EEG-Driven 3D Object Reconstruction with Style Consistency and Diffusion Prior

Xin Xiang
School of Computer Science and Technology
Hangzhou Dianzi University
Hangzhou

Wenhui Zhou
School of Computer Science and Technology
Hangzhou Dianzi University
Hangzhou

Guojun Dai
School of Computer Science and Technology
Hangzhou Dianzi University
Hangzhou

November 19, 2024

ABSTRACT

Electroencephalography (EEG) -based visual perception reconstruction has become an important area of research. Neuroscientific studies indicate that humans can decode imagined 3D objects by perceiving or imagining various visual information, such as color, shape, and rotation. Existing EEG-based visual decoding methods typically focus only on the reconstruction of 2D visual stimulus images and face various challenges in generation quality, including inconsistencies in texture, shape, and color between the visual stimuli and the reconstructed images. This paper proposes an EEG-based 3D object reconstruction method with style consistency and diffusion priors. The method consists of an EEG-driven multi-task joint learning stage and an EEG-to-3D diffusion stage. The first stage uses a neural EEG encoder based on regional semantic learning, employing a multi-task joint learning scheme that includes a masked EEG signal recovery task and an EEG based visual classification task. The second stage introduces a latent diffusion model (LDM) fine-tuning strategy with style-conditioned constraints and a neural radiance field (NeRF) optimization strategy. This strategy explicitly embeds semantic- and location-aware latent EEG codes and combines them with visual stimulus maps to fine-tune the LDM. The fine-tuned LDM serves as a diffusion prior, which, combined with the style loss of visual stimuli, is used to optimize NeRF for generating 3D objects. Finally, through experimental validation, we demonstrate that this method can effectively use EEG data to reconstruct 3D objects with style consistency.

1 Introduction

A noninvasive brain-computer interface (BCI) system typically controls electronic devices through voluntary modulation of EEG signals. However, most current studies investigating the relationship between brain activity and 3D object visual imagery tasks primarily rely on functional magnetic resonance imaging (fMRI). For example, Mind-3D [**?**] successfully decoded 3D visual information from the brain using fMRI signals, demonstrating the feasibility of this challenging task. Therefore, existing fMRI studies provide theoretical foundations and research support for exploring different neuroimaging techniques and understanding the neural modulation in 3D object imagery tasks, especially in the context of EEG-based decoding of 3D visual information from the brain. A large body of neuroscientific experimental research [1, 2, 3, 4, 5, 6, 7] has demonstrated that the brain can rapidly perceive 2D images and 3D objects in extremely short
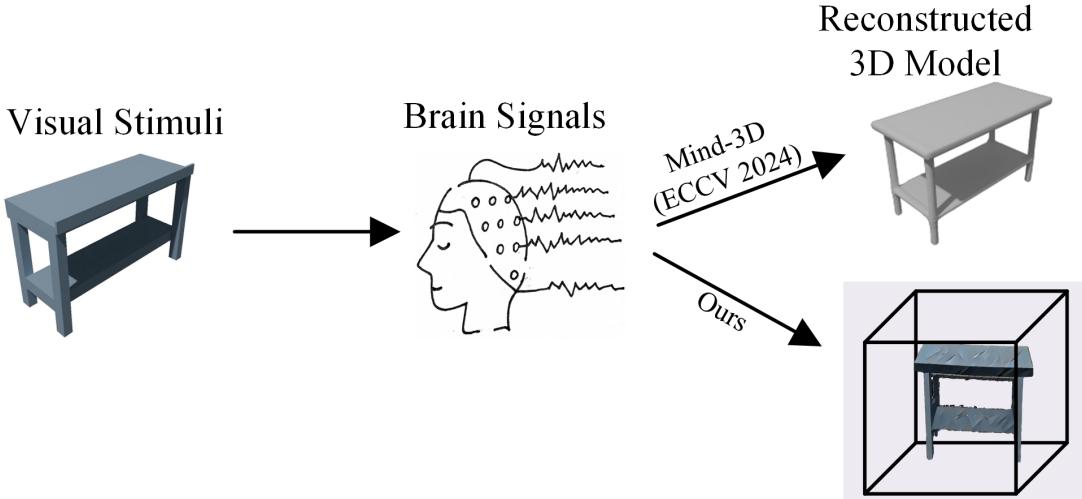
Figure 1: Our model reconstructs 3D objects using fMRI signals.

periods of time. One of the primary methods for studying human visual perception is using deep neural networks to reconstruct visual content that evokes subjective responses in stimulus experiments. Many studies [8, 9, 10, 11] have attempted to reconstruct visual information based on functional magnetic resonance imaging (fMRI). However, the collection of fMRI data requires expensive equipment, limiting its widespread use in practical applications. In contrast, EEG is a more cost-effective technique for capturing brain activity and is easier to collect. EEG data are typically recorded as a series of time-series electrophysiological signals by placing electrodes on the scalp. During this process, subjects are presented with stimulus images, while brain signals are simultaneously recorded.

Recently, several studies have explored human visual perception learning based on EEG signals. For example, [12] used traditional generative models to convert EEG signals into images, while [13] and [14] utilized latent diffusion models (LDM) to extract latent features for reconstructing visual stimulus images corresponding to EEG signals, achieving better alignment between EEG signals and 2D images. However, these attempts still have limitations in terms of pixel-level and semantic fidelity, and no research to date has utilized EEG signals to reconstruct color-consistent 3D objects. One reason for this is the difficulty in effectively capturing semantic information, and another is the relative complexity of the learning process in generative models. As a result, the reconstructed images often lack accurate perceptual information and struggle to achieve precise prediction and control of structural details.

Building upon the aforementioned neuroscientific theories and technological advancements, it can be inferred that reconstructing high-quality visual information from brain activity is feasible. In this study, we propose a method for reconstructing color-consistent 3D objects from EEG signals. To address the challenges of this task, we jointly train EEG signals with Ground Truth (GT) images. Specifically, we first train an implicit neural EEG encoder with the capability of perceiving 3D objects, allowing it to capture regional semantic features. Then, based on the latent EEG codes obtained in the first stage, we integrate a diffusion model, neural style loss, and NeRF to implicitly decode the 3D objects. Finally, through experimental validation, we demonstrate that our method is capable of reconstructing color-consistent 3D objects using EEG signals.

In order for the implicit neural EEG encoder to capture regional semantic features, in the first stage, we employ joint training for reconstruction and classification using EEG signals, enabling the EEG encoder to learn the regional semantic features of EEG. The reconstruction task aims to reconstruct the EEG signals, learning the temporal information associated with specific regions, while the semantic classification task is used to classify the semantic features of the EEG signals, training the encoder to recognize semantic characteristics. Through the joint learning of these two tasks, the proposed method can effectively capture regional semantic features.

In order for the implicit neural decoder to decode 3D objects, in the second stage, this paper integrates latent EEG codes, a diffusion model, and NeRF to decode 3D objects. The latent EEG codes guide the diffusion model to generate novel 2D views, and a style loss is used to transfer the colors from GT, ensuring that the colors of the novel views remain as consistent with the GT as possible. Subsequently, the novel 2D views from different perspectives are used to optimize NeRF. Unlike previous text-to-3D methods [15, 16], the proposed approach focuses on reconstructing color-consistent 3D objects based on latent EEG encodings.

The main contributions of this paper are as follows:

- We propose an EEG-based 3D object reconstruction framework with style-consistent semantic region awareness for reconstructing 3D objects that are consistent with the style of visual stimuli. The framework consists of an EEG multi-task joint learning stage and a style-semantic region-aware LDM fine-tuning and NeRF optimization stage. The former focuses on learning semantic- and location-aware latent EEG codes from EEG signals, while the latter uses these learned latent EEG codes as conditions to fine-tune LDM. The fine-tuned LDM serves as a diffusion prior, which, in combination with a style loss, optimizes NeRF and ultimately reconstructs style-consistent 3D objects.

- We design a neural EEG encoder based on regional semantic learning, employing a multi-task joint learning scheme that includes a masked EEG signal recovery task and an EEG-based visual classification task. According to the visual attention shifting mechanism, the fixation regions of the human eye change over time. Thus, the recovery task helps the EEG encoder learn the spatial location information of objects by reconstructing missing EEG data, while the classification task aids in learning the semantic information of object regions.

- This paper proposes a fine-tuning strategy for LDM with style-conditioned constraints and NeRF optimization. The strategy explicitly embeds semantic- and location-aware latent EEG codes and incorporates visual stimulus maps to fine-tune the LDM. The fine-tuned LDM serves as a diffusion prior, which, combined with the style loss of visual stimuli, is used to optimize NeRF. Finally, EEG data is employed to reconstruct 3D objects with color consistency.

2 Related Work

EEG-Based Image Synthesis   Over the past few years, text-to-image generation models have developed rapidly. For example, diffusion models have made significant progress in this field, as they are capable of extracting complex latent semantic features from text descriptions and generating high-quality object and scene images [17, 18, 19, 20, 13]. In this paper, we fine-tune the diffusion model using EEG-ImageNet [1] and Things-EEG2 [21] dataset to achieve the task of generating 2D images from EEG signals. For instance, DreamDiffusion [13] fine-tunes the diffusion model through global semantic alignment, but there still remains a certain gap between the generated images and the actual images. To enhance the realism of the generated images, the proposed method combines EEG signal reconstruction and classification to help the model capture regional semantic features, enabling the diffusion model to better learn the regional mapping relationship between EEG signals and images.

EEG-to-3D Object Generation   In recent years, significant progress has been made in text-to-3D object generation models [22, 15, 16] , but EEG-to-3D object generation has not yet been fully explored. Inspired by text-to-3D methods such as [15] and [16], this paper attempts to utilize latent EEG codes to guide NeRF in reconstructing 3D objects through the shape priors of a diffusion model. To further investigate the brain's ability to perceive color visual information from EEG, and drawing upon content and style transfer methods such as [23] and [24], we employ content and style losses to ensure that the reconstructed 3D objects maintain style consistency with ground truth (GT) images. Using this approach, we have, for the first time, achieved the reconstruction of style-consistent 3D objects from EEG signals, indirectly validating the neuroscientific theory that the human brain can perceive various types of visual information, such as color, shape, and texture, when observing objects.

3 The Neuroscientific Analysis of Our Method

Our dataset is sourced from [1], where each image is displayed for 0.5 seconds, during which EEG is collected simultaneously. Based on references [2, 25, 3, 6], it is known that the brain is capable of acquiring visual information within 0.5 seconds. Therefore, we hypothesize that EEG has already perceived specific 3D texture information within this 0.5-second window. Our work proposes this hypothesis and experimentally verifies its existence. To analyze how the brain captures visual perceptual information in such a short timeframe, we adopt a methodology incorporating 3D and color perception. This approach facilitates a more comprehensive explanation and understanding of the brain's visual perception process. Not only does it aid researchers in exploring perceptual mechanisms, but it also advances theoretical research pertaining to vision.

3.1 3D Perception

[26] conducted a series of experiments, including multiple participants who participated in two offline and three online sessions. These experiments successfully demonstrated the feasibility of distinguishing the neural correlates of imagined 3D objects in EEG. Specifically, they decoded the participants' imagined spheres, cones, pyramids, cylinders, and cubes. [27] discovered evidence of neural decoding for complex 3D object shapes. It utilized an evolutionary
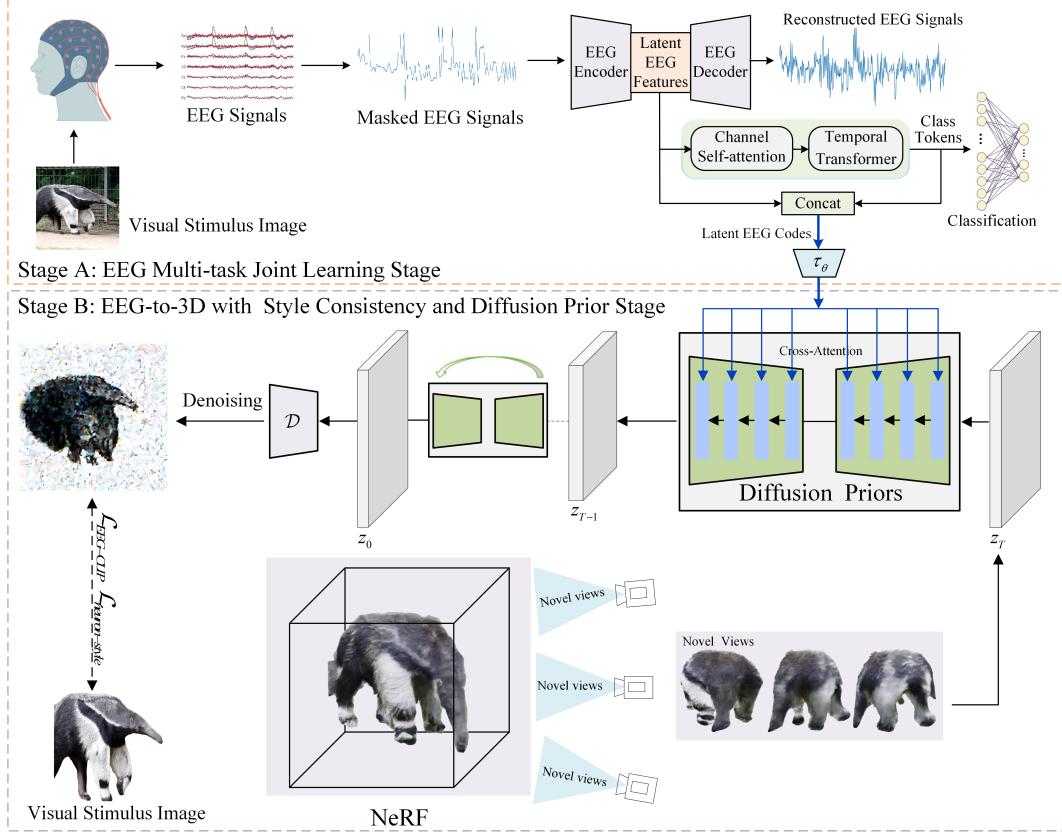
Figure 2: The overall network architecture of the proposed method.

stimulus strategy and linear/nonlinear response models to characterize responses to 3D shapes. The configuration representation of 3D shapes can provide specific knowledge about object structures, supporting guidance for complex physical interactions.

3.2 Color Perception

According to reference [29], using techniques such as intrinsic signal optical imaging, two-photon imaging, and EEG recording, the study detailed the hue map structures of different visual brain areas, revealing the neural mechanisms underlying the formation of color perception space. It was found that as the processing hierarchy of the visual cortex increases, the neurons encoding color in the brain gradually coordinate, ultimately forming a balanced mechanism that matches our subjective perception of hues.

These studies demonstrate that brain signals can perceive and process various visual information, such as color, shape, and texture.

4 Methodology

4.1 Overview

As shown in Figure 2, to reconstruct style-consistent 3D objects from EEG signals, we propose an EEG-to-3D architecture with style consistency and semantic awareness. The proposed method mainly consists of two components: 1) a semantic-aware neural EEG encoder for 3D objects, and 2) a style-consistent neural decoder for 3D objects. First, we utilize EEG signals for neural encoding of 3D objects. In this stage, we obtain latent EEG features and class tokens through joint training tasks involving EEG signal reconstruction and classification, and these two components are merged and passed through a linear layer to generate latent EEG codes. Next, we use latent EEG codes obtained in the previous stage to provide conditional decoding features for the reverse diffusion process of the diffusion model via a cross-attention mechanism, generating novel 2D views while using content and style losses to ensure that the style

of the novel 2D views remain as consistent as possible with GT. Finally, we use latent EEG codes to initialize LDM to generate novel 2D views from different perspectives to optimize NeRF, enabling the neural decoder to decode 3D objects.

4.2 Stage A: Multi-Task Joint Learning

During the initial phase, the process leverages EEG signals for neural encoding of 3D objects. Specifically, we first extract the regional semantic features of EEG by using a masked reconstruction model to capture the temporal features of EEG, thereby obtaining its regional information. Next, we perform a semantic classification task on EEG signals, enabling EEG encoder to learn semantic features. Through the EEG reconstruction task and semantic classification task, we capture the semantic features of regions, which serve as input for subsequent steps. Then, we explicitly introduce the semantic regional features of the original and reconstructed images by leveraging latent EEG codes containing latent information of semantic regions. We jointly fine-tune LDM to incorporate the performance of semantic regions, making the generated images more similar to GT in terms of regional and spatial relationships.

4.2.1 Reconstruction and Classification Tasks of EEG

EEG Reconstruction Task   Traditional models struggle to effectively extract meaningful features due to the complex spatial regional information characteristics of EEG. However, as demonstrated by the work and ideas of [30], it is possible to capture valuable information from the context of EEG signals. Therefore, we use a masked model that randomly masks portions of EEG signals and then reconstructs them to achieve this purpose. Combining the findings of [15] and [30], we mask a certain proportion of EEG signals based on their temporal features and use [31] method to convert EEG into 1D data to embed it into the network. By considering the contextual temporal cues to predict the missing signals, we can learn the regional information.

EEG-based Visual Classification Task   Prior studies have reconstructed EEG signals using an EEG encoder [30, 13]. However, obtained latent EEG codes currently only encompass temporal and spatial characteristics, lacking crucial semantic features. Due to the lengthy temporal nature of EEG signals, establishing global relationships for EEG data classification poses a challenge. Therefore, we employ a Temporal Transformer to correlate time-series features with their own features, extracting global features of EEG over the time sequence. This allows for the extraction of temporal features from EEG signals, classification of EEG signals, and subsequent acquisition of class tokens. By integrating these features with the output from the EEG encoder, we aim to further enhance the accuracy of semantic classification.

4.3 Stage B: Diffusion Prior and Style Consistency

As Stage B, we utilize the reference view x generated by EEG and latent EEG codes to reconstruct NeRF, and constrain the novel view through a diffusion prior conditioned on the latent EEG codes. The proposed EEG-to-3D method is expected to simultaneously meet the following requirements: 1) It can generate 3D objects using EEG that closely resemble the rendering appearance of the reference view x; 2) Novel view renderings should exhibit semantic consistency with the reference view x, while also maintaining style consistency; 3) The generated 3D objects should possess well-defined geometric structures.

Text-to-3D   DreamFusion [15] demonstrated its capabilities in the field of text-to-3D synthesis by leveraging a pretrained text-to-image diffusion model [18] as a strong image prior. DreamFusion achieves text-to-3D generation through two key components: a pretrained text-to-image diffusion-based generative model and a neural scene representation of the scene model. The scene model is a parametric function $x = g(\theta)$, which generates an image s at the specified camera pose. Here, g is a volumetric renderer, and θ is a coordinate-based multi-layer perceptron (MLP) representing a 3D volume. The diffusion model ϕ includes a learned denoising function $\epsilon_\theta(x_t; y, t)$, which predicts the sampled noise ϵ given the noisy image x_t, noise level t, and text embedding y. It provides the gradient direction to update θ such that all rendered images are pushed toward high-probability density regions conditioned on the text embedding under the diffusion prior. Specifically, DreamFusion introduces Score Distillation Sampling (SDS) to compute the gradient.

$$\nabla_\theta \mathcal{L}_{SDS}(\phi, g(\theta)) = \mathbb{E}_{t,\varepsilon}\left[\omega(t)(\epsilon_\theta(z_t; y, t) - \epsilon)\frac{\partial x}{\partial \theta}\right] \tag{1}$$

Here, $\omega(t)$ is a weighting function. We regard the scene model g and the diffusion model as modular components within the framework, with ϵ_θ serving as the denoising function.

EEG-to-3D with Diffusion Prior   To ensure the semantic coherence of the EEG-to-3D objects, we adopt a diffusion prior to impose additional constraints on the novel view renderings. Previous works on text-to-3D generation [15, 16] have applied $\mathcal{L}_{SDS}$ to leverage text-conditioned diffusion models as 2D diffusion priors. In the method proposed in this paper, we employ a diffusion prior conditioned on latent EEG codes to optimize the views generated by NeRF, gradually refining them from blurry to sharp. In this case, to apply $\mathcal{L}_{SDS}$, we use latent EEG codes obtained in the first stage as the latent prompt y, allowing us to perform $\mathcal{L}_{SDS}$ within the latent space of the diffusion model:

$$\nabla_\theta \mathcal{L}_{SDS}(\phi, g(\theta)) = \mathbb{E}_{t,\varepsilon}\left[\omega(t)(\epsilon_\theta(z_t; y, t) - \epsilon)\frac{\partial z}{\partial x}\frac{\partial x}{\partial \theta}\right] \quad (2)$$

here, y represents latent EEG codes, z_t denotes the noisy latent representation, $\frac{\partial z}{\partial x}$ refers to the gradient of the LDM encoder, and $\frac{\partial x}{\partial \theta}$ represents the gradient of the rendered image.

EEG-Image Information Aligning with NeRF   According to $\mathcal{L}_{SDS}$ in DreamFusion, this function is used to measure the similarity between novel views and text prompts. However, in this work, 3D objects are generated based on EEG signals, thus requiring further alignment between the EEG signals and novel views from different perspectives. To achieve this, we leverage a pretrained CLIP image encoder to align the EEG signals with the novel views. The corresponding loss function is as follows:

$$\mathcal{L}_{EEG-CLIP} = -E_{CLIP}(T/I) \cdot \rho(\varphi_{latent}(y)) \cdot E_{CLIP}(g(\theta)) \quad (3)$$

Here, $E_{CLIP}(T/I)$ includes the CLIP image encoder. By adding a similarity loss to the novel view image $g(\theta)$ obtained under the diffusion prior of EEG, our goal is to align the novel view image with the reference image.
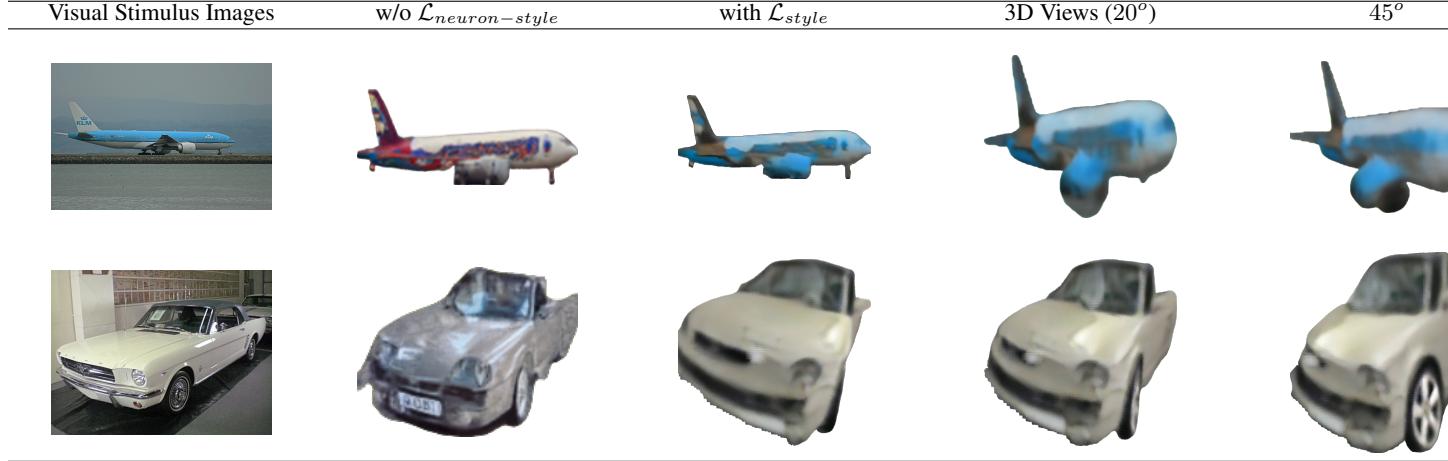
Visual Stimulus Images	w/o $\mathcal{L}_{neuron-style}$	with $\mathcal{L}_{style}$	3D Views (20°)	45°

Table 1: Typical 3D objects generated by EEG-driven models based on the EEG-ImageNet dataset [1].

Neural Style Transfer   As shown in Figure 3, the goal of neural style transfer is to transfer the style I_s of GT onto the content image I_c, generating a stylized image I_o, while ensuring that the reconstructed image preserves the color realism of the GT and the content remains unchanged. According to [32, 33], adding a new loss term to the optimization objective enhances the realism of stylized images produced by the neural style transfer algorithm, particularly in preserving the local structures of the content image. Building upon these references, we incorporate both style loss and content loss in the style transfer process. We utilize a pretrained VGG network to extract features from the target image, which are used to compute the content and style losses. By combining these two losses, the network generates new images that retain the content while adopting a specific style. To achieve the aforementioned style transfer, we use the pretrained VGG feature extractor $\phi(\cdot)$ to extract content and style features, denoted as $F_c = \phi(I_c)$ and $F_s = \phi(I_s)$, respectively. The output image I_o is optimized using the following objective function, which includes content loss $\mathcal{L}_{content}$ and style loss $\mathcal{L}_{style}$:

$$\mathcal{L}_{neuron-style} = argmin\{\mathcal{L}_{content}(\varsigma(I), F_c) + \lambda\mathcal{L}_{style}(\varsigma(I), F_s)\} \quad (4)$$

here, $\varsigma(\cdot)$ represents the feature extractor, and λ is the balancing factor between the content loss and style loss.

Loss Function with Diffusion Prior and Style Consistency   The overall loss in the second stage can be represented by $\mathcal{L}_{SDS}$, $\mathcal{L}_{EEG-CLIP}$, and $\mathcal{L}_{neuron-style}$. In this work, the diffusion prior generated under the condition of latent

6

Visual Stimulus Images	with $\mathcal{L}_{neuron-style}$	3D Views (20°)	45°	70°

Table 2: The method proposed in this study reconstructs 3D objects with typical style consistency using the Things-EEG2 [21] and fMRI-Image [?] datasets, respectively.

EEG codes is used to guide NeRF in generating 3D objects with consistent style. $\mathcal{L}_{SDS}$ is used to optimize NeRF to achieve better geometric shapes and details, $\mathcal{L}_{EEG-Image}$ is employed to align the EEG, text, and image, and $\mathcal{L}_{style}$ is used to adjust the reconstructed style of the 3D objects to ensure they remain as consistent as possible with GT color, thereby enhancing the realism and visual accuracy of the 3D reconstruction.

Loss Function between EEG and Latent Diffusion Models   Under the combined influence of reconstruction and classification based on EEG signals, an EEG encoder with temporal, spatial, and semantic features is obtained. Using this EEG encoder along with semantic embedding, we map EEG signals to a latent space constrained by semantic regions. We integrate cross-attention mechanisms and latent conditional features t, and utilize the denoising U-Net to replace the original temporal pixel space features in LDM with conditional latent features z_t, thereby optimizing the loss function. This forms the conditional loss function for LDM:

$$\mathcal{L}_{ldm} = E_{z,\varepsilon \sim\ N(0,1),t} \left[\|\varepsilon - \varepsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2 \right] \tag{5}$$

To endow LDM with regional semantic performance, we utilize the pre-trained Segment Anything (SAM) [?] to obtain the regional semantic maps of visual stimulus images and their corresponding reconstructed images. The regional
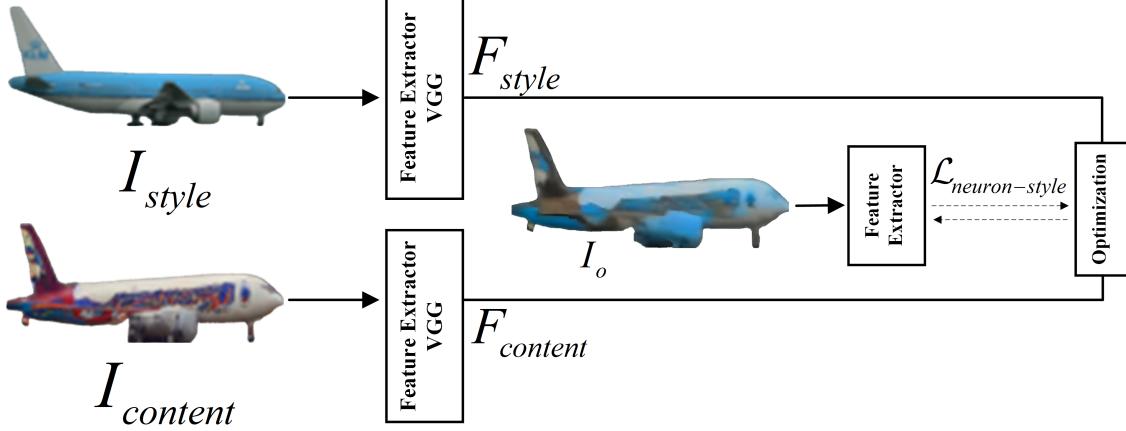
Figure 3: Neural style transfer. In Stage B of the proposed method, a color transfer loss is incorporated.

semantic loss is calculated by the cross-entropy loss function as follows,

$$\mathcal{L}_{region} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{M} p_{i,k} \cdot \log\left(\hat{p}_{i,k}\right) \tag{6}$$

where N is the number of pixels, M is the number of categories, $p = \text{SAM}\left(S\right)$ and $\hat{p} = \text{SAM}\left(S'\right)$ denote the regional semantic maps of the visual stimulus image S and its corresponding reconstructed image S', respectively. Combining (5) and (6), we finally obtain:

$$\mathcal{L}_{ldm-region} = \lambda_{ldm}\mathcal{L}_{ldm} + \lambda_{region}\mathcal{L}_{region}, \tag{7}$$

where λ_{ldm} and λ_{region} are balancing factors used to balance the influence between $\mathcal{L}_{ldm}$ and $\mathcal{L}_{region}$, with default values of 1. Subsequently, through backpropagation, we continuously fine-tune and update the weights of the LDM to endow it with the capability to capture semantic regions.

5 Experiments

5.1 Dataset and Implementation Details

EEG-ImageNet Dataset   The EEG-ImageNet dataset comes from the PeRCeiVe Lab [1], which contains visual stimulus EEG signals recorded from 6 subjects. The visual stimulus images encompass a total of 40 categories, with each category comprising 50 images sourced from ImageNet. Each image within the same category is displayed consecutively for 0.5 seconds, and there is a 10-second interval between each category. The collected dataset comprises a total of 11,964 segments (corresponding to visual stimuli for EEG). Each EEG segment includes 128 channels, and the EEG signals are in three frequency ranges: 14-70Hz, 5-95Hz, and 55-95Hz.

Things-EEG2 Dataset   To validate the effectiveness of the proposed method, we introduce an additional Things-EEG2 dataset [21]. This dataset includes ten subjects, with a total of 1,654 training sample categories. Each image is displayed for 100 ms with a 750 ms blank interval. All participants completed four equivalent experiments, resulting in 16,540 training images, each condition repeated four times, and 200 test images, each condition repeated 80 times.

fMRI-Image Dataset   The fMRI dataset comes from Mind-3D [?], and in the supplementary materials, the methods proposed in this paper are compared with Mind-3D. The methods in this paper reconstruct 3D objects with consistent style.

Implementation Details of EEG-to-3D   In this work, the latent EEG codes obtained in Stage A are used to guide LDM in generating diffusion priors, with Stable Diffusion version 1.5 employed. To accelerate NeRF training and rendering in Stage B, the proposed method adopts Instant-NGP [34]. We use $\mathcal{L}_{SDS}$, $\mathcal{L}_{EEG-Image}$, and $\mathcal{L}_{neuron-style}$ to compute the loss for generating 3D objects, and the Adam optimizer is used to update the model parameters, with the learning rate set to 0.001. Our hardware configuration utilizes an NVIDIA A100 GPU

Models	Acc.	FID [35]	IS [36]	Size (GB)	Inference time
EEGStyleGAN-ADA [37]	0.38	10.82	⌣	0.458	⌣
DreamDiffusion [13]	0.46	⌣	⌣	5.4	50 s/pic
EEG-Decoding [38]	0.47	⌣	⌣	⌣	⌣
EEGVis-CMR [14]	0.51	⌣	⌣	⌣	⌣
Brain2Image [39]	⌣	⌣	5.01	⌣	⌣
NeuroVision [40]	⌣	⌣	5.23	⌣	⌣
Ours	**0.67**	**2.17**	**25.41**	**7.8**	**50 s/pic**

Table 3: Quantitative metrics for 2D image reconstruction.

	Average Value	0°	20°	45°	70°
LPIPS [41]	0.3384	0.2403	0.3617	0.4539	0.4952
Contextual [42]	2.7943	2.3406	2.9274	3.1158	3.2932
Acc [43]	0.9468	0.8259	0.8664	0.8093	0.8917

Table 4: Quantitative LPIPS and Contextual metrics for different viewpoints of 3D objects. Acc refers to the scores obtained using CLIP. This study measures the results by evaluating the semantic similarity between the generated views and the reference views.

	Full	w/o $\mathcal{L}_{neuron-style}$	w/o $\mathcal{L}_{EEG-CLIP}$	w/o L_{region}
LPIPS	0.3384	0.5981	0.3061	0.5827
Contextual	2.7943	3.8915	3.2685	3.6412
Acc	0.9468	0.8134	0.5837	0.7258

Table 5: A quantitative comparison of different design choices from the perspectives of LPIPS and Contextual metrics.

5.2 Results

5.2.1 Quality with 3D and Style Consistency

As shown in Table 1 and 2, this paper reconstructs 3D objects from EEG signals under two conditions: with and without style transfer loss. For the reconstruction of 3D objects with style consistency, the proposed method integrates latent EEG codes, semantically accurate reconstructed images, and style transfer to generate 3D objects with consistent style. These reconstructed 3D objects not only exhibit high geometric similarity to GT objects but also maintain consistency in style. This demonstrates that our model effectively captures the core features of 3D objects in the process of translating EEG signals into complex visual information, providing evidence that EEG signals can perceive and process various visual information such as color, shape, and texture.

5.2.2 Quantity

As shown in Table 3 and Table 4, our quantitative analysis results are presented. We primarily used the following metrics to evaluate the performance of the reconstructed 2D and 3D models: 1) Fréchet Inception Distance (FID) [35], which assesses the difference between the reconstructed 2D images and real images; 2) Structural Similarity Index Measure (SSIM) [44], which measures the authenticity of the 2D images; 3) Inception Score (IS) [36], which evaluates the quality and diversity of the reconstructed images; 4) LPIPS [41], which assesses the 3D reconstruction quality of reference views; and 5) Contextual Distance [42], which measures the pixel-level similarity between novel view renderings and reference images.

Quantitative Analysis of 2D Among these metrics, 1), 2), and 3) are quantitative indicators for 2D images. As shown in Table 3, some values are missing due to the fact that certain metrics were not reported in the literature for the comparison methods, and the source code was not publicly available. For the Acc metric in the 50-way top-1 task, the accuracy of our proposed method is 21.4% higher than that of DreamDiffusion [13]. In terms of the IS metric, our model also significantly outperforms Brain2Image [39] and NeuroVision [40], indicating that the images generated by our model are more realistic and clearer. Compared to these models, our method demonstrates superior performance in both diversity and authenticity of the reconstructed images from EEG signals.

Quantitative Analysis of 3D Metrics 4) and 5) are qualitative indicators for 3D objects. As shown in Table 4, we evaluated the LPIPS and Contextual Distance metrics from different angles during the process of generating style-consistent 3D objects using EEG signals. The Acc metric here refers to the scores obtained using CLIP, where we assess the semantic similarity between the newly generated views and the reference views to measure the results.

Ablation Study of 3D We conducted three ablation experiments to quantitatively analyze the impact of each component on the visual quality of style-consistent 3D objects. Table 5 presents the results of the ablation studies. When all components are utilized, our network achieves the best performance.

1) $\mathcal{L}_{neuron-style}$. To enable the EEG reconstruction of style-consistent 3D objects, we introduced neural style loss, which ensures that the 2D images reconstructed from different viewpoints using EEG maintain style consistency with GT. This approach facilitates the generation of style-consistent 3D objects.

2) $\mathcal{L}_{EEG-CLIP}$. Since the method proposed in this paper aligns EEG signals with novel views from different perspectives, we introduce the alignment between CLIP's image encoder and EEG signals. $\mathcal{L}_{EEG-CLIP}$ aligns the novel view image $g(\theta)$ obtained under the diffusion prior with the EEG signals.

3) $\mathcal{L}_{region}$. To enable the diffusion model with semantic region-aware capabilities, we keep the inherent $\mathcal{L}_{ldm}$ property of the LDM unchanged. By integrating latent EEG codes with regional images, we fine-tune LDM, allowing it to reconstruct images with precise positional and spatial representation.

6 Conclusions

We propose an EEG-to-3D with style consistency approach, which, for the first time, utilizes EEG signals to generate high-fidelity 3D objects with consistent style. This method employs a two-stage strategy, integrating EEG signals, visual stimulus images, and style loss, to jointly fine-tune LDM and NeRF through collaborative training. The reconstructed 3D objects exhibit a high degree of realism and consistency in both geometry and color.

Limitations At stage A phase, the semantic accuracy of 2D images reconstructed from EEG signals still has room for improvement, which is expected to further enhance the clarity of the generated 3D objects. Nevertheless, it is noteworthy that the findings of this study have demonstrated that brain signals can perceive and encode various visual features, such as the color, shape, and texture of objects.

As shown in Tables 6 and 7, we present several key reconstructed images featuring style-consistent 3D objects. Our model utilizes EEG to directly reconstruct these style-consistent 3D objects, accurately restoring the spatial positions of the objects and the spatial relationships between multiple objects within the scene. As illustrated in Table 8, the method proposed in this paper has been validated on the fMRI dataset, demonstrating the ability to reconstruct style-consistent 3D objects using fMRI data.
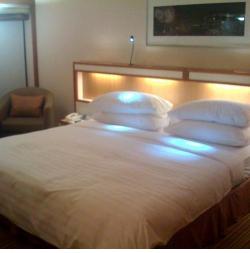
Visual Stimulus Images	with $\mathcal{L}_{neuron-style}$	3D Views (20°)	45°	70°

Table 6: Additional typical 3D objects generated by EEG-driven models based on the Things-EEG2 dataset [21].

References

[1] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6809–6817, 2017.

[2] Jay Hegdé. Time course of visual perception: coarse-to-fine processing and beyond. *Progress in neurobiology*, 84(4):405–439, 2008.

[3] Michele Fabre-Thorpe. Visual categorization: accessing abstraction in non–human primates. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1435):1215–1223, 2003.

[4] Rufin VanRullen and Christof Koch. Competition and selection during visual processing of natural scenes and objects. *Journal of vision*, 3(1):8–8, 2003.

[5] Richard A Abrams, David E Meyer, and Sylvan Kornblum. Speed and accuracy of saccadic eye movements: characteristics of impulse variability in the oculomotor system. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):529, 1989.

[6] Keith Rayner. Eye movement latencies for parafoveally presented words. *Bulletin of the Psychonomic Society*, 11(1):13–16, 1978.

[7] Lin Chen. Topological structure in visual perception. *Science*, 218(4573):699–700, 1982.

[8] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22710–22720, June 2023.

Visual Stimulus Images	with $\mathcal{L}_{neuron-style}$	3D Views (20°)	45°	70°

Table 7: Additional typical 3D objects generated by EEG-driven models based on the EEG-ImageNet dataset [1].

[9] Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777, 2023.

[10] Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. Minddiffuser: Controlled image reconstruction from human brain activity with semantic and structural diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5899–5908, 2023.

[11] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.

[12] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1809–1817, 2017.

[13] Yunpeng Bai, Xintao Wang, Yanpei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. DreamDiffusion: Generating high-quality images from brain EEG signals. In *European Conference on Computer Vision (ECCV)*. Springer, 2024.

[14] Zesheng Ye, Lina Yao, Yu Zhang, and Sylvia Gustin. Self-supervised cross-modal visual retrieval from brain activities. *Pattern Recognition*, 145:109915, 2024.

[15] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022.

[16] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22819–22829, 2023.

[17] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

[18] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Visual Stimulus Images	with $\mathcal{L}_{neuron-style}$	3D Views (20°)	45°	70°

Table 8: Additional typical 3D objects generated by ours models based on the fmri dataset [?].

[19] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022.

[20] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *Advances in Neural Information Processing Systems*, 35:33999–34011, 2022.

[21] Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022.

[22] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023.

[23] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 453–468, 2018.

[24] Tai-Yin Chiu and Danna Gurari. Pca-based knowledge distillation towards lightweight and content-style balanced photorealistic style transfer models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7844–7853, 2022.

[25] Michèle Fabre-Thorpe, Ghislaine Richard, and Simon J Thorpe. Rapid categorization of natural images by rhesus monkeys. *Neuroreport*, 9(2):303–308, 1998.

[26] Attila Korik et al. Real-time feedback improves imagined 3d primitive object classification from eeg. *Brain-Computer Interfaces*, pages 1–25, 2024.

[27] Yukako Yamane et al. A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature neuroscience*, 11(11):1352–1360, 2008.

[28] Ye Liu, Ming Li, Xian Zhang, Yiliang Lu, Hongliang Gong, Jiapeng Yin, Zheyuan Chen, Liling Qian, Yupeng Yang, Ian Max Andolina, et al. Hierarchical representation for chromatic processing across macaque v1, v2, and v4. *Neuron*, 108(3):538–550, 2020.
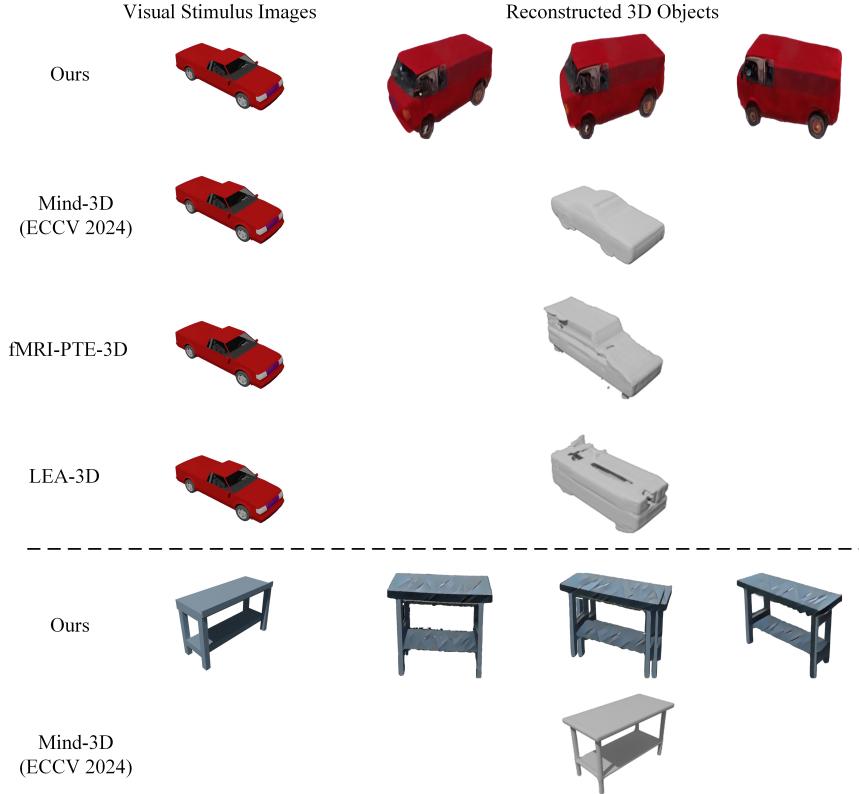
Figure 4: Our model reconstructs 3D objects using fMRI.

[29] Anupam K Garg, Peichao Li, Mohammad S Rashid, and Edward M Callaway. Color and orientation are jointly coded and spatially organized in primate primary visual cortex. *Science*, 364(6447):1275–1279, 2019.

[30] Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. Seeing beyond the brain: Conditional diffusion model with sparse masked modeling for vision decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22710–22720, 2023.

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[32] Tai-Yin Chiu and Danna Gurari. Pca-based knowledge distillation towards lightweight and content-style balanced photorealistic style transfer models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7844–7853, June 2022.

[33] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[34] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022.

[35] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[37] Prajwal Singh, Dwip Dalal, Gautam Vashishtha, Krishna Miyapuram, and Shanmuganathan Raman. Learning robust deep visual representations from EEG brain recordings. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7553–7562, 2024.

[38] Matteo Ferrante, Tommaso Boccato, Stefano Bargione, and Nicola Toschi. Decoding visual brain representations from electroencephalography through knowledge distillation and latent diffusion models. *Computers in Biology and Medicine*, page 108701, 2024.

[39] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images. In *25th ACM International Conference on Multimedia*, page 1809–1817, 2017.

[40] Sanchita Khare, Rajiv Nayan Choubey, Loveleen Amar, and Venkanna Udutalapalli. NeuroVision: perceived image regeneration using cProGAN. *Neural Computing and Applications*, page 5979–5991, 2022.

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[42] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 768–783, 2018.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

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