Joint Learning for Visual Reconstruction from the Brain Activity: Hierarchical Representation of Image Perception with EEG-Vision Transformer

Ali Akbari
Department of Electrical Engineering
Sharif University of Technology
Tehran, Iran
ali.akbari@ee.sharif.edu

Kosar Sanjar
Department of Electrical Engineering
University of Tehran
Tehran, Iran
kosar.sanjar@gmail.com

Muhammad Yousefnezhad
Departments of Computing Science and Psychiatry
University of Alberta
Alberta, Canada
myousefnezhad@ualberta.ca

Maryam S. Mirian
Department of Medicine
University of British Columbia
British Columbia, Vancouver
maryam.mirian@ubc.ca

Emad Arasteh
Department of Neonatology
University Medical Center Utrecht
Utrecht, the Netherlands
earasteh@umcutrecht.nl

Editors: Marco Fumero, Clementine Domine, Zorah Lähner, Donato Crisostomi, Luca Moschella, Kimberly Stachenfeld

Abstract

Reconstructing visual stimuli from brain activity is a challenging problem, particularly when using EEG data, which is more affordable and accessible than fMRI but noisier and lower in spatial resolution. In this paper, we present Hierarchical-ViT, a novel framework designed to improve the quality and precision of EEG-based image reconstruction by integrating hierarchical visual feature extraction, vision transformer-based EEG (EEG-ViT) processing, and CLIP-based joint learning. Inspired by the hierarchical nature of the human visual system, our model progressively captures complex visual features—such as edges, textures, and shapes—through a multi-stage processing approach. These features are aligned with EEG signals processed by the EEG-ViT model, allowing for the creation of a shared latent space that enhances contrastive learning. A StyleGAN is then employed to generate high-resolution images from these aligned representations. We evaluated our method on two benchmark datasets, EEGCVPR40 and ThoughtViz,

Proceedings of the II edition of the Workshop on Unifying Representations in Neural Models (UniReps 2024).

achieving superior results compared to existing approaches in terms of Inception Score (IS), Kernel Inception Distance (KID), and Fréchet Inception Distance (FID) for EEGCVPR, and IS and KID for the ThoughtViz dataset. Through an ablation study, we underscored the feasibility of hierarchical feature extraction, while the multivariate analysis of variance (MANOVA) test confirmed the distinctiveness of the learned feature spaces. In conclusion, our results show the feasibility and uniqueness of using hierarchical filtering of perceived images combined with EEG-ViT-based features to improve brain decoding from EEG data.

1 Introduction

Visual reconstruction is a critical area of research in both neuroscience and machine learning (ML), as it provides insights into how the brain processes and represents perceptible information [Takagi and Nishimoto, a]. In this regard, understanding the neural correlates of visual perception is vital for decoding brain activity, which has implications for both cognitive science and clinical applications [Pollen]. Traditionally, visual reconstruction methods have heavily relied on functional magnetic resonance imaging (fMRI) data, which allows for high spatial resolution images of brain activity [Rakhimberdina et al.]. However, fMRI-based approaches come with limitations, including high costs, limited accessibility, and (often) low temporal resolution, making them less practical for continuous monitoring or real-time applications [Glover, Wilson et al.]. To tackle the challenges mentioned above, there is a growing interest in EEG-based methods of visual reconstruction, as EEG offers a balance between temporal resolution and cost-effectiveness [Wilson et al.].

The field of image reconstruction from brain activity has recently advanced noticeably with the adoption of sophisticated generative models, including generative adversarial networks (GANs) and diffusion models with the Stable Diffusion [Ozcelik and VanRullen] ability to generate high-resolution images, showcasing the power of large pre-trained models [Takagi and Nishimoto, b] and offering new insights into visual processing. Moreover, recent advancements in AI and machine learning have opened possibilities for effective image reconstruction from EEG signals, addressing the need for cost-effective and efficient approaches in neuroimaging [Li et al., Guenther et al.]. Joint learning is one of these recent techniques that enables continuous data of EEG to be paired with other modalities like perceived images. Such pairing can help to synthesize new aspects of EEG-Image dynamics for more effective brain decoding [Song et al., Xu et al.]. The CLIP (Contrastive Language-Image Pretraining) model [Radford et al.] is a joint learning framework where a single model learns visual and language representations by predicting which caption matches which image. By extending this capability to EEG data, the model learns shared representations across different modalities, enhancing the potential for accurate visual reconstructions from EEG signals [Singh et al., a].

Regardless of their feasibility in aligning EEG data with perceived images, multimodal learning models like CLIP cannot provide a measure of a biologically plausible representation of decoded brain activity. This may be one important reason that most of the existing models for generating images from EEG data have primarily excelled in distinguishing between different classes of perceived images, rather than reconstructing the actual visualized image with high accuracy [Kavasidis et al., Jiang et al., a, Khare et al., Spampinato et al.]. On the other hand, several researchers have demonstrated the advantages of biologically inspired computational modeling for both advanced deep learning models' performance as well as more feasible biologically-inspired tools [Kim et al., Luppi et al., Collins and Shenhav]. Over the last two decades, several models for decoding visual perception mechanisms have been proposed that support the concept of hierarchical image processing in the brain, where different layers (e.g., V1, V2, V3, V4 in the visual cortex) process different aspects of visual stimuli. [Pohl et al., Bracci et al., D'Souza et al.].

In this paper, we propose a model (called here "Hierarchical-ViT") combining the vision transformer (ViT), hierarchical visual feature extraction, and contrastive learning to improve visual reconstruction from EEG signals. We utilized the EEG-ViT [Yang] model for EEG feature extraction, leveraging their self-attention mechanisms to capture complex temporal and spatial patterns in brain activity. These EEG features are integrated with hierarchical visual features, inspired by the human visual system's layered processing of visual stimuli. The combined EEG and visual features are aligned in a shared latent space using the CLIP framework, enhancing the model's ability to accurately reconstruct images from EEG data. Finally, a StyleGAN [Karras et al.] model is employed for high-resolution image generation, allowing for greater control and realism in the reconstructed visuals.

The paper is structured as follows: The Related works section reviews existing approaches and models for image reconstruction from brain activity. Then, the method section details our proposed approach. The experiment and results section presents the evaluation of the proposed method's performance. The discussion section interprets the results and explores their implications for future research. Finally, the conclusion summarizes the key contributions and potential directions for further development in EEG-based visual reconstruction.

2 Related works

Recent improvements in generative AI have encouraged researchers to develop new encode-decoder frameworks for image reconstruction by VAE, GAN, or latent diffusion models [Huang et al., a, Gong et al., Ozcelik and VanRullen]. Although the diffusion and VAE models have great advantages in terms of stability and versatility, GAN models are well-known for their ability to generate realistic images [Peng]. This can be helpful, especially in the endeavor of generating natural image generation by brain decoding where metrics like IS, KID, and FID are the major criteria of models' performance.

Researchers have developed attention-based GAN architectures that can reconstruct complex natural object images from EEG data, outperforming traditional cross-modality encoder-decoder networks [Habashi et al.]. These models often incorporate additional components such as perceptual loss and auxiliary classifiers to improve the quality and relevance of the generated images [Mishra and Bhavsar]. Other approaches have utilized contrastive learning methods to extract features from EEG signals, which are then used to condition GANs for image synthesis. These techniques have shown promise in reconstructing various types of visual stimuli, including objects, digits, and characters, even when working with small-scale EEG datasets [Hartmann et al.].

While GAN-based image reconstruction methods applied to EEG data have shown promising advancements, they still face challenges in achieving the same level of quality as similar techniques applied to fMRI data, particularly in terms of IS and FID [Yang and Modesitt]. Therefore, there is still room for improvement to elevate the performance of EEG data decoding for natural image reconstruction.

3 Method

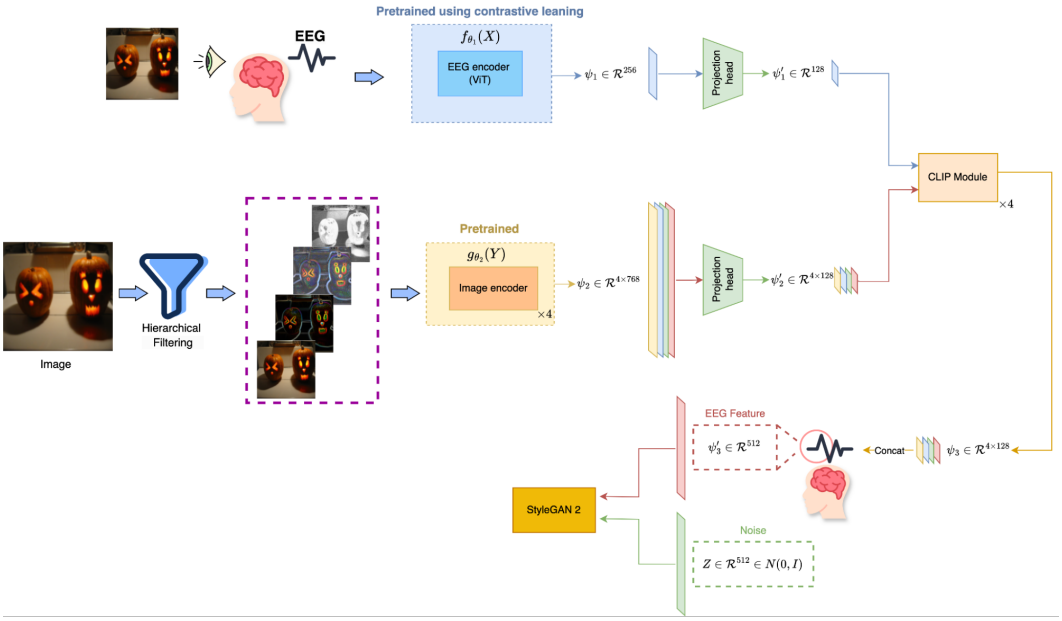
We considered two already-established AI and cognitive science-known facts to enhance image reconstruction from EEG signals:

1. The self-attention mechanism of transformer models generally excels LSTM methods in terms of long-term dependencies and flexible context modelings. Moreover, in the case of EEG data analysis, Vision transformers (ViT) can extract more feasible spatial-temporal features compared to regular attention-based models [Yang and Modesitt].
2. The human visual system processes information through a hierarchical structure, where different specialized brain regions manage progressively more complex aspects of visual stimuli [Lerner et al.].

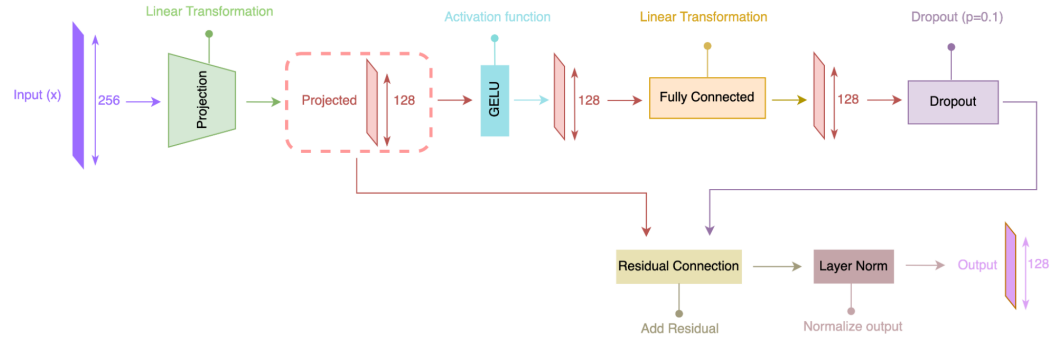
Based on these two facts, we propose two modifications to existing GAN-based image reconstruction models:

1. An EEG-ViT [Dosovitskiy et al.] for feature extraction from EEG instead of LSTM and CNN models to uncover longer-term spatial-temporal dynamics of the brain.
2. Hierarchical feature extraction from images with joint space learning to improve the biological plausibility of signal processing compared to earlier methods.

These modifications form the foundation of our proposed image reconstruction framework, as illustrated in Fig. 1. While classification accuracy remains important, metrics such as IS, KID, and FID are more reflective of the quality of the reconstructed images. In this paper, we hypothesize that leveraging the EEG-ViT model alongside biologically inspired image feature extraction will enhance these three metrics. Accordingly, our goal is to achieve a joint learned representation of EEG and image features to improve the quality of reconstructed images. All experiments and models were



(a)



(b)

Figure 1: Hierarchical-ViT scheme. (a) Overview of the proposed framework for (training phase of) the EEG-based visual reconstruction using hierarchical feature extraction and contrastive learning. (b) The detailed structures of the "Projection head" are drawn in the general scheme. Detailed structures of the other components are drawn in the subsequent subsections through Fig. 2 and the appendix section.

trained on a server equipped with one V100 GPU card and 200 GB of RAM. The entire analysis took approximately 100 hours.

3.1 Notations

We refer to the feature vectors extracted from EEG and image as psi in Fig. 1. The EEG signal is referred to as X and the image as Y. The EEG signal is in space $X \in \mathbb{R}^{N \times C \times T}$, where N is the number of EEG trials, C is the number of channels, and T is the number of time points. The corresponding labels are marked as L. The EEG feature extractor is shown by $f_{\theta_1}(X)$ and the image feature extractor as $g_{\theta_2(X)}$, where θ_i is the learned weight during the training. The projection head for g_θ and f_θ are g_γ and f_γ. We define the problem as follows: Given a data set of samples of $\{X, Y, L\}$, we want to train a deep neural network pipeline to reconstruct Y, given X and L.

3.2 Transformer-based feature extraction from EEG data

Transformer models offer significant advantages over LSTM and CNN-based approaches for EEG feature extraction [Hu et al.]. By leveraging self-attention, they can capture long-range dependencies more effectively, improving the understanding of complex temporal patterns in brain activity [Siddhad et al.]. In this paper, we selected EEG-ViT models for their unique capabilities to effectively and simultaneously capture both the spatial relationships between EEG channels and the temporal dynamics of brain activity [Yang and Modesitt, Patel et al.].

During pre-training, the triplet margin loss is used, while Cross Entropy Loss is applied in the CLIP setting to maximize the similarity between image and EEG pairs. We use the triplet loss for feature learning with semi-hard triplets. We utilized a Vision transformer (ViT)-based model, EEG-ViT, for EEG feature extraction. The Transformer treats the EEG data as image-like inputs, enabling it to model spatial-temporal relationships more effectively.

The EEG encoder $f_{\theta_1}(X)$ maps the EEG signals into a feature space $\psi_1 \in \mathbb{R}^{N \times 256}$. This encoded feature representation captures both local and global patterns from the input EEG signals. We apply a projection head to transform these features into a lower-dimensional latent space $\psi_1' \in \mathbb{R}^{N \times 128}$. To optimize this process, we pre-train the EEG encoder using contrastive loss. The triplet margin loss function is employed to align the EEG features with their corresponding visual representations. The loss function is given by:

$$\theta = \arg\min_{\theta} \mathbb{E}\left[||f_\theta(X^a) - f_\theta(X^p)||_2^2 - ||f_\theta(X^a) - f_\theta(X^n)||_2^2 + \delta\right]$$

where X^a, X^p, and X^n represent anchor, positive, and negative samples, respectively. This helps ensure that the EEG features are closely aligned with the correct visual stimuli in the latent space.

3.3 Hierarchical visual features from the perceived natural images

Hierarchical models of the visual system are considered to correspond to four major feedforward v1-v4 layers by some neuroscientists [Riesenhuber and Poggio]. In the early processing stage of model area V1, boundaries and their orientations are detected, followed by a grouping process in model area V2. Contextual boundary patterns are also processed at a broader spatial level in model areas V2 and V4, allowing for sensitivity to contour curvature [Angelucci and Bressloff].

In this paper, we took advantage of simple feedforward hierarchical filtering of perceived images to use for joint learning with EEG features. The four hierarchical filters are as follows:

1. **V1 (Edge detection)**: The Sobel filter computes the gradient magnitude of an image to detect edges. The gradient magnitude is given by:

   $$G = \sqrt{(S_x^2 + S_y^2)}$$

   Where:

   - S_x and S_y are the gradient in the x and y direction, respectively.

2. **V2 (Texture and contour detection)**: Local Binary Patterns (LPBs) and contour detection simulate V2's role in recognizing textures, contours, and boundary details, processing the information from the prior layers.

3. **V3 (Motion and color processing)**: The HSV [Smith, 1978] (Hue, Saturation, Value) color model separates color information into three channels. The saturation channel is extracted from the converted RGB color space to HSV color space and used as the main component extracted in V3.

   Detailed explanations are discussed in the Supplementary section. The outcomes of V4 were shown heuristically to be redundant in terms of the jointly learned feature space of the original image. Therefore, we used V1-V3 added to the original image for feature extraction from the image.

   We utilize a pre-trained ViT [Wu et al., Deng et al.], $g_{\theta_2}(Y)$ and fine-tune it on our image data set. Four models are trained on each one of the resulting image data sets after filtering in V1, V2, and V3 filters and the original images. The objective function is the contrastive loss. This gives us an embedding space $\psi \in \mathbb{R}^{256}$.

3.4 CLIP-based joint learning of image and EEG

Recent studies have explored innovative approaches to bridge the gap between EEG signals and visual representations. Palazzo et al. employed a contrastive learning strategy with triplet loss to train an EEG encoder, aligning it with image features generated by a pre-trained image encoder [48]. Zesheng et al. [49] employed CLIP for joint representation learning by generating image representations via a GAN before training the EEG encoder using contrastive methods for image retrieval. On the other hand, Singh et al. [Singh et al., a] directly applied a pre-trained image encoder for EEG-based image retrieval tasks, streamlining the process and potentially improving efficiency.

We used the fine-tuned $g_{\theta_2}(Y)$ and pre-trained $f_{\theta 1}(X)$ in the CLIP module to align their embedding spaces, $\psi_1' \in \mathbb{R}^{128}$ and $\psi_2' \in \mathbb{R}^{128}$. We freeze g_{θ_2} and only allow f_θ and g_γ and f_γ to be updated in this process, due to the higher accuracy rate that g_θ possesses in our framework. This setup allows the EEG encoder to learn to better align its representation, resulting in better accuracy rates. Fig. 2 depicts the general graphical scheme of joint learning of EEG-Image feature spaces in our work.

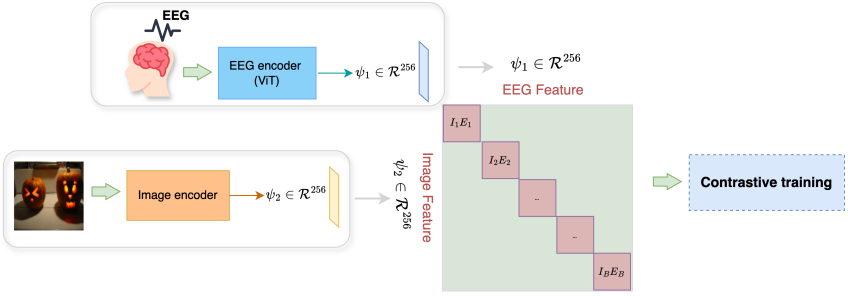


Figure 2: Joint learning of image and EEG features. The general scheme of CLIP-based joint learning in our work. This architecture allows us to align the representations of the EGG signals and the corresponding images. We trained four different CLIP modules, corresponding to V1, V2, V3 generated images and the original images. Each one of the CLIP modules is equipped with its corresponding pre-trained image encoder. Therefore, each EEG encoder learns a different embedding space.

We trained four different CLIP modules, corresponding to V1, V2, V3 generated images and the original images. Each one of the CLIP modules is equipped with its corresponding pre-trained image encoder. Therefore, each EEG encoder learns a different embedding space.

3.5 Image generation by Style-GAN model

StyleGAN includes several key features that improve image generation. It uses style blending with hybrid regularization and two random latent codes, giving users control over the style of the images for greater customization [Huang et al., b]. The model is capable of producing high-resolution images and managing complex tasks like face and landscape generation, reducing common issues like blurring and distortion seen in traditional GANs. Additionally, StyleGAN introduces stochastic variation by adding uncorrelated Gaussian noise to each layer of the network, which helps generate diverse and varied details. This allows the generated image to have some random variations in detail while maintaining overall structural consistency, increasing the diversity and realism of the image [Karras et al.].

We concatenated the resulting EEG features from the CLIP model with a normal vector $z \in \mathbb{R}^{512}$, resulting in a vector $\mathbb{R}^{1024}$. The model is trained to reconstruct the images, corresponding to the EEG feature vector.

3.6 Unique feature space and ablation study of hierarchical image features

While previous research has shown the biological plausibility of hierarchical feature extraction from perceived images [Horikawa and Kamitani, DiCarlo et al., Serre et al., Riesenhuber and Poggio], to the best of our knowledge, the importance of these features for improving brain decoding models has not been thoroughly explored. For this aim, we first evaluated the uniqueness of the feature

space out of joint learning for V1-V3 extracted features compared to the feature space learned from the original image. We performed a MANOVA test on the generated feature spaces to evaluate the uniqueness of each feature space. In the next step, we applied an ablation study to evaluate the effect of these hierarchical features on the quality and precision of the generated images. An ablation study is a methodical technique in machine learning research used to analyze the influence of individual components or features on a model's performance. The process entails selectively removing or modifying certain elements of the model, retraining it, and then evaluating the impact of these alterations on its overall performance [Meyes et al.]. We compared the results of Hierarchical-ViT with the CLIP-ViT model in which only features of the original image are utilized for joint learning with EEG features.

4 Experiment and results

The first part of this section discusses the two datasets. The second part discusses the outcomes of feature extraction. After that, hierarchical feature extraction is discussed. Joint learning, image synthesis, generated images, and ablation study results are the next subsequent presented parts of this section.

4.1 Datasets

EEGCVPR This dataset [Spampinato et al.] is a subset of ImageNet [Deng et al.], including data of 40 object classes, with 50 images per class, for a total of 2,000 images. We utilized the version of 5-95 Hz of the EEGCVPR dataset to include the biggest possible bandwidth of frequency components in the data. The recording protocol presented visual stimuli to users in a block-based setting, showing images of each class consecutively in a single sequence. Each EEG segment contains data from 128 channels, recorded for 0.5 seconds at a 1 kHz sampling rate. The resultant EEG signal consists of 440 time samples, after discarding the first and the last time samples.

ThoughtViz This dataset [Tirupattur et al.] includes 10 different categories of objects. The images were shown to the participants and were asked to imagine the image that was shown to them. The EEG signal consists of 14 channels. The sampling frequency of the device is 128 Hz. After pre-processing [Tirupattur et al.], each epoch has 32 time steps.

4.2 Jointly learned EEG-image features by CLIP

Feature extraction from images is performed using a Sobel filter for edge detection (V1), LBP for texture features (V2), and color intensity (V3). These images are fed into the 3 different CLIP models, allowing them to learn different weights and feature spaces. The original image is also fed into another clip model. In Fig. 3, one example of the effect of each V1-V3 filter on the original image is depicted.

The EEG-ViT architecture consists of 3 layers. The MLP dimension is set to 64 and the number of attention heads, and the attention dimension is set to 16. A dropout rate of 0.5 is applied consistently across models. To further reduce the risk of overfitting, FTsurrogate and smooth time masking are employed for EEG data augmentation, techniques that have been shown to enhance performance in BCI tasks involving EEG signals. We used Adam optimizer alongside cosine learning rate scheduler, ϕ of FTsurrogate equal to 1, and 0.5 probability of data augmentation for both of the data sets.

Fig. 4 depicts the nonlinearly mapped two-dimensional feature space from EEG-ViT by t-SNE [Maaten and Hinton] for the EEGCVPR dataset. The feature spaces of the ThoughtViz dataset are also depicted in the Supplementary section.

4.3 Generated images by StyleGAN

A group of images generated for each of the datasets by our model is shown in Fig. 5.

Table 1 summarizes the classification accuracy and quality of the generated images among the two data sets for our method and already established methods in the field of image generation from EEG. Our Hierarchical-ViT model, combined with StyleGAN, generates high-resolution images from EEG data for both the EEGCVPR40 and ThoughtViz datasets. The hierarchical extraction of visual

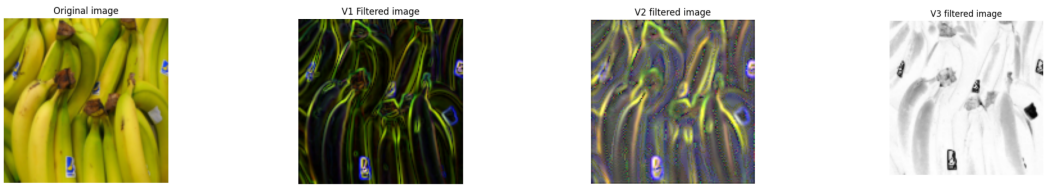


Figure 3: The filtering effect on a sample image

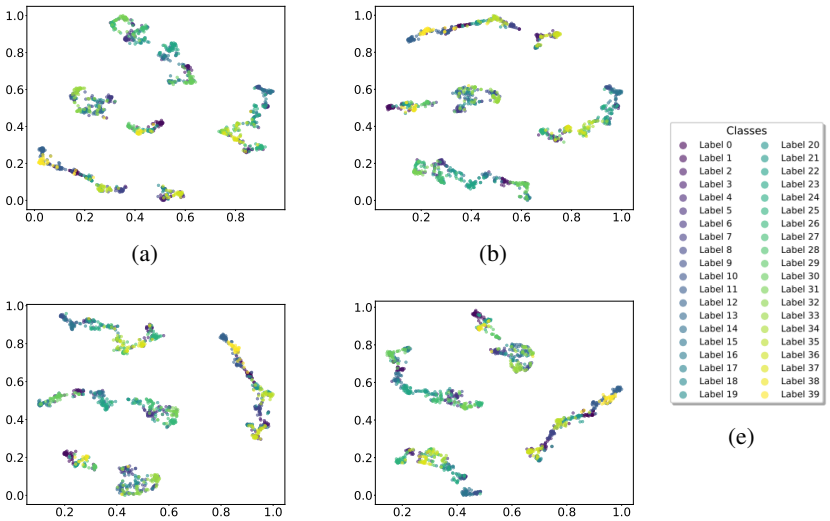


Figure 4: The result related to effects V1-V3 filtering on the feature space. The figures illustrate the first two dimensions of the t-SNE map [Maaten and Hinton] (horizontal as the first dimension). CLIP-based Jointly learned feature space of (a) original image, (b) V1-filtered image, (c) V2-filtered, and image (d) V3-filtered image.

(a) Real Images (b) Generated Images

Figure 5: Comparison of real perceived images and generated images by Hierarchichal-ViT model for EEGCVPR [Spampinato et al.] and ThoughtViz [Tirupattur et al.] datasets. The first two rows are related to EEGCVPR and the final row is from ThoughtViz.

features, aligned with EEG representations, improves the quality of generated images. Results show superior outcomes in IS, FID, and KID for the EEGCVPR dataset and IS and KID for ThoughtViz.

4.4 Comparison of feature spaces and ablation study

We compared the four feature spaces depicted in Fig. 4 by multivariate analysis of variance (MANOVA) [Tabachnick and Fidell]. The MANOVA test showed that the V1, V2, and V3 feature spaces differ from the feature space related to the original image significantly (Pillai's Trace

= 2.9960, F(384, 13437) = 26053.44, p < 1e-5]). Therefore, each of the V1-V3 feature spaces had unique feature representations and were not redundantly adding dependent variables to the Hierarchical-ViT model. The detailed outcomes of MANOVA statistics are mentioned in the Supplementary section. For the ablation study (as mentioned in section 3.6), we evaluated the outcomes of the IS, FID, and KID by using only a learned representation of the original image. As mentioned in Table 1, the IS, FID, and KID of our model is 12.17, 122.91, and 0.059, while for the same model, these results degraded to 11.17, 126.88, and 0.062. Therefore, this ablation study also shows the feasibility of adding hierarchical feature extraction for more accurate decoding of visual processing from brain activity.

Dataset	**Reference**	**Discriminative Model**	**Classification**		**Generative Model**	**Quality Metrics**		
			Accuracy	**K-Means**		**IS ↑**	**FID ↓**	**KID ↓**
EEG CVPR	[Kavasidis et al.]	-	-	-	Brain2Image-VAE	4.49	-	-
	[Palazzo et al.]	-	-	-	Brain2Image-GAN	5.07	-	-
	[Spampinato et al.]	LSTM Encoder	0.829	-	-	-	-	-
	Jiang et al. [a]	DML	0.977	-	-	-	-	-
	Zheng et al.	LSTM-CNN	0.944	-	Improved-SNGAN	5.53	-	-
	Jiang et al. [b]	BioLSTM	0.991	-	-	-	-	-
	Khare et al.	NeuroVision	0.988	-	-	5.15	-	-
	Singh et al. [a]	EEGLSTM	0.983	0.96	EEGStyleGAN-ADA	10.82	174.13	0.065
		EEG-ViT (ours)	**0.72**	**0.70**	Hierarchical-ViT (ours)	**12.17**	**122.91**	**0.059**
ThoughtViz	[Tirupattur et al.]	ThoughtViz	0.729	-	ThoughtViz	5.43	-	-
	Mishra and Bhavsar	SiameseCNN	0.741	-	NeuroGAN	6.02	-	-
	Singh et al. [b]	EEG2Image	0.55	-	EEG2Image	6.78	-	-
	Singh et al. [a]	EEGLSTM	0.741	0.72	EEGStyleGAN-ADA	9.23	109.49	0.039
		EEG-ViT (ours)	**0.85**	**0.84**	Hierarchical-ViT (ours)	**10.20**	**167.92**	**0.037**

Table 1: Between class discrimination by extracted EEG features and quality of the generated images. The discriminative metrics are calculated based on accuracy and k-means[Macqueen, 1967]. The k-means algorithm is applied to the features generated by our model. This table compares various approaches and loss functions for extracting generated images.

5 Discussion

This paper introduced a new approach for reconstructing images from EEG signals by combining transformer-based EEG feature extraction, hierarchical visual processing, and joint learning using the CLIP framework. The method improved the mapping of EEG signals to visual stimuli, leading to enhancements in the precision and quality of the generated images. On the EEGCVPR40 dataset, our model achieved an IS of 12.17 and a KID of 0.059, outperforming all the earlier established methods listed in Table 1. Similarly, with the ThoughtViz dataset, our model achieved an IS of 10.20 and a KID of 0.037, surpassing the other approaches. In the case of FID, our model outperformed all the other approaches by the value of 122.91 for the EEGCVPR dataset, while the EEGStyleGAN-ADA [Singh et al., a] exceeded our FID value (109.42 vs 167.92) for the ThoughtViz dataset. This shows the superiority of our model in 2 of 3 metrics to all other methods, with even better FID for a longer duration of the EEG dataset (EEGCVPR compared to ThoughtViz). These outcomes demonstrate the benefits of integrating hierarchical visual processing with transformer-based EEG feature extraction.

A key aspect of this approach is the use of hierarchical visual features, modeled after the layered structure of the human visual system. Results from the MANOVA test confirm that the features generated at each stage (V1, V2, V3) are unique and non-redundant, showing significant differences from the original image feature space (with p-value<1e-5). This highlights that feature spaces of hierarchical images are not redundant considering the original images' feature spaces. Moreover, the effectiveness of these hierarchical features was further confirmed by the mentioned ablation study showing lower quality images in terms of IS (11.17 vs. 12.17), FID (126.88 vs. 122.91), and KID (0.062 vs. 0.059).

One important limitation of our model is the lower classification accuracy for the EEGCVPR dataset compared to other methods (while our model reached the highest accuracy for ThoughtViz). It is worth mentioning that our model reached the highest image quality metrics on this dataset while attaining the lowest between-class discrimination. This can be a starting point for future works on the possible trade-off between class accuracy vs. precision of generated images in the brain-decoding

field of research. Regardless of the promising results of this study, some challenges remain for the future steps. The complexity of transformer architectures can increase the risk of overfitting, especially with smaller datasets. While the hierarchical visual processing approach has proven to be beneficial, further work is needed to test its generalizability across different EEG datasets and a broader range of stimuli. Future studies could also explore alternative architectures or more efficient data augmentation methods to address the balance between model complexity and dataset size. Moreover, new studies could also focus on better fine-tuning of Hierarchical-ViT model considering the length of EEG datasets in case of any relation between FID metric and the model's parameters.

6 Conclusion

This paper introduced a novel EEG-based image reconstruction method that integrates hierarchical feature extraction with transformer-based EEG processing and CLIP-inspired joint representation learning. By mimicking the human visual system's layered architecture, our approach effectively captured essential visual features like edges, textures, and contours, enhancing the alignment between EEG signals and visual representations. Our method outperformed existing techniques on datasets such as EEGCVPR40 and ThoughtViz, achieving superior Inception Score and Fréchet Inception Distance. Ablation studies confirmed the significance of hierarchical feature extraction in improving image quality and model robustness, while MANOVA tests validated the unique contributions of these features at various stages. Despite these promising results, challenges remain in generalizing to smaller, diverse datasets. Future research will focus on addressing model complexity and dataset scalability, alongside exploring efficient data augmentation strategies to enhance performance and mitigate overfitting.

In conclusion, the combination of hierarchical feature extraction with advanced EEG processing marks an advancement in EEG-to-image synthesis, while paving the way for more accurate brain decoding systems.

References

Yu Takagi and Shinji Nishimoto. Improving visual image reconstruction from human brain activity using latent diffusion models via multiple decoded inputs, a. URL https://arxiv.org/abs/2306.11536v1.

D. A. Pollen. On the neural correlates of visual perception. 9(1):4–19. ISSN 1047-3211. doi: 10.1093/cercor/9.1.4.

Zarina Rakhimberdina, Quentin Jodelet, Xin Liu, and Tsuyoshi Murata. Natural image reconstruction from fMRI using deep learning: A survey. 15:795488. ISSN 1662-4548. doi: 10.3389/fnins.2021.795488. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8722107/.

Gary H. Glover. Overview of functional magnetic resonance imaging. 22(2):133–139, vii. ISSN 1558-1349. doi: 10.1016/j.nec.2010.11.001.

Holly Wilson, Xi Chen, Mohammad Golbabaee, Michael J. Proulx, and Eamonn O'Neill. Feasibility of decoding visual information from EEG. 11(1):33–60. ISSN 2326-263X. doi: 10.1080/2326263X.2023.2287719. URL https://doi.org/10.1080/2326263X.2023.2287719. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/2326263X.2023.2287719.

Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. 13(1):15666. URL https://www.nature.com/articles/s41598-023-42891-8. Publisher: Nature Publishing Group UK London.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity, b. URL https://www.biorxiv.org/content/10.1101/2022.11.18.517004v3. Pages: 2022.11.18.517004 Section: New Results.

Dongyang Li, Chen Wei, Shiying Li, Jiachen Zou, and Quanying Liu. Visual decoding and reconstruction via EEG embeddings with guided diffusion. URL http://arxiv.org/abs/2403.07721.

Sven Guenther, Nataliya Kosmyna, and Pattie Maes. Image classification and reconstruction from low-density EEG. 14(1):16436. URL https://www.nature.com/articles/s41598-024-66228-1. Publisher: Nature Publishing Group UK London.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. URL http://arxiv.org/abs/2308.13234.

Jonathan Xu, Bruno Aristimunha, Max Emanuel Feucht, Emma Qian, Charles Liu, Tazik Shahjahan, Martyna Spyra, Steven Zifan Zhang, Nicholas Short, Jioh Kim, Paula Perdomo, Ricky Renfeng Mao, Yashvir Sabharwal, Michael Ahedor Moaz Shoura, and Adrian Nestor. Alljoined1 – a dataset for EEG-to-image decoding. URL http://arxiv.org/abs/2404.05553.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. URL https://arxiv.org/abs/2103.00020v1.

Prajwal Singh, Dwip Dalal, Gautam Vashishtha, Krishna Miyapuram, and Shanmuganathan Raman. Learning robust deep visual representations from EEG brain recordings. pages 7538–7547. IEEE Computer Society, a. ISBN 9798350318920. doi: 10.1109/WACV57701.2024.00738. URL https://www.computer.org/csdl/proceedings-article/wacv/2024/189200h538/1W0dmHguKU8.

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*, MM '17, pages 1809–1817. Association for Computing Machinery. ISBN 978-1-4503-4906-2. doi: 10.1145/3123266.3127907. URL https://doi.org/10.1145/3123266.3127907.

Jianmin Jiang, Ahmed Fares, and Sheng-Hua Zhong. A context-supported deep learning framework for multimodal brain imaging classification. 49(6):611–622, a. URL https://ieeexplore.ieee.org/abstract/document/8680664/?casa_token=9lyVb_10IHkAAAAA:lEWPivmzJ3FGoRF4LFVTw3EzJuTNIBEGcGiT17ApDUasQEokh7xGm7Ktew6A81fcm32IK1tZ. Publisher: IEEE.

Sanchita Khare, Rajiv Nayan Choubey, Loveleen Amar, and Venkanna Udutalapalli. NeuroVision: perceived image regeneration using cProGAN. 34(8):5979–5991. ISSN 0941-0643, 1433-3058. doi: 10.1007/s00521-021-06774-1. URL https://link.springer.com/10.1007/s00521-021-06774-1.

Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Mubarak Shah, and Nasim Souly. Deep learning human mind for automated visual classification. URL https://arxiv.org/abs/1609.00344v2.

Hong Ji Kim, Byeol Kim Lux, Eunjin Lee, Emily S. Finn, and Choong-Wan Woo. Brain decoding of spontaneous thought: Predictive modeling of self-relevance and valence using personal narratives. 121(14):e2401959121. ISSN 1091-6490. doi: 10.1073/pnas.2401959121.

Andrea I. Luppi, Joana Cabral, Rodrigo Cofre, Pedro A. M. Mediano, Fernando E. Rosas, Abid Y. Qureshi, Amy Kuceyeski, Enzo Tagliazucchi, Federico Raimondo, Gustavo Deco, James M. Shine, Morten L. Kringelbach, Patricio Orio, ShiNung Ching, Yonatan Sanz Perl, Michael N. Diringer, Robert D. Stevens, and Jacobo Diego Sitt. Computational modelling in disorders of consciousness: Closing the gap towards personalised models for restoring consciousness. 275:1–16. ISSN 1095-9572. doi: 10.1016/j.neuroimage.2023.120162. Place: Netherlands Publisher: Elsevier Science.

Anne G. E. Collins and Amitai Shenhav. Advances in modeling learning and decision-making in neuroscience. 47(1):104–118. ISSN 1740-634X. doi: 10.1038/s41386-021-01126-y.

Kilian M. Pohl, Sylvain Bouix, Motoaki Nakamura, Torsten Rohlfing, Robert W. McCarley, Ron Kikinis, W. Eric L. Grimson, Martha E. Shenton, and William M. Wells. A hierarchical algorithm for MR brain image parcellation. 26(9):1201–1212. ISSN 0278-0062. doi: 10.1109/TMI.2007.901433.

Stefania Bracci, Jakob Mraz, Astrid Zeman, Gaëlle Leys, and Hans Op de Beeck. The representational hierarchy in human and artificial visual systems in the presence of object-scene regularities. 19(4): e1011086. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1011086.

Rinaldo D. D'Souza, Quanxin Wang, Weiqing Ji, Andrew M. Meier, Henry Kennedy, Kenneth Knoblauch, and Andreas Burkhalter. Hierarchical and nonhierarchical features of the mouse visual cortical network. 13(1):503. ISSN 2041-1723. doi: 10.1038/s41467-022-28035-y.

Richard Yang. ruiqiRichard/EEGViT. URL https://github.com/ruiqiRichard/EEGViT. original-date: 2023-05-25T18:47:48Z.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. URL http://arxiv.org/abs/1812.04948.

Wei Huang, Hongmei Yan, Chong Wang, Xiaoqing Yang, Jiyi Li, Zhentao Zuo, Jiang Zhang, and Huafu Chen. Deep natural image reconstruction from human brain activity based on conditional progressively growing generative adversarial networks. 37(3):369–379, a. ISSN 1673-7067. doi: 10.1007/s12264-020-00613-4. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7954952/.

Changwei Gong, Changhong Jing, Xuhang Chen, Chi Man Pun, Guoli Huang, Ashirbani Saha, Martin Nieuwoudt, Han-Xiong Li, Yong Hu, and Shuqiang Wang. Generative AI for brain image computing and brain network computing: a review. 17:1203104. ISSN 1662-4548. doi: 10.3389/fnins.2023.1203104. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10293625/.

Yingying Peng. A comparative analysis between GAN and diffusion models in image generation. 5: 189–195. doi: 10.62051/0f1va465.

Ahmed G. Habashi, Ahmed M. Azab, Seif Eldawlatly, and Gamal M. Aly. Generative adversarial networks in EEG analysis: an overview. 20(1):40. ISSN 1743-0003. doi: 10.1186/s12984-023-01169-w. URL https://doi.org/10.1186/s12984-023-01169-w.

Rahul Mishra and Arnav Bhavsar. EEG classification for visual brain decoding via metric learning. In *Bioimaging*, pages 160–167. URL https://pdfs.semanticscholar.org/9e1b/bb38c39177c3dde1a15ee17fca9fa3a4a776.pdf.

Kay Gregor Hartmann, Robin Tibor Schirrmeister, and Tonio Ball. EEG-GAN: Generative adversarial networks for electroencephalograhic (EEG) brain signals. URL http://arxiv.org/abs/1806.01875.

Ruiqi Yang and Eric Modesitt. ViT2eeg: Leveraging hybrid pretrained vision transformers for EEG data. URL http://arxiv.org/abs/2308.00454.

Yulia Lerner, Talma Hendler, Dafna Ben-Bashat, Michal Harel, and Rafael Malach. A hierarchical axis of object processing stages in the human visual cortex. 11(4):287–297. ISSN 1047-3211. doi: 10.1093/cercor/11.4.287. URL https://doi.org/10.1093/cercor/11.4.287.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. URL http://arxiv.org/abs/2010.11929.

Zhangfang Hu, Lingxiao He, and Haoze Wu. A multi-feature fusion transformer neural network for motor imagery EEG signal classification. 31(4). URL https://www.engineeringletters.com/issues_v31/issue_4/EL_31_4_54.pdf.

Gourav Siddhad, Anmol Gupta, Debi Prosad Dogra, and Partha Pratim Roy. Efficacy of transformer networks for classification of raw EEG data. 87:105488. ISSN 17468094. doi: 10.1016/j.bspc.2023.105488. URL http://arxiv.org/abs/2202.05170.

Prince Patel, Hari Kishan Kondaveeti, Santosh Kumar Satapathy, and Namya Vyas. Detection of schizophrenia based on EEG signal using vision transformer techniques. In *2023 2nd International Conference on Futuristic Technologies (INCOFT)*, pages 1–6. doi: 10.1109/INCOFT60753.2023.10425650. URL https://ieeexplore.ieee.org/abstract/document/10425650.

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. 2(11):1019–1025. ISSN 1097-6256. doi: 10.1038/14819.

Alessandra Angelucci and Paul C. Bressloff. Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate v1 neurons. In S. Martinez-Conde, S. L. Macknik, L. M. Martinez, J. M. Alonso, and P. U. Tse, editors, *Progress in Brain Research*, volume 154 of *Visual Perception*, pages 93–120. Elsevier. doi: 10.1016/S0079-6123(06)54005-1. URL https://www.sciencedirect.com/science/article/pii/S0079612306540051.

Alvy Ray Smith. Color gamut transform pairs. *SIGGRAPH Comput. Graph.*, 12(3):12–19, aug 1978. ISSN 0097-8930. doi: 10.1145/965139.807361. URL https://doi.org/10.1145/965139.807361.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. URL http://arxiv.org/abs/2006.03677.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. pages 248–255. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/document/5206848/. Conference Name: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops) ISBN: 9781424439928 Place: Miami, FL Publisher: IEEE.

Jialu Huang, Jing Liao, and Sam Kwong. Unsupervised image-to-image translation via pre-trained stylegan2 network. 24:1435–1448, b. URL https://ieeexplore.ieee.org/abstract/document/9380493/?casa_token=p8jcEJ0D91AAAAAA:pldvelNbqOsMeMa1wfYMKeOnYZoDnFhIVUMImT_ugkTghOc3rYwrXF72pakakxDqEJAgvqkr. Publisher: IEEE.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. 8(1):15037. ISSN 2041-1723. doi: 10.1038/ncomms15037. URL https://www.nature.com/articles/ncomms15037. Publisher: Nature Publishing Group.

James J. DiCarlo, Davide Zoccolan, and Nicole C. Rust. How does the brain solve visual object recognition? 73(3):415–434. ISSN 1097-4199. doi: 10.1016/j.neuron.2012.01.010.

Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. 29(3):411–426. ISSN 1939-3539. doi: 10.1109/TPAMI.2007.56. URL https://ieeexplore.ieee.org/document/4069258. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau, and Tobias Meisen. Ablation studies in artificial neural networks. URL http://arxiv.org/abs/1901.08644.

Praveen Tirupattur, Yogesh Singh Rawat, Concetto Spampinato, and Mubarak Shah. ThoughtViz: Visualizing human thoughts using generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, MM '18, pages 950–958. Association for Computing Machinery. ISBN 978-1-4503-5665-7. doi: 10.1145/3240508.3240641. URL https://doi.org/10.1145/3240508.3240641.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. 9(86):2579–2605. ISSN 1533-7928. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Barbara G. Tabachnick and Linda S. Fidell. Multivariate analysis of variance (MANOVA). In Miodrag Lovric, editor, *International Encyclopedia of Statistical Science*, pages 902–904. Springer. ISBN 978-3-642-04898-2. doi: 10.1007/978-3-642-04898-2_394. URL https://doi.org/10.1007/978-3-642-04898-2_394.

Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, and Mubarak Shah. Generative adversarial networks conditioned by brain signals. In *Proceedings of the IEEE international conference on computer vision*, pages 3410–3418. URL http://openaccess.thecvf.com/content_iccv_2017/html/Palazzo_Generative_Adversarial_Networks_ICCV_2017_paper.html.

Xiao Zheng, Wanzhong Chen, Mingyang Li, Tao Zhang, Yang You, and Yun Jiang. Decoding human brain activity with deep learning. 56:101730. URL https://www.sciencedirect.com/science/article/pii/S1746809419303118. Publisher: Elsevier.

Jianmin Jiang, Ahmed Fares, and Sheng-Hua Zhong. A brain-media deep framework towards seeing imaginations inside brains. 23:1454–1465, b. ISSN 1941-0077. doi: 10.1109/TMM.2020.2999183. URL https://ieeexplore.ieee.org/document/9105088. Conference Name: IEEE Transactions on Multimedia.

Prajwal Singh, Pankaj Pandey, Krishna Miyapuram, and Shanmuganathan Raman. EEG2image: Image reconstruction from EEG brain signals. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, b. doi: 10.1109/ICASSP49357.2023.10096587. URL https://ieeexplore.ieee.org/document/10096587. ISSN: 2379-190X.

J Macqueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967.