}_{dis}$ may be found as follows

$$\begin{aligned}\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log(D(x))] + \\ E_{z\sim p_z(z)}[\log(D(1-D(G(z))))],\end{aligned} \tag{9}$$

where x is a real image and $G(z)$ is a generated image, the value of the latent variable z is chosen at random from a normal distribution during the training phase. Furthermore, by normalizing z, it is embedded in a continuous, bounded space without borders, limiting it to a unit hypersphere.

The reconstruction loss $\mathcal{L}_{RE}$ combine the pixel-level loss $\mathcal{L}_{pix}$ and the discrimination loss $\mathcal{L}_{dis}$ as

$$\mathcal{L}_{RE} = \frac{1}{T}(\mathcal{L}_{pix} + \mathcal{L}_{dis}). \tag{10}$$

After that, we are able to give the complete loss function of our method as

$$\mathcal{L} = \alpha\mathcal{L}_{RE} + \beta\mathcal{L}_{CM}, \tag{11}$$

where α and β are two hyper-parameters used to find a balance between the effects of $\mathcal{L}_{RE}$ and $\mathcal{L}_{CM}$, and the values of two parameters are chose randomly from $\{0.05, 0.1, 0.5, 1, 5, 10\}$.

C. Implementation Details

1) Visual Image Generator: The GAN architecture used in our FAA-GAN method mainly refers to [33], whose model is built on a framework and implementation that are publicly available[1]. The generator in our method is derived from the decoder network of the pretrained autoencoder, which has one linear and three deconvolutional layers, each with batch normalization and ReLU activation. The linear layer maps the image latent features to the first deconvolutional layer that expects 256 feature channels, and the three deconvolutional layers maps to 128, 64 and 1 feature channels. Kernel sizes are 4×4 and stride is 2. The hashing encode layer that follow the image encoder maps image feature channels to a 10-dimensional code.

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

2) Multi-Subject Functional Alignment: This module consists of S MLP networks, where S equals to the quantity of subjects for each dataset. Each MLP network consists of four fully connected (FC) layers, mapping the activity patterns at each time point into 1024, 256, 128 and 64 features, respectively. Each layer followed by ReLU activation functions. The regularized parameter ϵ is set to be 10^{-8}. Similar as the image generator network, we add a hash hidden layer, mapping the 64 feature into a 10-dimension code.

3) Reconstructed Image Discriminator: Three convolutional layers are consisted in the discriminator, followed by batch normalization and ReLU activations. Kernel sizes are 4×4 and stride is 2. The layers map from 1 to 64, 128 and 256 feature channels. After the convolutional layers, a linear layer maps the final activations to a single value reflecting the discriminator decision. In this paper, Adam [44] is used as the optimizer and we set the learning rate at 0.0005.

IV. Experimental Results

A. Datasets

Two publicly available datasets are used in this paper to verify the proposed FAA-GAN method, including, a) OpenNeuro dataset[2], and b) Handwritten character dataset[3]. Both of them are multi-subject task-based fMRI datasets. The properties of datasets used in our investigations are displayed in Table I, and more information is provided below.

a) OpenNeuro Dataset: In this paper, we select the Visual Object Recognition (DS105) dataset [1] from OpenNeuro platform for our empirical research. There are 8 categories of grayscale images, bottles, cats, chairs, faces, houses, scissors, shoes, and the meaningless pattern, respectively. The visual task included 6 subjects and each subject saw images from all the categories and underwent 12 runs. In our experiment, we down-sample the resolution of the stimuli images from 400×400 to 100×100, to improve the model's computational efficiency. Non-practical edge gradation is also set to 0. To confirm the reconstruction effect of real stimulus images, meaningless labels and resting time points are eliminated, and seven categories are reserved. Data preprocessing is done via the open source toolbox EasyfMRI[4]. Through EasyfMRI, we complete the main processing steps, including slice timing, smoothing, normalization, and anatomical alignment. The fusiform face area (FFA) and the parahippocampal place area (PPA) are two areas of the ventral temporal cortex that were developed specifically to represent different categories. In addition to the above steps, as our method is proposed for analyzing multi-subject data, temporal alignment is also need to ensure that at the same t-th time point, all subjects must perform the same type of cognitive task [26], [36].

The leave-one-subject-out cross-validation strategy is used for the effectiveness estimation of our multi-subject reconstruction method. Five subjects' data is used for training, whereas only one subject's data is used for testing. In the training phase, we use data consisting of 5 (subjects) × 120 (samples) × 7 (categories) = 4200 samples. And in the test phase, data from the test subject consisting of 1(subject) × 120 (samples) × 7 (categories) = 840 samples are used. For the convenience of observation, we output the visualization results into 12 groups. Each group consists of 10(samples) × 7(categories) = 70 samples. To avoid the contingency caused by random division, we conduct the experiment ten times and then determine the average results.

[2]http://openneuro.org

[3]http://sciencesanne.com/research/

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

TABLE I
Properties of the Datasets Used in the Experiments

Datasets	Instances	Subjects	Categories	Pixels	Voxels
DS105	5040	6	7	100×100	2294
Chars	1080	3	6	28×28	2420

b) Handwritten Character Dataset: Handwritten character dataset (Chars) contains 360 gray-scale handwritten character images (equal number of B's, R's, A's, I's, N's, and S's) [18]. Each side of the image has a resolution of 56×56 pixels, reduced to 28×28 for the sake of computation. Stimuli were shown in a Siemens Trio 3T MRI system with overall fixation (TR = 1.74s, TE = 30ms, voxel size = 2mm^3). To obtain an estimate that is closer to the true value of the BOLD response to individual character instances, the characters were repeated twice [18], [21], [33].

As with DS105's training and testing, we employ leave-one-subject-out cross validation. In each phase, two subjects' data are used for training purposes, but only one is used for testing at each level. In each phase, two subjects' data are used for training purposes, and another one is used for testing. In the test phase, we use data from the test subject consisting of 1(subject) × 60(samples) × 6(categories) = 360 samples. Similar to the DS105 dataset, we output the visualization results into 12 groups. Each group consists of 1(subject) × 5(samples) × 6(categories) = 30 samples. We repeat the procedure ten times and then take the average of the results to get the final result. This is done to get rid of the chance that comes from random division.

B. Comparison Methods

Our proposed method is compared with four deep learning methods that also used for reconstruction, including

Deep canonically correlated autoencoder (DCCAE) [45]: DCCAE uses two autoencoders to learn deep representations from different modalities of data. DCCAE primarily examines the same modal data reconstruction errors and bottleneck correlation, ignoring the reconstruction errors of different modalities of data.

Deep generative multiview model (DGMM) [21]: This deep learning model reconstructs the visual images based on neural responses. It's nearly like a nonlinear continuation of the BCCA. DGMM, like previous methods, discards the temporal information present in fMRI data.

Deep convolutional generative adversarial network (DCGAN) [33]: DCGAN uses a deep convolutional generative adversarial network to generate arbitrary images based on the stimulus images. DCGAN, on the other hand, does not take

TABLE II
RECONSTRUCTION RESULTS OF DIFFERENT SUBJECTS ON DS105 (CATEGORY=SHOES) BEFORE & AFTER ALIGNMENT

	Before Alignment			After Alignment		
	PCC↑	SSIM↑	Euc_dis↓	PCC↑	SSIM↑	Euc_dis↓
Subject 1	0.672±0.151	0.485±0.135	0.604±0.088	0.693±0.128	0.493±0.166	0.606±0.097
Subject 2	0.687±0.135	0.494±0.101	0.603±0.154	0.696±0.175	0.498±0.079	0.602±0.149
Subject 3	0.673±0.118	0.489±0.096	0.624±0.160	0.692±0.236	0.496±0.137	0.604±0.202
Subject 4	0.693±0.219	0.492±0.142	0.625±0.133	0.698±0.098	0.494±0.084	0.608±0.173
Subject 5	0.703±0.177	0.474±0.167	0.627±0.201	0.697±0.135	0.491±0.150	0.607±0.155
Subject 6	0.679±0.074	0.466±0.112	0.606±0.144	0.696±0.079	0.493±0.093	0.601±0.126
Average	0.685±0.146	0.483±0.126	0.615±0.147	0.695±0.142	0.494±0.118	0.605±0.150
Variance	$1.48e$-4	$1.22e$-4	$1.34e$-4	$5.47e$-6	$6.17e$-6	$7.87e$-6

into account the temporal information that is included in fMRI data, just like other approaches do not.

Temporal information guided generative adversarial network (TIGAN) [35]: TIGAN is our previous work, which is a GAN-based stimuli reconstruction method that takes temporal information into consideration. However, in this work, we don't think about the heterogeneity of response space among different subjects.

Shared autoencoder (SAE) [46]: SAE is used in this work to improve the reconstruction of natural images from fMRI voxels, utilizing a shared autoencoder to alternate encoding and decoding based on shared semi-supervised learning.

Instance-conditioned generative adversarial network (IC-GAN) [47]: IC-GAN reconstructs semantically accurate images with maintained low-level detailed information using fMRI patterns. A self-supervised learning model reconstructs target image instance features, while the noise vector uses inter-sample variance.

C. Evaluation Metrics

In this paper, we use three separate evaluation metrics, the Pearson's correlation coefficient (PCC), the structural similarity index (SSIM), and the Euclidean distance (Euc_dis), to analyze the efficacy of various methods for visual stimuli reconstruction. 1) PCC shows the correlation between the shown images and reconstructed ones. The larger this value is, the higher the correlation between the two images; 2) SSIM considers about the image texture when calculate the similarity, so it also reflects the perception of human brains partly [21]. From -1 to 1, the SSIM can take on a variety of values. The closer it is to 1, the more similar the two images are; 3) The distance between the original and reconstructed versions of an image is calculated via Euc_dis. As this value decreases, feature-space accuracy improves, meaning the reconstructed image more closely matches the original.

D. Experimental Results

1) Multi-Subject Correlation Analysis: At the beginning of this part, we conduct an experiment to analyze one of the main contributions of this paper, namely the effectiveness of the functional alignment module. In this part, we analyze the correlation between different subjects in our reconstruction task. As aforementioned, the heterogeneous response patterns across different subjects affect the stability and robustness of the results. So we conduct an experiment to verify that the functional alignment is valid.

As can be seen in Table II, all the subjects achieve acceptable reconstruction results on DS105 (category="shoes"). Comparing the performance before and after alignment, we can obtain a better average performance after alignment, which can verify the effectiveness of functional alignment. We also calculate the variance of different metrics among subjects, as shown in the last row of the table. For each metric, the variance after alignment is significantly smaller than before alignment. This also confirms that functional alignment can achieve more stable reconstruction performance on different subjects' neural data. Further, we calculate the correlation between the reconstruction results of different subjects, the result can be seen in Fig. 2. As is shown in the figure, after alignment, the correlation between subjects is higher. The correlation is in the range of $(0.50, 0.55)$ before alignment and increased to $(0.67, 0.72)$ after alignment.

2) Quantitative Analysis: Tables III and IV show the quantitative results of different methods on two datasets. The performance on DS105 are reported in Table III, which leads us to several observations below. First, our FAA-GAN outperforms the other deep learning-based comparison methods in terms of performance. Second, FAA-GAN outperforms DCCAE significantly. The benefits are likely due to GAN's greater generative ability, cross-modal retrieval, and multi-subject data alignment. Third, our method performs more evenly than DGMM and SAE. This may be due to functional alignment and the discriminant model's performance advantage over deep neural networks. Finally, cross-modal retrieval helps mine the image-response relationship compared with GAN-based

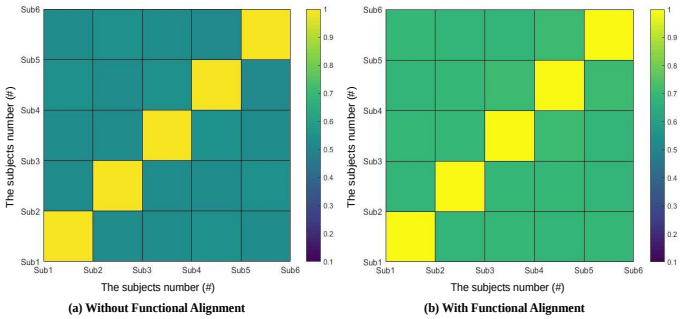

Fig. 2. Correlation of reconstruction results of different subjects on $DS105$ dataset (category="shoes"). (a) Results without functional alignment. (b) Results with functional alignment.

TABLE III
QUANTITATIVE PERFORMANCES OF COMPARED METHODS ON THE $DS105$ DATASET.

Model	PCC↑	p-value	SSIM↑	p-value	Euc_dis↓	p-value
DCCAE	0.594±0.133	$1.9474e$-21***	0.403±0.172	$3.1363e$-22***	0.741±0.136	$5.1284e$-21***
DGMM	0.638±0.098	$2.8877e$-19***	0.466±0.125	$1.9536e$-13***	0.649±0.079	$8.6753e$-16***
DCGAN	0.649±0.172	$8.3987e$-19***	0.481±0.083	$2.9422e$-12***	0.639±0.179	$1.0771e$-12***
TIGAN	0.686±0.157	$1.0893e$-7***	0.488±0.145	$4.1410e$-9***	0.612±0.116	$9.6304e$-5***
SAE	0.684±0.063	$1.5035e$-10***	0.489±0.038	$1.3009e$-7***	0.614±0.052	$4.0888e$-5***
IC-GAN	0.693±0.099	$3.8766e$-5***	0.493±0.065	$9.6802e$-5***	0.609±0.071	0.0212*
FAA-GAN (Ours)	**0.698±0.091**	—	**0.499±0.119**	—	**0.601±0.087**	—

↑: The higher the value is, the better performance the method get. ↓: The lower the value is, the better performance the method get.
The p-value of the improvement of FAA-GAN over the method: $*$ indicating ($p<0.05$), $**$ indicating ($p<0.01$), $***$ indicating ($p<0.001$).

TABLE IV
QUANTITATIVE PERFORMANCES OF COMPARED METHODS ON THE HANDWRITTEN CHARACTER DATASET.

Model	PCC↑	p-value	SSIM↑	p-value	Euc_dis↓	p-value
DCCAE	0.352±0.155	$2.1754e$-25***	0.185±0.174	$6.7282e$-33***	0.761±0.093	$9.8456e$-32***
DGMM	0.497±0.164	$4.9025e$-9***	0.339±0.058	$1.0150e$-12***	0.651±0.118	$3.0544e$-15***
DCGAN	0.496±0.125	$3.0680e$-9***	0.342±0.091	$4.3433e$-7***	0.644±0.174	$4.8757e$-12***
TIGAN	0.501±0.191	$1.8298e$-7***	0.341±0.076	$1.2857e$-7***	0.641±0.122	$3.4158e$-11***
SAE	0.503±0.093	$4.9853e$-5***	0.343±0.069	$5.5897e$-6***	0.638±0.105	$1.8247e$-8***
IC-GAN	0.499±0.079	$6.7078e$-7***	0.341±0.087	$2.0590e$-9***	0.642±0.077	$1.0261e$-10***
FAA-GAN (Ours)	**0.509±0.139**	—	**0.348±0.107**	—	**0.631±0.085**	—

↑: The higher the value is, the better performance the method get. ↓: The lower the value is, the better performance the method get.
The p-value of the improvement of FAA-GAN over the method: $*$ indicating ($p<0.05$), $**$ indicating ($p<0.01$), $***$ indicating ($p<0.001$).

reconstruction methods like DCGAN, TIGAN, and IC-GAN. Functional alignment also improves subject-to-subject consistency. Table IV shows the reconstruction performance for the handwritten character dataset. When comparing DCCAE and DGMM, we refer to the experimental settings in [21]. Similar to the DS105 dataset, the quantitative results for all three evaluative metrics are likewise better. We perform a pairwise t-test based on the experimental results, and the p-values and indications are shown in Tables III and IV. The resulting p-values show that significant performance improvement has been achieved.

3) Qualitative Analysis: Fig. 3 and Fig. 4 show the qualitative results of image reconstruction on the DS105 and the handwritten character datasets, respectively. In each figure, the top row represents the original images that were shown to the subjects during the experiments, while the following rows show the results of different compared reconstruction methods. Fig. 3 shows DS105 (category = "shoes") visualization reconstructing results. The example demonstrates that compared to the other methods, our FAA-GAN yields superior reconstructions, especially for natural images. Inaccurate contour description is a limitation of DCCAE and DGMM. Their inability to reconcile the original and reconstructed images is one possible explanation. The three GAN-based algorithms outperform DCCAE and DGMM; however, they are not as good as FAA-GAN. As for SAE, although acceptable reconstructed images are generated, the results still don't match our FAA-GAN in some details.

Fig. 4 depicts the visualization results of the reconstructed handwritten characters. The figure shows how closely the reconstructed characters resemble the originals. The performance of DCCAE does not match our approach. Complex

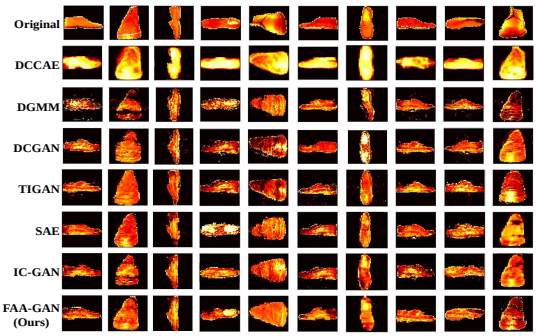

Fig. 3. Qualitative performances of compared methods on the $DS105$ dataset (category = "shoes").

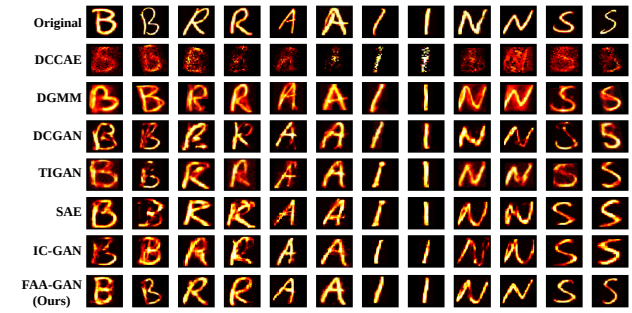

Fig. 4. Qualitative performances of compared methods on the handwritten character dataset.

noises have an impact on reconstruction results, which frequently lack the core properties of the original images. Further, the other compared methods also produce imprecise reconstruction performance. Despite producing better results than DCCAE and close to each other, these methods lose some detailed information.

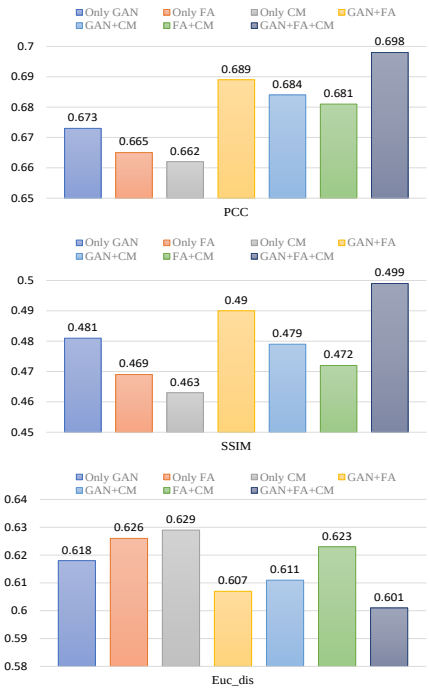

Fig. 5. Reconstruction results on the $DS105$ dataset with/without different components in our model.

4) Ablation Study: As mentioned above, there are three key components in our method, i.e., the multi-subject functional alignment module, the cross-modal retrieval module, and the GAN-based reconstruction module, respectively. Herein, an ablation study is conducted to determine the effectiveness of each component of our strategy and, by extension, its contribution to the model's overall performance.

With one module: With one module means only one component was used to reconstruct the images. Only GAN means that there is no functional alignment across subjects, and cross-modal hashing retrieval also be replaced by Bayesian inference in [21]. Only FA means that there is no cross-modal retrieval and no discriminator loss, only pixel-wise loss is retained to complete the reconstruction. Only CM refers to no functional alignment and GAN architecture in the method.

With two modules: Here, two of the three components are included to evaluate the impact on our model, while the rest component is absent. 1) "GAN+FA": means "No CM". We only learn a linear regression between the neural representation and the image latent features. At the same time, the effect in the absence of β can be verified; 2) "GAN+CM": means "No FA". To explore the influence of the multi-subject functional alignment module, we disregard the heterogeneity among subjects and don't do the functional alignment; 3) "FA+CM": means "No GAN". We replace the GAN with the pre-trained autoencoder, and there is no discriminator. At the same time, the effect in the absence of α can be verified.

With three modules: With three modules means that we use all the components in our method, that is to say, we use the complete model to compare with the circumstances above to show the ablation study results.

Fig. 5 shows the reconstruction performance on the DS105 dataset under the different circumstances we set. We can mainly observe potential trends and draw conclusions from the figure below. First, comparing the three main components in our method, GAN contributes more to the final performance, which may be due to its generative ability, which can generate images similar to the inputs. Second, the more components the model has, the better results can be obtained. Last but not least, the best performance is achieved by FAA-GAN, which integrates all three components, indicating that the entire procedure helps to enhance reconstruction accuracy.

TABLE V
RECONSTRUCTION RESULTS OF DIFFERENT FUNCTIONAL ALIGNMENT METHODS ON DS105

	PCC↑	SSIM↑	Euc_dis↓
wo-alignment	0.649±0.227	0.449±0.073	0.635±0.092
HA [11]	0.675±0.084	0.468±0.119	0.619±0.219
SRM [36]	0.682±0.146	0.481±0.060	0.612±0.164
SHA [26]	0.688±0.099	0.480±0.144	0.609±0.085
DHA [25]	**0.698±0.091**	**0.499±0.119**	**0.601±0.087**

5) Comparison of Different Functional Alignment Methods: As one of the important module of our proposed method, different functional alignment methods may lead to quite different influence to the final reconstruction results. In this paper, we use several different functional alignment methods [11], [25], [26], [36] and compared the performance on DS105.

Table V illustrates that the reconstruction outcome without alignment (wo-alignment) is not as satisfying as the techniques with alignment, indicating that our functional alignment strategy in the reconstruction model succeeded. Because they are the initial versions of functional alignment and the updated approaches produce superior outcomes, HA and SRM are slightly less expensive than other alignment methods. As a supervised method, SHA may have failed to play an optimal function in our unsupervised reconstruction technique. Finally, the nonlinear fitting ability of learned parameters of DHA, which we applied in this work, resulted in improved outcomes.

V. DISCUSSION

In this part, we first analyze the computational complexity of different deep learning-based reconstruction methods, including space complexity and time complexity. Then, we evaluate the effects of regularization parameters in our model. Finally, we highlight our work's weaknesses and propose directions for additional field research.

A. Analysis of Computational Complexity

In this part, we analyze the computational complexity of our visual image reconstruction task. Two main indexes are considered, i.e., space complexity and time complexity, respectively. For space complexity, the network structure of our model has been described in the preceding part (See Implementation Details). For time complexity, the runtime of the proposed method is compared with the previous methods using the DS105 dataset. Fig. 6 illustrates the runtime of all the methods, where the runtime of other methods is scaled based on the FAA-GAN (the runtime of the proposed method is considered

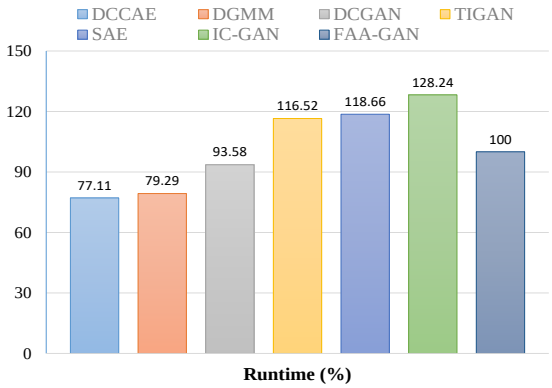

Fig. 6. Runtime analysis of different methods.

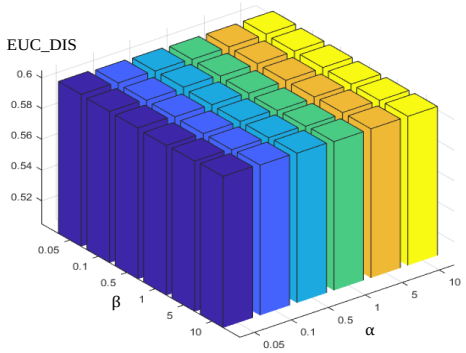

Fig. 7. Reconstruction results (Euc_dis) of the DS105 dataset vis different values of α and β.

the unit). As depicted in this figure, DCCAE and DGMM spend less time because of their simple architecture compared with the improved versions. Then, IC-GAN generated the longest runtime because it is proposed based on multi-layer BigGAN, and as it is used for generating colorful images, it considers the RGB channels. Further, for the other four methods, including FAA-GAN, the runtime is about the same and acceptable.

B. Effects of Regularization Parameters

In our proposed method, the optimized objective function is mainly composed of two parts, the reconstruction loss $\mathcal{L}_{RE}$ and the cross-modal focal loss $\mathcal{L}_{CM}$. In order to balance the two loss parts and control their influence in the model, we set two hyperparameters, α and β. We conduct image reconstruction experiments on the DS105 dataset with different regularization parameters chosen from $\{0.05, 0.1, 0.5, 1, 5, 10\}$, and the results are displayed in Fig. 7.

As can be seen from the figure, we can observe that with different parameters, our proposed FAA-GAN method can obtain relatively stable reconstruction results. And the best regularization parameter can be chosen from $\alpha = 1$ and $\beta = \{1, 5\}$, where FAA-GAN achieves better results.

C. Limitations and Future Work

In our opinion, the current work still has several shortcomings that need to be addressed. First, the solution of FAA-GAN being presented is made up of three main modules, which will make the demand on memory for visual reconstruction greater. As a result, model compression is a significant area that should be focused on for practical applications. Second, issues with data collection meant that the task-based fMRI datasets used here had a relatively small sample size. Several transfer learning-based methods have been proposed [48]–[50]. For large sample sizes or multi-site fMRI datasets, this is a barrier that needs to be overcome before algorithms can be applied. Third, the method described in this paper doesn't effectively utilize whole-brain structural data. In future research, we will construct information-based methods on whole-brain structural data. It highlights the information-valid area in brain data and improves visual image reconstruction. At last, the functional alignment module in FAA-GAN assumes that the training set's stimuli are already aligned in time, which needs additional time to align the testing subject to the learned common space. In the future, in order to improve time efficiency, we will consider developing a framework that doesn't need temporal alignment to adapt to more application scenarios.

VI. Conclusion

This paper proposes FAA-GAN, a multi-subject visual image reconstruction method. Our FAA-GAN method reconstructs visual images based on neural activity patterns via GAN architecture, and multi-subject functional alignment is taken into account. The GAN module, the functional alignment module, and the cross-modal retrieval module are the three major modules in our method. The proposed FAA-GAN not only provides an image reconstruction model that mines the relationship between visual stimulus and elicited brain activities, but it also addresses the issue of subject heterogeneity. Cross-modal hashing retrieval is also introduced to compute the association between different modalities of data and increase cross-modal retrieval accuracy. Through the experiments on Visual Object Recognition and handwritten character datasets, we have seen that our FAA-GAN can get better results than the state-of-the-art 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 News*, vol. 502, no. 7472, p. 428, 2013.

[3] L. Gao, J. Yu, L. Zhu, S. Wang, J. Yuan, G. Li, J. Cai, X. Qi, Y. Sun, and Y. Sun, "Dynamic reorganization of functional connectivity during post-break task reengagement," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 157–166, 2022.

[4] G. Gaziv, R. Beliy, N. Granot, A. Hoogi, F. Strappini, T. Golan, and M. Irani, "Self-supervised natural image reconstruction and large-scale semantic classification from brain activity," *NeuroImage*, p. 119121, 2022.

[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, p. 352, 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, pp. 435–456, 2014.

[7] X. Ma, S. Qiu, and H. He, "Time-distributed attention network for EEG-based motor imagery decoding from the same limb," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 496–508, 2022.

[8] 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.

[9] G. N. Dimitrakopoulos, I. Kakkos, Z. Dai, J. Lim, J. J. deSouza, A. Bezerianos, and Y. Sun, "Task-independent mental workload classification based upon common multiband EEG cortical connectivity," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1940–1949, 2017.

[10] Usman, Ayub, Sheikh, Manuel, Carreiras, David, and Soto, "Decoding the meaning of unconsciously processed words using fMRI-based MVPA," *NeuroImage*, 2019.

[11] J. V. Haxby, J. S. Guntupalli, A. C. Connolly, Y. O. Halchenko, B. R. Conroy, M. I. Gobbini, M. Hanke, 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.

[12] A. Lorbert and P. J. Ramadge, "Kernel hyperalignment," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[13] M. Yousefnezhad and D. Zhang, "Local discriminant hyperalignment for multi-subject fMRI data alignment," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 59–65.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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, 2013.

[19] 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.

[20] 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.

[21] 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, 2018.

[22] 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, 2019.

[23] 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*, vol. 13, no. 4, pp. 827–840, 2020.

[24] H. Xu, A. Lorbert, P. J. Ramadge, J. S. Guntupalli, and J. V. Haxby, "Regularized hyperalignment of multi-set fMRI data," in *2012 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2012, pp. 229–232.

[25] M. Yousefnezhad and D. Zhang, "Deep hyperalignment," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[26] M. Yousefnezhad, A. Selvitella, L. Han, and D. Zhang, "Supervised hyperalignment for multisubject fMRI data alignment," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 3, pp. 475–490, 2020.

[27] P.-H. Chen, J. S. Guntupalli, J. V. Haxby, and P. J. Ramadge, "Joint SVD-hyperalignment for multi-subject fMRI data alignment," in *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2014, pp. 1–6.

[28] 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.

[29] H. Zheng, L. Yao, M. Chen, and Z. Long, "3D contrast image reconstruction from human brain activity," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 12, pp. 2699–2710, 2020.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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 Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 1104–1118, 2022.

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

[37] P.-H. Chen, X. Zhu, H. Zhang, J. S. Turek, J. Chen, T. L. Willke, U. Hasson, and P. J. Ramadge, "A convolutional autoencoder for multi-subject fMRI data aggregation," *arXiv preprint arXiv:1608.04846*, 2016.

[38] J. C. Gower and G. B. Dijksterhuis, *Procrustes problems*. OUP Oxford, 2004, vol. 30.

[39] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[40] Y. Cao, B. Liu, M. Long, and J. Wang, "Cross-modal hamming hashing," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 202–218.

[41] C. Li, S. Gao, C. Deng, W. Liu, and H. Huang, "Adversarial attack on deep cross-modal hamming retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2218–2227.

[42] X. Xu, J. Tian, K. Lin, H. Lu, J. Shao, and H. T. Shen, "Zero-shot cross-modal retrieval by assembling autoencoder and generative adversarial network," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 1s, pp. 1–17, 2021.

[43] E. Yu, J. Ma, J. Sun, X. Chang, H. Zhang, and A. G. Hauptmann, "Deep discrete cross-modal hashing with multiple supervision," *Neurocomputing*, vol. 486, pp. 215–224, 2022.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[45] 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.

[46] "Reconstructing natural images from human fMRI by alternating encoding and decoding with shared autoencoder regularization," *Biomedical Signal Processing and Control*, vol. 73, p. 103397, 2022.

[47] 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 *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8.

[48] 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.

[49] 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.

[50] T. M. Yousefnezhad, A. Selvitella, D. Zhang, A. Greenshaw, and R. Greiner, "Shared space transfer learning for analyzing multi-site fMRI data," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 990–16 000, 2020.