Learning Robust Deep Visual Representations from EEG Brain Recordings

Prajwal Singh[†*], Dwip Dalal[†], Gautam Vashishtha[†], Krishna Miyapuram[‡], Shanmuganathan Raman[†*]
CVIG Lab[†], BRAIN Lab[‡]
IIT Gandhinagar, India
{singh_prajwal, dwip.dalal, gautam.pv, kprasad, shanmuga}@iitgn.ac.in

Abstract

Decoding the human brain has been a hallmark of neuroscientists and Artificial Intelligence researchers alike. Reconstruction of visual images from brain Electroencephalography (EEG) signals has garnered a lot of interest due to its applications in brain-computer interfacing. This study proposes a two-stage method where the first step is to obtain EEG-derived features for robust learning of deep representations and subsequently utilize the learned representation for image generation and classification. We demonstrate the generalizability of our feature extraction pipeline across three different datasets using deep-learning architectures with supervised and contrastive learning methods. We have performed the zero-shot EEG classification task to support the generalizability claim further. We observed that a subject invariant linearly separable visual representation was learned using EEG data alone in an unimodal setting that gives better k-means accuracy as compared to a joint representation learning between EEG and images. Finally, we propose a novel framework to transform unseen images into the EEG space and reconstruct them with approximation, showcasing the potential for image reconstruction from EEG signals. Our proposed image synthesis method from EEG shows 62.9% and 36.13% inception score improvement on the EEGCVPR40 and the Thoughtviz datasets, which is better than state-of-the-art performance in GAN [1].

1. Introduction

The field of Brain-Computer Interface (BCI) has witnessed a surge in interest and research due to its potential to revolutionize the way machines are controlled through human cognition. This rapidly advancing field holds promise for transforming various domains of human-computer interaction by leveraging our understanding of human brain activity [9, 10, 27]. In this context, Electroencephalography (EEG) has emerged as a key method for recording brain activity and has garnered significant attention in the scientific community [13]. EEG offers several advantages as a non-invasive technique, providing dense temporal information about brain activity. Its practical applications span a wide range of domains, including the identification of hand motor movements [44], neurorehabilitation [1], and even the decoding of speech from brain signals [5]. Notably, extracting visual information from EEG signals has been a long-standing research focus within the BCI field [4, 28, 37, 43]. Recent progress in methods for extracting visual information from EEG signals has paved the way for exciting possibilities, such as the synthesis of images using learned EEG representations [15, 39, 40, 41]. However, existing approaches in this domain have encountered limitations regarding the quality of synthesized images and reliance on label supervision.

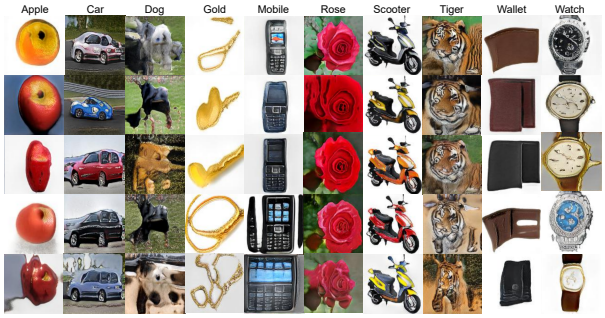


Figure 1: **EEG to Image.** Sample images generated from EEGStyleGAN-ADA using EEG signals where each image is generated with different EEG signals across different classes, Thoughtviz dataset [22, 41].

Our proposed work addresses these limitations and significantly advances EEG-based image synthesis.

1. We introduce an EEGStyleGAN-ADA framework, which improves image synthesis from EEG signals by leveraging learned EEG representations in contrastive settings. This approach generates higher-quality images, overcoming the drawbacks of previous methods and improving the state-of-the-art (SOTA) FID score by 62.9% and 36.13% on EEGCVPR40 [39] and ThoughtViz dataset [22, 41].
2. We investigate the impact of employing a similar archi-

*This work is supported by Prime Minister Research Fellowship (PMRF-2122-2557) and Jibaben Patel Chair in Artificial Intelligence.

[1] github.com/prajwalsingh/EEGStyleGAN-ADA

Figure 2: **EEG to Image.** Sample images generated from EEGStyleGAN-ADA using EEG signals where each image is generated with different EEG signals across different classes, EEGCVPR40 dataset [39].

tecture for feature extraction across all EEG datasets, thus reducing the overt dependency on the nature of the data distribution when modeling the architecture.

3. To showcase the adaptability of the acquired representations from the proposed EEG feature extraction framework, we present results involving zero-shot classification performance and a novel image-to-image translation method designed to reconstruct previously unseen images directly from the EEG space.
4. Our work further presents an innovative framework for joint representation learning that bridges the two different modalities, i.e., EEG and images, drawing inspiration from the existing Contrastive Language-Image Pre-Training (CLIP) method [32]. By fusing EEG signals and visual cues, our objective is to craft an enriched and comprehensive representation aimed to amplify performance across a spectrum of tasks, notably image classification and retrieval.

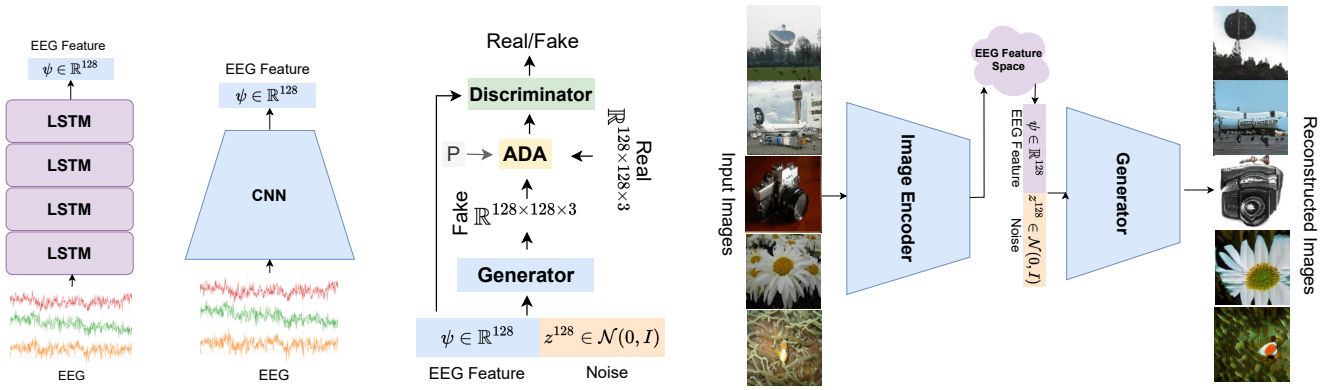
We have also conducted several experiments and ablation studies exploring EEG feature extraction using supervised and metric learning-based methods across different architectures. This rigorous evaluation allows us to assess the effectiveness of our proposed approaches and shed light on the underlying mechanisms of EEG-based representation learning. Our work addresses the challenges faced in EEG-based image synthesis and representation learning tasks and offers novel frameworks and experimental insights. By improving the quality of the synthesized images, enabling joint representation learning, creating novel frameworks for EEG-based image reconstruction, and conducting comprehensive evaluations, we aim to push the boundaries of what is possible in harnessing EEG signals for visual tasks.

2. Related Works

In the past decade, deep learning-based methods have made it possible to learn representation from complex data such as EEG, images, or text. The initial work by [39] proposed an EEGCVPR40 [39] dataset along with an LSTM-based EEG classification network for feature learning. Following this, the work by Kavasidis and Palazzo *et al.* [19, 31] uses a Generative Adversarial Network (GAN) [8] for synthesizing images from EEG features which learned using the LSTM based network proposed by [39]. Along with GAN, Kavasidis *et al.* [19] also used a Variational Autoencoder (VAE) [21] for synthesizing images from EEG signals. Their work concluded that the GAN-based method outperform the VAE in synthesizing the photorealistic images. Tirupattur *et al.* [41] proposed a GAN network that learns from a small-size dataset [22]. They have added a trainable Gaussian layer in the network that learns mean μ and variance σ of the EEG feature, preventing the discriminator network from overfitting. The work by Mishra *et al.* [24] uses an attention-based GAN network along with a trainable Gaussian layer for synthesizing images from small-size EEG dataset [22]. Both the work [41, 22] use the pre-trained image classification network for training the generator in GAN. In contrast, the work by Singh *et al.* [38] uses a metric learning-based approach for feature EEG extraction and modifies the GAN training strategy to use Differentiable Data Augmentation (DiffAug) [46] method for overcoming the problem of small-size EEG dataset. This also reduces the network complexity, i.e., a trainable Gaussian layer and a pre-trained image encoder are not required for training the generator in GAN.

3. Method

This work aims to address the problem of EEG to image reconstruction by addressing three questions: 1) What are the different strategies for extracting visual information from the EEG data? 2) Can we reconstruct the images with rich information from extracted EEG features? and 3) How can we jointly train the EEG-Image model for tasks such as EEG-based image retrieval? To address these questions, we have performed experiments and ablation studies across three different datasets EEGCVPR40 [39], ThoughtViz [22, 41] and Object [17] using different architectures with varying



(a) (b) (c) (d)

Figure 3: (a) and (b) shows the LSTM and CNN architecture used for $128D$ feature extraction from EEG signal. (c) StyleGAN-ADA [18] architecture with modification for EEG-based conditioning. (d) Illustrate the framework that transforms unseen images into learned EEG space, and then images are reconstructed from EEG features using the pre-trained generator network.

loss functions.

3.1. Feature Extraction from EEG Data

Extracting the features from the EEG data is an important step for problems like classification, reconstruction, or synthesizing images from EEG. Owing to the importance of feature extraction, several supervised or self-supervised methods have been used in the past. The following works [39, 40, 48, 15, 20, 47], use the supervised classification method for feature extraction. Supervised classification methods are preferable if the test data distribution overlaps with the train data distribution, which is not always the case with the EEG dataset. The issue can be overcome with self-supervised/metric-based learning and is addressed in these works [14, 26, 23, 38]. In [14, 26], a pre-trained image encoder is used to extract the features from the image, and an EEG encoder is trained to learn the feature distribution of images using regression and KL-divergence [16]. The work [23, 38] uses the metric learning-based method for feature learning, where triplet loss [36] is used to train the EEG encoder.

$$\min_{\theta} \mathbb{E}\left[||f_\theta(x^a) - f_\theta(x^p)||_2^2 - ||f_\theta(x^a) - f_\theta(x^n)||_2^2 + \delta\right] \quad (1)$$

Where f_θ is our encoder and $x_i \in \mathbb{R}^{N\times C}$ is EEG input of N time-steps and C channels. In this work, we have also used the triplet loss for feature learning with semi-hard triplets. The semi-hard triplets prevent the encoder network from generating similar representations for all the data, and it enforces the learning of discriminative features. In Eq.1, x^a is a anchor, x^p is a positive sample and x^n is a negative sample. The δ is the margin distance between the positive and negative samples. The semi-hard triplets have the following property: $||f_\theta(x^a) - f_\theta(x^p)|| < ||f_\theta(x^a) - f_\theta(x^n)|| < (||f_\theta(x^a) - f_\theta(x^p)|| + \delta)$.

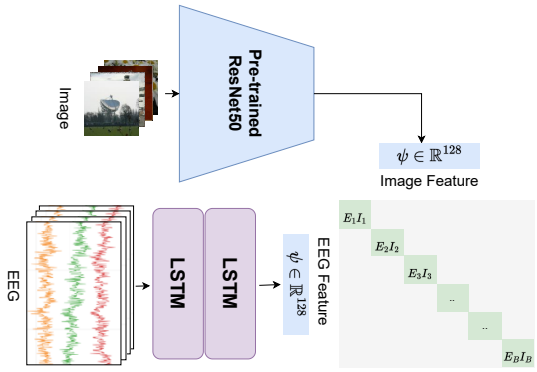


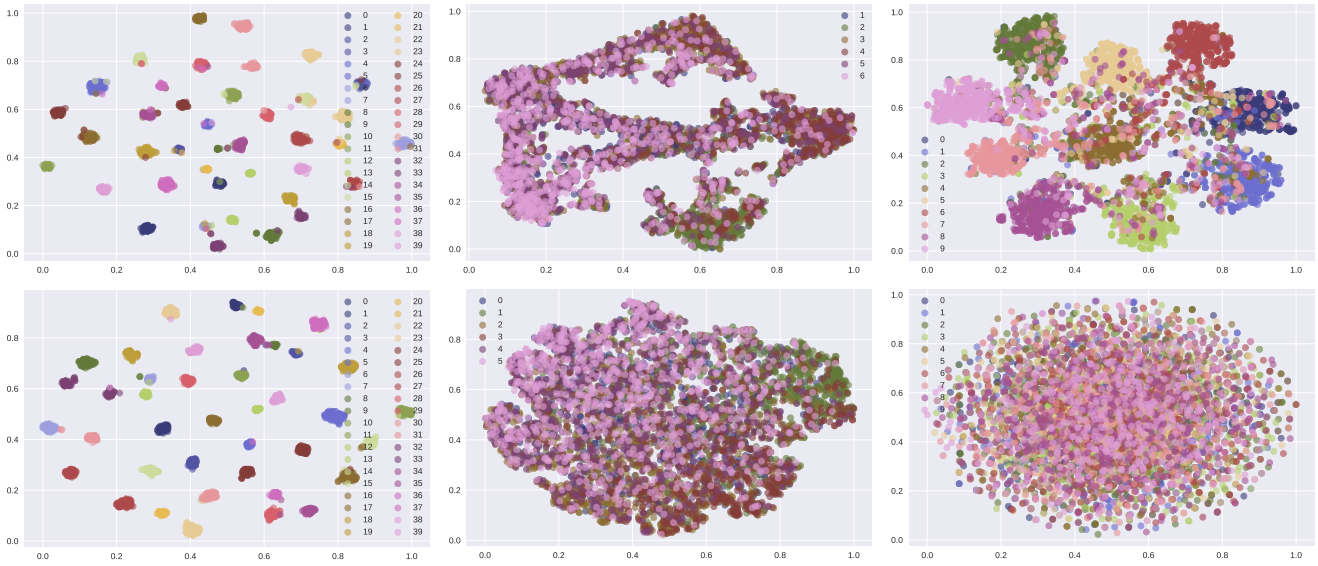
Figure 4: **EEGClip.** Illustrate the architecture used for joint representation learning of EEG and image based on [32].

3.2. Generating Image from EEG Feature

The first work by Kavasidis and Palazzo *et al.* [31, 19] uses the GAN-based method to synthesize images using EEG features. Following this, [20, 48, 7] proposes modifying the GAN architecture to improve the image synthesis quality. In recent years, image synthesis with GAN has reached the limit of photorealistic images, which is indistinguishable from real images [18, 35]. Building on this, we proposed a framework for synthesizing images from EEG features using StyleGAN-ADA network [18]. As shown in Fig.3, it takes feature vector and noise sampled from iso-tropic Gaussian distribution as input and synthesizes the desired image. StyleGAN-ADA [18] uses adaptive discriminator augmentation, which helps the discriminator learn with limited data by augmenting real images at training time.

3.3. Joint Space Learning for EEG and Image

So far, the work has been using different networks for learning EEG and image representation space with the help



(a) EEGCVPR40 Dataset [40] (b) Object Dataset [17] (c) ThoughtViz Dataset [22, 41]

Figure 5: Figure illustrates EEG cluster learned with LSTM (top) and CNN (bottom) architecture using triplet loss. K-Means score (a) $0.96, 0.98$ (b) $0.41, 0.35$ (c) $0.72, 0.12$. It shows that with a decrease in the timestep size of the EEG signal, the performance of CNN is also degrading. We kept the LSTM and CNN architecture the same across all the experiments to show the architecture's generalizability, which was not the case with previous methods where architectures were tailored according to the dataset.

of supervised and self-supervised methods. The modality of EEG signals and images is completely different, which makes learning the joint representation a non-trivial task. Few works have addressed the problem of joint representation learning between EEG and image. The work [40, 14, 26] uses a pre-trained image encoder to generate a representation for images equivalent to an EEG signal and train an EEG encoder network to regress the image feature vector. In work by Palazzo *et al.* [30], they trained the EEG encoder in a contrastive setting with a triplet loss instead of regressing the image feature vector. This work utilizes the CLIP [32] based method for joint representation learning of EEG and image data.

We used a pre-trained ResNet50 [11] as an image encoder and a multilayer LSTM network as an EEG feature encoder. During training, we freeze the weight of ResNet50 and only update the weights of the LSTM network. We have used CLIP-based loss for training the complete pipeline. As shown in Fig.4, each EEG-image pair is treated as a positive sample (diagonal elements) and the rest as a negative sample (non-diagonal elements). Similar to ours [45] uses the CLIP [32] for joint representation learning, but their problem statement differs from ours where they aim to learn representation for image encoding by training GAN and later train the EEG encoder using the contrastive method for EEG based image retrieval. We have used a pre-trained image encoder for the EEG-based image retrieval tasks.

4. Experiment and Results

In the first part of this section, we will discuss all the datasets used for training and testing. In the second part, we explain all the training regimes used for EEG feature learning and StyleGAN-ADA [18] along with ablation studies. The later part of this section discusses the joint space representation learning EEG-Image CLIP [32] model.

4.1. Datasets

We have used three datasets for training and testing the EEG representation learning.

EEGCVPR40 Dataset [39]. The dataset consists of EEG-Image pair for 40 classes, which is a subset of the ImageNet [6] dataset. While recording the brain activity EEG signals, participants were shown 50 images from each class for 0.5 seconds. The EEG device consists of 128 channels, and after preprocessing, the length of each EEG signal becomes 440 timesteps. Approximately 11800 EEG-Image pairs are there in the final dataset after discarding the bad samples.

ThoughtViz Dataset [22, 41]. This is a small-scale dataset curated by Kumar *et al.* [22]. It consists of 10 different object classes, which is a subset of the ImageNet [6]. To collect the dataset, each participant was asked to visualize one of these 10 different classes. The EEG device is of 14 channels, and a total of 23 participants' brain activity was recorded. After pre-processing, each EEG signal becomes 32 timesteps. In this work, we give it the alias ThoughtViz

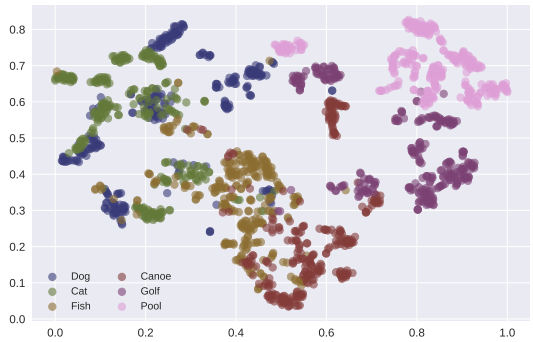


Figure 6: **Unseen EEG Clustering.** Shows a t-SNE plot of 6 unseen categories features with a k-means accuracy of 0.62. It also shows the zero-shot classification generalizability of EEG features using the proposed method.

Method	SVM	kNN	K-Means
Cogni-Net [26]	0.78	0.725	-
EEGLSTM (ours)	**0.93**	**0.86**	0.625

Table 1: **Unseen EEG**. Shows the feature generalization capability of a network when trained using triple-loss. In this case, the network is trained on 34 classes and tested 6 unseen classes of EEGCVPR40 [40] dataset.

because it was first used by Tirupattur *et al.* [41] for EEG to image synthesis work and to avoid confusion with other Object [17] dataset.

Object Dataset [17]. The dataset consists of 6 classes, each with 12 images that are shown to 10 participants, and EEG signals are recorded using a 128 channel device. The 6 classes include a human body (HB), human face (HF), animal body (AB), animal face (AF), fruit vegetable (FV), and an inanimate object (IO). After pre-processing, the final EEG data is of size 124×32.

4.2. Extracting EEG Features

Supervised. We first train the feature extraction network using a supervised setting, where we use the label associated with each signal to penalize the network if it makes a wrong prediction. We have used two different architectures, as shown in Fig.3(a,b). We train both CNN and LSTM networks for all three datasets with label supervision. We observe that networks trained in supervision have very low k-means accuracy, which shows the learned features are not linearly separable except for the CNN network trained on the EEGCVPR40 [39] dataset.

Triplet Loss. In this regime, we train both LSTM and CNN networks shown in Fig.3(a,b) using triplet loss [36] for all three datasets. Training networks using triplet loss helps them learn discriminative features, which leads to better k-means accuracy, as shown in Table 2. We have also shown the t-SNE [42] plot of the learned representation of all three

Dataset	Method	Accuracy	K-Means	SVM
EEGCVPR40 [39]	LSTM Encoder [40]	0.829	0.45	0.47
	DML [14]	0.977	-	-
	LSTM-CNN [48]	0.944	-	-
	BioLSTM [15]	0.991	-	-
	NeuroVision [20]	0.988	-	-
	EEGLSTM (ours)	**0.983**	0.961	0.962
Object [17]	BioLSTM [15]	0.611	-	-
	ERP-LSTM [47]	0.66	-	-
	EEGLSTM (ours)	0.41	0.40	0.401
ThoughtViz [22, 41]	ThoughtViz [41]	0.729	0.18	0.19
	SiameseCNN [23]	0.899	-	-
	EEG2Image [38]	0.55	0.52	-
	EEGLSTM (ours)	**0.741**	0.721	0.724

Table 2: **Clustering and Linear evaluation**. Comparison of different methods and loss types used for feature extraction from EEG signals across different datasets. In some cases, the triplet loss-based method outperforms the network trained with label supervision.

datasets in Fig.5 using the LSTM and CNN network. We further finetuned the network on all three datasets and reported the accuracy for comparison with other methods. It is vital to note that in order to show generalization in the EEG feature extraction task similar to using ResNet50 [11] for image feature extraction across various datasets and applications, we have used the same LSTM/CNN architecture across all the EEG datasets. This explains the reason for having lower fine-tuning accuracy in some datasets. To further support this claim, for EEGCVPR40 [39] dataset under the same training regime, it has k-means accuracy of 98% for CNN and 96% for LSTM architectures.

Unseen Data. In order to show the generalizability of our learned representations using triplet loss across unseen classes, we have compared our method with [26] for this. In this regime, the network is trained on 34 classes from EEGCVPR40 [39] dataset and tested on the remaining 6 classes, which are a dog, cat, fish, canoe, golf, and pool. Compared to [26], a pre-trained image network is not required. As shown in Table 1, our method performs better and has a higher SVM [12] and kNN [25] score. We have also shown t-SNE [42] plot for all the 6 unseen classes learned features Fig.6.

Image to Image. To study the effect of feature generalization, we performed another experiment, mapping the visual image features into the learned EEG manifold. We have learned the EEG feature space using triplet loss. The concept of mapping visual features stems from the notion that these mapped features can emulate the neural processes involved in human scene understanding [39]. These transformed image features were then used for image synthesis using EEGStyleGAN-ADA to show the generalization ability of the proposed network. The qualitative result of image synthesis is shown in Fig.7. The EEG features generated from unseen images can reconstruct the images with high fidelity.

	Method	IS ↑	FID ↓	KID ↓
EEGCVPR40	Brain2Image-VAE [19]	4.49	-	-
	Brain2Image-GAN [31, 19]	5.07	-	-
	NeuroVision [20]	5.15	-	-
	Improved-SNGAN [48]	5.53	-	-
	DCLS-GAN [7]	6.64	-	-
	EEGStyleGAN-ADA (ours)	**10.82**	174.13	0.065
ThoughtViz	AC-GAN [29]	4.93	-	-
	ThoughtViz [41]	5.43	-	-
	NeuroGAN [24]	6.02	-	-
	EEG2Image [38]	6.78	-	-
	EEGStyleGAN-ADA (ours)	**9.23**	109.49	0.039

Table 3: Comparison of Inception Score (on all classes) of EEGCVPR40 [39] and ThoughtViz dataset [22, 41]. For EEGStyleGAN-ADA, we have also calculated Frechet Inception Distance (FID) and Kernel Inception Distance (KID).

4.3. Image Synthesis

EEG Based Conditioning. We have significantly improved EEG to image generation by training StyleGAN-ADA [18], a state-of-the-art generative model. To tailor it specifically for EEG data, we introduced modifications to the existing StyleGAN-ADA pipeline, resulting in a framework we named EEGStyleGAN-ADA (Fig. 3(c)). Our approach incorporates a pre-trained LSTM network with triplet loss to extract EEG features, which are concatenated with the noise vector sampled from an isotropic Gaussian distribution. This combined input is then fed into the EEGStyleGAN-ADA network for image synthesis. For training the network, we employed the 'cifar' hyperparameters, leveraging data available from the EEGCVPR40 [39] and ThoughtViz [22, 41] datasets. The synthesized images produced by our proposed framework, as depicted in Fig. [1, 2], exhibit diversity and maintain a high level of fidelity when compared to previous methods.

To quantitatively evaluate the performance of our approach, we employed the Inception score [34], a commonly used metric in generative models. Comparative analysis with existing EEG to image synthesis networks, as summarized in Table 3, reveals that our proposed method outperforms them in terms of the Inception score. Furthermore, we report the Frechet Inception Distance (FID) [3] and Kernel Inception Distance (KID) [2] scores, providing additional insights into the quality and diversity of the generated images.

Class Based Conditioning. To show the effectiveness of the proposed EEGStyleGAN-ADA, we performed an ablation study where instead of giving an EEG signal, we used one-hot class conditions only. The EEGCVPR40 [39] dataset consists of 40 classes, each with 30-40 images, making it difficult to learn using a conditional GAN. To further verify this claim, we perform the experiment with the current state-of-the-art, NoisyTwins [33], for the long-tail conditional generation. As shown in Fig.9, the best FID score we achieved is 105.5, and the qualitative result shows even after using the SOTA model for image synthesis using one-hot class conditioning on the EEGCVPR dataset lacks photo-realistic effect. This implies that the photorealistic images we are synthesizing with the proposed EEGStyleGAN-ADA are the best among all the GAN state-of-the-art methods for EEG-based image generation.

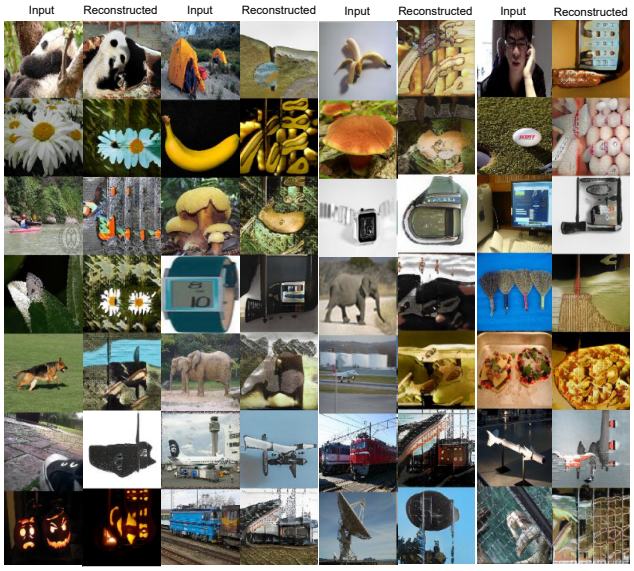


Figure 7: **Image to Image.** Shows the result of Image to Image translation. Here, instead of using an EEG signal from EEGCVPR40 [39] dataset, its equivalent images are used, which are transformed into EEG representation space, and later, the image is reconstructed with approximation using a pre-trained generative network.

4.4. Joint Representation Space Learning

This section of our work presents EEGClip, a novel framework for joint representation learning between EEG signal and images built upon the CLIP Model [32]. To evaluate the effectiveness of our method, we conducted experiments using the EEGCVPR40 dataset [39], which offers a significantly larger number of (EEG, Image) pairs compared to the ThoughtViz [22, 41] and Object [17] datasets. We performed several experiments to investigate the impact of batch size and the number of training epochs on learning joint representations. Due to computational constraints, we considered batch sizes of $16, 32, 64$ and trained the model for varying numbers of epochs, ranging from $64, 128, 256, 512, 1024, 2048$. The results of these experiments, as summarized in Table 4, present the top-K recall rates for $K \in \{1, 5, 10\}$. Our findings indicate that the proposed EEGClip framework achieved superior performance when trained with a batch size of 64 and for 2048 epochs. This configuration yielded the highest recall rates for different values of K. To provide a visual representation of the

Batch Size\Epochs		Fine Tuning Top-K Accuracy [@1 / @5 / @10]					
		64	128	256	512	1024	2048
EEG	16	0.26/0.45/0.60	0.37/0.59/0.71	0.51/0.75/0.81	0.59/0.80/0.85	0.69/0.88/0.92	0.73/0.90/0.93
	32	0.32/0.53/0.68	0.43/0.68/0.79	0.53/0.80/0.87	0.61/0.87/0.90	0.72/0.92/0.95	0.77/0.94/0.96
	64	0.34/0.54/0.69	0.44/0.67/0.82	0.57/0.80/0.85	0.68/0.89/0.93	0.76/0.94/0.97	0.79/0.96/0.98
Image	16	0.78/0.88/0.91	0.83/0.90/0.92	0.87/0.93/0.95	0.89/0.96/0.98	0.91/0.97/0.98	0.92/0.97/0.99
	32	0.80/0.91/0.94	0.84/0.92/0.95	0.90/0.96/0.98	0.92/0.97/0.99	0.92/0.96/0.98	0.93/0.96/0.99
	64	0.84/0.94/0.95	0.88/0.95/0.98	0.91/0.97/0.98	0.93/0.98/0.99	0.94/0.97/0.99	0.95/0.99/1.0

Table 4: **EEGClip.** Results of EEGClip network finetuned on EEGCVPR40 [39] dataset for both EEG and image classification.

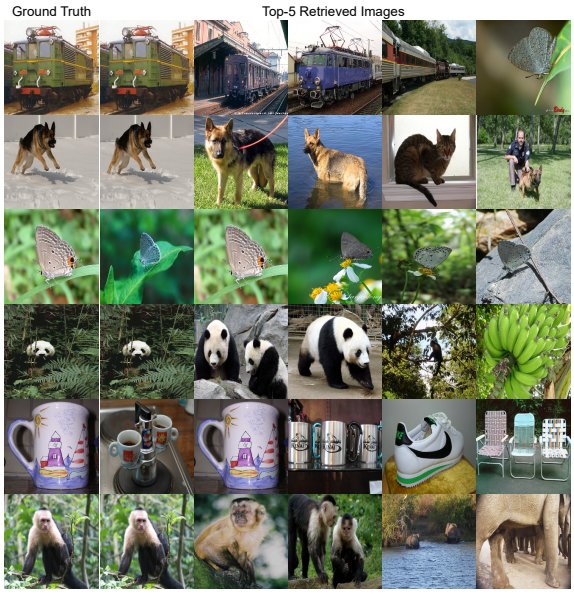


Figure 8: **Image Retrieval using EEG.** Shows the top-5 retrieved images for the given EEG signal from test data of EEGCVPR40 [39]. The pre-trained weights of EEGClip are used to extract the feature for both the image and EEG. Here, ground truth shows the expected image equivalent to the given EEG.

effectiveness of EEGClip, we present image retrieval results from EEG in Fig.8. These results demonstrate the ability of our framework to retrieve relevant images based on EEG input accurately.

4.5. Experiment on EEGCVPR40 Filter Dataset (5-95 Hz)

Based on the section 4.4, for EEGClip, the optimal performance is achieved when using a batch size of 64. Consequently, we employed the same batch size when conducting experiments on the EEGCVPR40 filter dataset (5-95 Hz) [39]. We have compared our method with Palazzo *et al.* [30], where they conducted experiments on the same filter dataset and achieved an accuracy of 60.4% for EEG classification and 94% for image classification. Our best-performing model, which corresponds to a batch size of 64 and 2048 epochs (as shown in Table 5), achieved an accuracy of 64% for EEG classification and 94% for image classification.

4.6. Quantitative Performance of Image Retrieval Task

We employed the EEGClip model (Fig. 4) based on EEG inputs for the Image Retrieval task. The outcomes obtained from our experiments have been presented in the main paper. However, to provide a more comprehensive analysis of the results, we utilized two specific metrics: mean Reciprocal Rank (MRR) and mean Average Precision (mAP) for ranking evaluation. The MRR metric assesses the effectiveness of the retrieval model in accurately ranking the unique visually-cued instances, while the mAP metric evaluates the retrieval model's ability to capture all relevant visual cues. These relevant visual cues correspond to images that belong to the same semantic class as the correct match. For the EEGCVPR40 [39] dataset, our approach yielded an MRR of 0.7427 and an mAP of 0.6689. These scores were achieved using the model associated with a batch size of 64 and trained for 2048 epochs.

5. Discussion

In this paper, we addressed the problem of EEG-to-image reconstruction and presented a comprehensive method to extract visual information from EEG data, synthesize images using extracted EEG features, and jointly train an EEG-Image model for tasks such as EEG-based image retrieval. We conducted experiments and ablation studies on three different datasets: EEGCVPR40 [39], ThoughtViz [22, 41], and Object [17], using various architectures and loss functions.

We first discussed different strategies for feature extraction from EEG data, including supervised and self-supervised methods. We compared supervised classification methods with self-supervised/metric-based learning approaches and found that the latter yielded more discriminative and generalizable features, particularly using triplet loss. We demonstrated the same through improved k-means accuracy, t-SNE visualizations, and zero-shot classification.

Next, we explored the generation of images from EEG features using the GAN framework. For this, we proposed

Fine Tuning Top-K Accuracy [@1 / @5 / @10]							
Batch Size\Epochs		64	128	256	512	1024	2048
EEG	64	0.28/0.39/0.48	0.37/0.53/0.69	0.45/0.66/0.81	0.53/0.77/0.86	0.59/0.82/0.89	0.64/0.86/0.92
Image	64	0.78/0.88/0.92	0.83/0.92/0.95	0.87/0.94/0.97	0.90/0.95/0.98	0.92/0.96/0.99	0.94/0.98/0.99

Table 5: **EEGClip.** Results of EEGClip network finetuned on EEGCVPR40 filter dataset (5-95 Hz) [39] for both EEG and image classification.

Figure 9: **Class Based Conditioning.** To show the complexity of generating images using EEG from EEGCVPR40 [39] dataset. We train a variation of StyleGAN2 known as NoisyTwins [33] for learning the image generation on long-tail conditional datasets or conditional datasets with fewer images, which is the case with the EEGCVPR40 dataset. Instead of using EEG, we have used one-hot encoding for conditioning. The best FID score we achieved is 105.5 with class label condition.

EEGStyleGAN-ADA, which incorporated EEG features and noise vectors to synthesize diverse and high-fidelity images. Our method outperformed previous EEG-to-image synthesis networks, with 62.9% and 36.13% inception score improvement on the EEGCVPR40 [39] dataset and Thoughtviz [22, 41] dataset, which is better than state-of-the-art performance using GAN. We have shown that achieving the photorealistic effect is not trivial with the help of class-based conditioning 4.3.

Furthermore, we investigated joint representation space learning for EEG and image using the proposed EEGClip framework. We achieved significant improvements in joint representation learning by freezing the weights of a pre-trained image encoder and training an EEG feature encoder using CLIP-based loss. Upon examining the effects of batch size and epoch numbers, we observed a direct correlation between increased batch size and enhanced performance, peaking at a batch size of 64 and 2048 epochs, yielding scores of 79% and 95% for top@1 for EEG and image, respectively. However, extending the epoch count beyond this point showed no significant improvement. EEGClip has shown 5.96% improvement over the previous state-of-the-art joint representation learning method.

Limitations. The proposed work has a few limitations. 1) Although we have used the same architecture for EEG feature extraction across all datasets, it is still an open problem to achieve SOTA performance using a single architecture. 2) In EEG-based image synthesis, we outperform the previous methods. Still, the quality of images in a limited dataset regime can be improved with better GAN training strategies, and we can further utilize the same for EEG-based image reconstruction.

6. Conclusion

In this study, our primary objective was to enhance the quality of image synthesis from EEG signals. To achieve this, we introduced EEGStyleGAN-ADA, a framework capable of leveraging small and large-sized EEG datasets to generate high-resolution images (128×128) directly from EEG signals that outperform the previous state-of-the-art.

In addition to image synthesis, we proposed a joint representation learning framework that bridges the gap between EEG and image representation. By combining the power of both modalities, we obtained a richer and more comprehensive representation, enabling us to perform image retrieval tasks using EEG signals. Our exhaustive experiments substantiated the effectiveness of this joint representation learning approach and showcased its potential in real-world applications. Furthermore, we have also shown the triplet loss-based feature extraction method's zero-shot classification capability.

Our future endeavors will further improve EEG-based image synthesis, explore novel techniques, and refine existing frameworks. Additionally, we aim to expand our investigation to encompass the emerging field of EEG-based video analysis, leveraging temporal dynamics to generate and analyze visual content.

References

[1] Mahdi Bamdad, Homayoon Zarshenas, and Mohammad A Auais. Application of bci systems in neurorehabilitation: a scoping review. *Disability and Rehabilitation: Assistive Technology*, 10(5):355–364, 2015.

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[3] Naresh Babu Bynagari. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Asian Journal of Applied Science and Engineering*, 8:25–34, 2019.

[4] Thomas A Carlson, Hinze Hogendoorn, Ryota Kanai, Juraj Mesik, and Jeremy Turret. High temporal resolution decoding of object position and category. *Journal of vision*, 11(10):9–9, 2011.

[5] Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech from non-invasive brain recordings. *arXiv preprint arXiv:2208.12266*, 2022.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Ahmed Fares, Sheng-hua Zhong, and Jianmin Jiang. Brain-media: A dual conditioned and lateralization supported gan (dcls-gan) towards visualization of image-evoked brain activities. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1764–1772, 2020.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[9] Andrea M Green and John F Kalaska. Learning to move machines with the mind. *Trends in neurosciences*, 34(2):61–75, 2011.

[10] Christoph Guger, Werner Harkam, Carin Hertnaes, and Gert Pfurtscheller. Prosthetic control by an eeg-based brain-computer interface (bci). In *Proc. aaate 5th european conference for the advancement of assistive technology*, pages 3–6. Citeseer, 1999.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998.

[13] Herbert H Jasper. Report of the committee on methods of clinical examination in electroencephalography: 1957. *Electroencephalogr Clin Neurophysiol*, 10:370–375, 1958.

[14] Jianmin Jiang, Ahmed Fares, and Sheng-Hua Zhong. A context-supported deep learning framework for multimodal brain imaging classification. *IEEE Transactions on Human-Machine Systems*, 49(6):611–622, 2019.

[15] Jianmin Jiang, Ahmed Fares, and Sheng-Hua Zhong. A brain-media deep framework towards seeing imaginations inside brains. *IEEE Transactions on Multimedia*, 23:1454–1465, 2020.

[16] James M Joyce. Kullback-leibler divergence. In *International encyclopedia of statistical science*, pages 720–722. Springer, 2011.

[17] Blair Kaneshiro, Marcos Perreau Guimaraes, Hyung-Suk Kim, Anthony M Norcia, and Patrick Suppes. A representational similarity analysis of the dynamics of object processing using single-trial eeg classification. *Plos one*, 10(8):e0135697, 2015.

[18] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

[19] 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.

[20] Sanchita Khare, Rajiv Nayan Choubey, Loveleen Amar, and Venkanna Udutalapalli. Neurovision: perceived image regeneration using cprogan. *Neural Computing and Applications*, 34(8):5979–5991, 2022.

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] Pradeep Kumar, Rajkumar Saini, Partha Pratim Roy, Pawan Kumar Sahu, and Debi Prosad Dogra. Envisioned speech recognition using eeg sensors. *Personal and Ubiquitous Computing*, 22:185–199, 2018.

[23] Rahul Mishra and Arnav Bhavsar. Eeg classification for visual brain decoding via metric learning. In *BIOIMAGING*, pages 160–167, 2021.

[24] Rahul Mishra, Krishan Sharma, RR Jha, and Arnav Bhavsar. Neurogan: image reconstruction from eeg signals via an attention-based gan. *Neural Computing and Applications*, pages 1–12, 2022.

[25] Antonio Mucherino, Petraq J. Papajorgji, and Panos M. Pardalos. *k-Nearest Neighbor Classification*, pages 83–106. Springer New York, New York, NY, 2009.

[26] Pranay Mukherjee, Abhirup Das, Ayan Kumar Bhunia, and Partha Pratim Roy. Cogni-net: Cognitive feature learning through deep visual perception. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4539–4543. IEEE, 2019.

[27] Gernot R Muller-Putz and Gert Pfurtscheller. Control of an electrical prosthesis with an ssvep-based bci. *IEEE Transactions on biomedical engineering*, 55(1):361–364, 2007.

[28] Maryam Mustafa, Lea Lindemann, and Marcus Magnor. Eeg analysis of implicit human visual perception. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 513–516, 2012.

[29] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.

[30] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah. Decoding brain representations by multimodal learning of neural

activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3833–3849, 2020.

[31] 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, 2017.

[32] 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.

[33] Harsh Rangwani, Lavish Bansal, Kartik Sharma, Tejan Karmali, Varun Jampani, and R Venkatesh Babu. Noisytwins: Class-consistent and diverse image generation through stylegans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5987–5996, 2023.

[34] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[35] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[37] Pradeep Shenoy and Desney S Tan. Human-aided computing: utilizing implicit human processing to classify images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 845–854, 2008.

[38] 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. IEEE, 2023.

[39] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4503–4511, 2016.

[40] 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.

[41] 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*, pages 950–958, 2018.

[42] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[43] Changming Wang, Shi Xiong, Xiaoping Hu, Li Yao, and Jiacai Zhang. Combining features from erp components in single-trial eeg for discriminating four-category visual objects. *Journal of neural engineering*, 9(5):056013, 2012.

[44] Fangzhou Xu, Fenqi Rong, Jiancai Leng, Tao Sun, Yang Zhang, Siddharth Siddharth, and Tzyy-Ping Jung. Classification of left-versus right-hand motor imagery in stroke patients using supplementary data generated by cyclegan. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:2417–2424, 2021.

[45] Zesheng Ye, Lina Yao, Yu Zhang, and Silvia Gustin. See what you see: Self-supervised cross-modal retrieval of visual stimuli from brain activity. *arXiv preprint arXiv:2208.03666*, 2022.

[46] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020.

[47] Xianglin Zheng, Zehong Cao, and Quan Bai. An evoked potential-guided deep learning brain representation for visual classification. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part V 27*, pages 54–61. Springer, 2020.

[48] Xiao Zheng, Wanzhong Chen, Mingyang Li, Tao Zhang, Yang You, and Yun Jiang. Decoding human brain activity with deep learning. *Biomedical Signal Processing and Control*, 56:101730, 2020.