DECODING NATURAL IMAGES FROM EEG FOR OBJECT RECOGNITION

Yonghao Song[1], **Bingchuan Liu**[1], **Xiang Li**[1], **Nanlin Shi**[1], **Yijun Wang**[2], **Xiaorong Gao**[1*]
[1] Department of Biomedical Engineering, Tsinghua University
[2] Institute of Semiconductors, CAS
gxr-dea@tsinghua.edu.cn

ABSTRACT

Electroencephalography (EEG) signals, known for convenient non-invasive acquisition but low signal-to-noise ratio, have recently gained substantial attention due to the potential to decode natural images. This paper presents a self-supervised framework to demonstrate the feasibility of learning image representations from EEG signals, particularly for object recognition. The framework utilizes image and EEG encoders to extract features from paired image stimuli and EEG responses. Contrastive learning aligns these two modalities by constraining their similarity. Our approach achieves state-of-the-art results on a comprehensive EEG-image dataset, with a top-1 accuracy of 15.6% and a top-5 accuracy of 42.8% in 200-way zero-shot tasks. Moreover, we perform extensive experiments to explore the biological plausibility by resolving the temporal, spatial, spectral, and semantic aspects of EEG signals. Besides, we introduce attention modules to capture spatial correlations, providing implicit evidence of the brain activity perceived from EEG data. These findings yield valuable insights for neural decoding and brain-computer interfaces in real-world scenarios. Code available at https://github.com/eeyhsong/NICE-EEG.

1 INTRODUCTION

Our daily life relies on accurately and rapidly identifying objects in complex visual environments (DiCarlo & Cox, 2007). Researchers have pursued decoding natural images from brain activity, aiming to deepen our understanding of the brain and create user-friendly brain-computer interfaces (BCIs) (Kamitani & Tong, 2005; Kay et al., 2008; Gao et al., 2021). Functional magnetic resonance imaging (fMRI), which records blood oxygen level-dependent signals, has been a popular choice for categorizing objects observed by humans (Du et al., 2023; Horikawa & Kamitani, 2017; Allen et al., 2022). Despite its high spatial resolution, fMRI typically requires several seconds to provide a stable response to a single stimulus, limiting its real-time applicability in daily interactions (Lin et al., 2022). Similarly, Magnetoencephalogram (MEG), with high time resolution, has been employed for this purpose, but it is hindered by cost and large devices (Cichy et al., 2014; Hebart et al., 2023).

Electroencephalogram (EEG) has emerged as a valuable tool for decoding images based on visual-evoked brain activities (Spampinato et al., 2017). EEG has high time resolution, low cost, and good portability, but the low signal-to-noise ratio takes problems (Pan et al., 2022; Kobler et al., 2022). Some efforts have yielded impressive classification results and captured saliency maps using EEG signals (Palazzo et al., 2021). However, its flawed block-design experiments group images of the same class within one block, resulting in classification that relies on block-level temporal correlation rather than stimulus-related activity (Li et al., 2021). Some studies have achieved above-chance performance but were restricted to single-subject data (Ahmed et al., 2021). Moreover, available datasets often feature a limited number of image categories, posing challenges for real-world image decoding tasks. The latest study has significantly expanded the situation by amassing a large and diverse EEG dataset comprising 16,740 image stimuli spanning 1,854 concepts, employing rapid serial visual presentation (Gifford et al., 2022). While demonstrating the feasibility of image-to-EEG encoding and category separability, this work primarily focused on analyzing electrodes in the occipital and parietal regions, overlooking the inferior temporal cortex plays a necessary role in object recognition (Dapello et al., 2023). In addition, the work focuses on the peak of pairwise decoding about 110 ms after stimuli onset. Generally, this latency involves visual conduction and primary visual processing by a large

margin, rather than semantic understanding Xu et al. (2023). In short, the issues of multi-class object recognition and biological plausibility warrant further exploration.

Pattern recognition plays a pivotal role in EEG analysis. Existing work has predominantly relied on supervised learning, often dealing with limited data from a few categories (Cheng et al., 2022). Applying machine learning methods directly, such as support vector machines and deep neural networks, has proven challenging in obtaining intrinsic representations (Fu et al., 2022; Liu et al., 2021). People have performed self-supervised learning and leveraged information from other modalities in computer vision research (Wang et al., 2022; Saharia et al., 2022). Text replaces the traditional hard labels to enable self-supervised learning to produce superior image representations, often leading to impressive zero-shot results on downstream tasks (Radford et al., 2021). Researchers have also used contrastive learning to obtain consistent embedding of neural recordings across multiple animals (Schneider et al., 2023). Building upon these insights, we explore the intriguing possibility of recognizing visual objects from EEG signals in a self-supervised manner. Another issue pertains to the limitations of commonly employed EEG feature extractors, typically structured around convolutional layers applied separately along temporal and spatial dimensions of raw EEG signals (Schirrmeister et al., 2017; Lawhern et al., 2018; Song et al., 2023). This organization disrupts the correlations between electrode channels, hindering our perceiving spatial properties of brain activity (Ding et al., 2023).

To address the above limitations, we introduce a self-supervised framework to decode image representations from EEG signals, focusing on object recognition. Our approach uses contrastive learning as the bridge connecting image stimuli and EEG responses. We feed image-EEG pairs to the model and process them by an image encoder and an EEG encoder separately. These modalities are aligned by constraining their cosine similarity, resulting in the ability to perform zero-shot EEG decoding after training, achieving remarkable classification accuracy for previously unseen image categories. With the framework, we provide an overall resolving of the visual object recognition processing in our brain from spatial, temporal, spectral, and semantic aspects with comparative experiments. The results are consistent with established neuroscientific knowledge, further demonstrating the feasibility of EEG-based image decoding. Furthermore, we present two spatial modules with self-attention and graph attention that integrate with the EEG encoder (Dosovitskiy et al., 2021; Veličković et al., 2018), helping us to emphasize key brain regions relevant to object recognition.

Our main contributions can be summarized as follows:

- We propose a self-supervised framework for EEG-based object recognition with contrastive learning. Remarkable zero-shot performance has been achieved on large and rich datasets.
- We demonstrate the feasibility of investigating natural image information from EEG signals. Extensive experiments affirm the biological plausibility, which brings a resolving of human object recognition from temporal, spatial, spectral, and semantic aspects.
- We apply two plug-and-play modules to capture spatial correlations among EEG channels, offering evidence that the model discerns the spatial dynamics of object recognition.

2 Related Works

Decoding visual information from our brain has been a long-standing pursuit in neuroscience and computer science (Miyawaki et al., 2008; Jia et al., 2021). While some progress has been made in decoding steady-state visual stimuli, the accurate and rapid decoding of semantic information in natural images has remained a challenge (Shi et al., 2023). fMRI has been widely used to estimate semantic and shape information from visual processing in the brain (Ho et al., 2023; Takagi & Nishimoto, 2023). However, the demand for high-speed and practical applications in brain-computer interaction necessitates alternative approaches. EEG has emerged as a promising option due to its high temporal resolution and portability (Willett et al., 2021). But general performance on different subjects and biological plausibility remain unsolved (Ahmed et al., 2021). Additionally, prior approaches have often relied on supervised learning methods with limited image categories, overlooking the intrinsic relationship between image stimuli and brain responses (Shen et al., 2019; Li et al., 2021; Liu et al., 2023). These limitations hinder their practicality in real-world scenarios with generalization to new categories. Motivated by these challenges, we present an EEG-based image decoding framework that employs self-supervised learning, enabling the model to achieve zero-shot generalization in object recognition tasks, further demonstrating the feasibility.

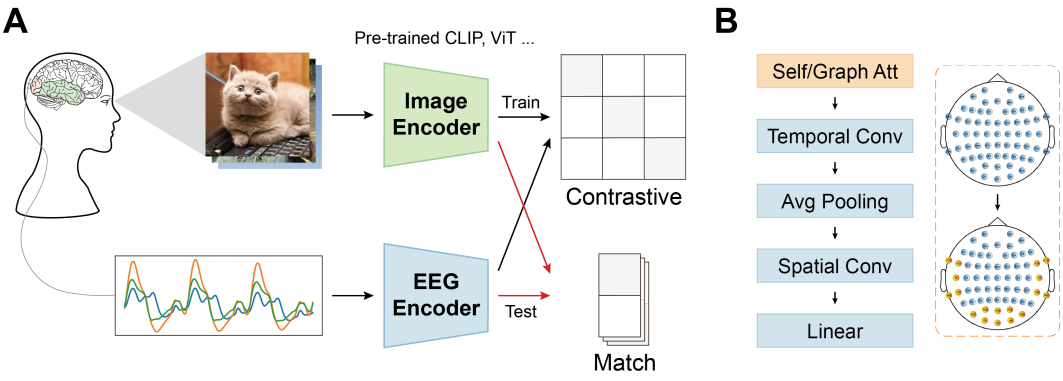

Figure 1: (A) Overall framework for EEG-based object recognition. During training, image-EEG pairs are processed by an image encoder (pre-trained) and an EEG encoder. The objective is to increase the similarity between matched pairs while decreasing it for unmatched pairs. During testing, a few unseen images of target concepts (classes) are processed in advance into templates. Then, the model obtains results by matching test data to templates. (B) Architecture of the EEG encoder. Temporal-spatial convolution is used with spatial modules, made with self and graph attention, to reveal spatial features of brain activity. The linear layer is used to adjust the feature dimension.

3 METHODS

3.1 OVERVIEW

We propose a self-supervised framework, Natural Image Contrast EEG (NICE), to decode images from EEG signals. The overall framework is depicted in Fig. 1(A).

During training, stimulus-response pairs comprising images and EEG signals are fed into the framework. The image and EEG encoder extract features from their respective modalities. Contrastive learning is employed to align these features, optimizing the similarity between matched pairs and reducing it for unmatched pairs. This self-supervised approach enables the model to learn shared information between the two modalities instead of traditional supervised learning with predefined labels. Before the test process, we use a few images that belong to the image concepts (classes) to be classified, and have not appeared in the EEG collection experiments. These images are processed and averaged to obtain one template for each concept. Next, the EEG encoder processes the test signals, and their similarity with all templates is calculated for matching results.

For the EEG encoder, we adapt temporal-spatial convolution (TSConv), which is widely used in EEG analysis, as shown in Fig. 1(B). In addition, we integrate two plug-and-play spatial modules with self-attention and graph attention. These modules are instrumental in preserving the spatial characteristics of EEG channels, reflecting intrinsic patterns of brain activity. Note that the image encoder has been pre-trained on other image datasets. The leading studies in computer vision help us with larger sample space, when we cannot easily collect numbers of brain responses.

3.2 NETWORK ARCHITECTURE

3.2.1 EEG ENCODER

Researchers usually arrange the raw EEG trials into two dimensions $C \times T$, in which C denotes electrode channels, and T denotes time samples. Convolution along the two dimensions is widely used in deep learning-based EEG analysis models. In the same way, we present a very concise network architecture as shown in Table 1. One-dimension convolution is first applied to capture the temporal features with k kernels of size $(1, m_1)$ and stride of $(1, 1)$. An average pooling layer with a kernel size of $(1, m_2)$ and stride of $(1, s)$ is introduced to alleviate overfitting and smooth the temporal features. Another one-dimension convolution is then used for spatial features keeping k kernels of size $(ch, 1)$ and stride of $(1, 1)$, where ch usually equals the number of electrodes. After

Table 1: Architecture of EEG encoder with SA or GA followed by TSConv

Layer	In	Out	Kernel	Stride	Dimension
SA & GA	$[(b,1,C,T) \rightarrow (b,1,C,T)]$, b is batch				
Temporal Conv	1	k	$(1,m_1)$	(1, 1)	$(b,k,C,T-m_1+1)$
Avg Pooling	k	k	$(1,m_2)$	$(1,s)$	$(b,k,C,(T-m_1-m_2+1)/s+1)$
Spatial Conv	k	k	$(C,1)$	$(1,1)$	$(b,k,1,(T-m_1-m_2+1)/s+1)$
Flatten&Linear	$[k*((T-m_1-m_2+1)/s+1) \rightarrow$ shape of image feature$]$				

each convolutional layer, batch normalization and exponential linear units (ELU) activation functions are used for better training and nonlinearity (Clevert et al., 2016). Finally, a linear layer is added as a projector to transform the features to the same size as the output of the image encoder.

3.2.2 Image Encoder

The image features are extracted with the pairing EEG features for contrastive learning. There have been some excellent computer vision models that can obtain image features with semantic discrimination. We prefer directly using models pre-trained on other image datasets to get a large sample space. The feature extraction parts of Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021), Vision Transfomer (ViT) (Dosovitskiy et al., 2021), and Residual Neural Networks (ResNet) (He et al., 2016) are tried separately in our framework for demonstration. All the images for stimuli can be processed in advance, thus speeding up our model training.

3.2.3 Contrastive Learning

The framework is constructed with contrastive learning, shown in Algorithm 1. The outputs of the image and EEG encoder are normalized separately. Then, the dot product is used to evaluate the similarity of all image-EEG feature pairs. A scaled temperature parameter is used to adjust distribution probability. InfoNCE loss (Radford et al., 2021; van den Oord et al., 2019) is employed as the objective function to increase the similarities between matched pairs and decrease those between unmatched pairs. After adequate training, the EEG encoder can extract representations similar to corresponding images. The self-supervised strategy allows us to learn inherent patterns from EEG signals without labels, rather than directly separate different classes with supervised learning.

Algorithm 1 Natural Image Contrast EEG framework

Input: (Image, EEG) - stimulus & response
Model: Enc_img - CLIP & Enc_eeg - TSConv

E - $(batch, channel, electrode, sample)$ ▷ batch of input EEG
I - $(batch, channel, height, width)$ ▷ batch of input images
τ - learned temperature parameter

# extract normalized representations from the raw image and EEG
E_f = Norm(Linear(Enc_eeg(E)))
I_f = Norm(Enc_img(I)) ▷ can be obtained before training

# scaled pairwise cosine similarity
logits = dot(E_f, I_f.t) * exp(τ)

# symmetric loss function
labels = arange(batch)
loss_e = cross_entropy_loss(logits, labels, axis=0)
loss_i = cross_entropy_loss(logits, labels, axis=1)
loss = (loss_e + loss_i) / 2

3.3 PLUG-AND-PLAY MODULE

3.3.1 SELF-ATTENTION

We utilize two approaches, self-attention (SA) Vaswani et al. (2017), and graph attention (GA) Veličković et al. (2018), to supplement the EEG encoder as "spatial filters" to encapsulate electrode correlations and reflect the spatial dynamics of brain activity. We use SA on the electrode channels to evaluate the spatial correlations of EEG data. The input $x_{in} \in \mathbb{R}^{C \times T}$ is linearly transformed into equal-sized with weight metrics W_q, W_k, W_v:

$$x'_{in} = \text{Softmax}(\frac{W_q x_{in} \cdot (W_k x_{in})^T}{\sqrt{d}}) \cdot W_v x_{in} \tag{1}$$

where $x'_{in} \in \mathbb{R}^{C \times T}$ denotes the output after the SA module, d, the time length of EEG data, is a scaled factor to accommodate the Softmax function.

3.3.2 GRAPH ATTENTION

We employ the GA module to update each electrode with all the others, using the implementation from Brody et al. (2022). We treat each electrode as a node $n_i \in \mathbb{R}^{1 \times T}$, $i = 1, ..., ch$, which has edges to all the other electrodes $\mathcal{N}_i$. The process to update one electrode is as follows:

$$n'_i = \alpha_{i,i} W n_i + \sum_{j \in \mathcal{N}_i} \alpha_{i,j} W n_j \tag{2}$$

where n'_i denotes each node after processing, $\alpha_{i,j}$ is attention coefficients indicating importance of node j features to node i, and W is the weight of linear transformation. Calculate $\alpha_{i,j}$ as:

$$\alpha_{i,j} = \frac{\exp(a^T \text{LeakyReLU}(W[n_i \,\|\, n_j]))}{\sum_{k \in \mathcal{N}_i \cup \{i\}} \exp(a^T \text{LeakyReLU}(W[n_i \,\|\, n_k]))} \tag{3}$$

where $a \in \mathbb{R}^{2T}$ denotes the weight of a feedforward layer for attention calculation, $()^T$ is the transposition operator, and $\|$ is the concatenation operator. LeakyReLU with a slope of 0.2 is used for nonlinearity in calculation. Residual connection is employed by integrating the input and output of the SA and GA modules separately, contributing to stable training (He et al., 2016).

4 EXPERIMENTS AND RESULTS

4.1 DATASETS AND PREPROCESSING

The dataset (Gifford et al., 2022) contains EEG data from ten participants with a time-efficient rapid serial visual presentation (RSVP) paradigm. The training set includes 1654 concepts$\times$10 images$\times$4 repetitions. The test set includes 200 concepts$\times$1 image$\times$80 repetitions. Images for training and testing appear in a pseudo-randomized order, and a target image is used to reduce eye blinks and other artifacts. Each image displays 100 ms, followed by a 100 ms blank screen. Raw EEG data filtered to [0.1, 100] Hz has 63 channels and a sample rate of 1000 Hz.

For preprocessing, we epoched EEG data into trials ranging from 0 to 1000 ms after stimuli onset. Baseline correction was performed with the mean of 200 ms pre-stimulus data. All electrodes were preserved for analysis with down-sampling to 250 Hz, and multivariate noise normalization was performed with training data (Guggenmos et al., 2018). We averaged all EEG repetitions of one image to ensure the signal-to-noise ratio and compared the impact of repetitions. Images were resized to 224$\times$224 and normalized before being processed by the image encoder.

4.2 EXPERIMENT DETAILS

Our method is implemented with PyTorch on a Geforce 4090 GPU. We randomly select 740 trials from training data as the validation set in each run of the code. Best models are saved when the validation loss reaches a minimum of 200 epochs in the training process. We compute the results of the test set once after training. It takes about 5 minutes per subject to train with a batch size of 1000, since we get the image features in advance. The k in TSConv is set to 40, m_1 to 25, m_2 to 51, and s to 5 by pre-experiments. Adam optimizer is used with the learning rate, β_1 and β_2 of 0.0002, 0.5, and 0.999, respectively. Wilcoxon Signed-Rank Test is employed to analyze the statistical significance.

Table 2: Overall accuracy (%) of 200-way zero-shot classification: top-1 and top-5

	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5		Subject 6		Subject 7		Subject 8		Subject 9		Subject 10		Ave	
Method	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5	top-1	top-5
Subject dependent - train and test on one subject																						
BraVL	6.1	17.9	4.9	14.9	5.6	17.4	5.0	15.1	4.0	13.4	6.0	18.2	6.5	20.4	8.8	23.7	4.3	14.0	7.0	19.7	5.8	17.5
NICE	12.3	36.6	10.4	33.9	13.1	39.0	16.4	47.0	8.0	26.9	14.1	40.6	15.2	42.1	20.0	49.9	13.3	37.1	14.9	41.9	**13.8**	**39.5**
NICE - SA	13.3	40.2	12.1	36.1	15.3	39.6	15.9	49.0	9.8	34.4	14.2	42.4	17.9	43.6	18.2	50.2	14.4	38.7	16.0	42.8	14.7	41.7
NICE - GA	15.2	40.1	13.9	40.1	14.7	42.7	17.6	48.9	9.0	29.7	16.4	44.4	14.9	43.1	20.3	52.1	14.1	39.7	19.6	46.7	15.6	42.8
Subject independent - leave one subject out for test																						
BraVL	2.3	8.0	1.5	6.3	1.4	5.9	1.7	6.7	1.5	5.6	1.8	7.2	2.1	8.1	2.2	7.6	1.6	6.4	2.3	8.5	1.8	7.0
NICE	7.6	22.8	5.9	20.5	6.0	22.3	6.3	20.7	4.4	18.3	5.6	22.2	5.6	19.7	6.3	22.0	5.7	17.6	8.4	28.3	**6.2**	**21.4**
NICE - SA	7.0	22.6	6.6	23.2	7.5	23.7	5.4	21.4	6.4	22.2	7.5	22.5	3.8	19.1	8.5	24.4	7.4	22.3	9.8	29.6	7.0	23.1
NICE - GA	5.9	21.4	6.4	22.7	5.5	20.1	6.1	21.0	4.7	19.5	6.2	22.5	5.9	19.1	7.3	25.3	4.8	18.3	6.2	26.3	5.9	21.6

Table 3: Classification accuracy (%) with different EEG encoder and Image encoder: top-1 (top-5)

Method	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7	Subject 8	Subject 9	Subject 10	Ave	std
EEG encoder												
ShallowNet	6.0 (16.0)	3.5 (21.5)	8.5 (28.0)	16.0 (41.0)	4.0 (16.5)	9.5 (32.0)	13.0 (31.0)	10.0 (24.5)	2.5 (9.0)	6.0 (22.5)	7.9 (24.2)	4.3 (9.3)
DeepNet	12.0 (31.0)	7.5 (27.5)	10.5 (33.0)	14.0 (36.5)	4.5 (20.0)	10.0 (33.0)	9.5 (30.0)	12.0 (41.0)	9.0 (31.5)	10.0 (33.5)	9.9 (31.7)	2.6 (5.5)
Conformer	11.0 (39.0)	8.5 (28.0)	11.0 (33.5)	14.5 (38.0)	6.0 (25.5)	10.0 (30.0)	13.0 (40.0)	13.0 (35.5)	10.5 (31.0)	13.5 (37.5)	11.1 (33.8)	2.6 (5.0)
EEGNet	11.5 (38.0)	11.0 (35.5)	14.5 (37.5)	15.0 (43.5)	11.0 (30.0)	13.0 (42.5)	14.0 (35.5)	15.0 (42.5)	11.0 (37.0)	14.5 (41.5)	13.1 (38.4)	1.8 (4.2)
TSConv	12.3 (36.6)	10.4 (33.9)	13.1 (39.0)	16.4 (47.0)	8.0 (26.9)	14.1 (40.6)	15.2 (42.1)	20.0 (49.9)	13.3 (37.1)	14.9 (41.9)	**13.8 (39.5)**	**3.3 (6.5)**
Image encoder												
ResNet	9.5 (20.0)	7.0 (11.5)	7.5 (16.0)	10.0 (18.0)	6.5 (15.0)	8.0 (14.0)	10.0 (20.0)	14.0 (25.0)	6.0 (15.0)	9.0 (18.5)	8.8 (17.3)	2.2 (3.6)
ResNet*	8.0 (13.0)	6.0 (11.0)	6.0 (11.5)	5.0 (9.5)	4.5 (10.0)	8.5 (12.0)	6.0 (10.5)	10.0 (14.5)	6.0 (12.0)	9.0 (13.5)	6.9 (11.8)	1.8 (1.6)
ViT	8.5 (18.0)	5.5 (15.0)	10.5 (20.5)	7.0 (20.5)	6.0 (15.0)	8.0 (20.5)	7.5 (19.5)	11.5 (21.5)	7.0 (17.0)	7.5 (17.5)	7.9 (18.5)	1.8 (2.2)
ViT*	13.5 (27.0)	13.0 (28.5)	13.0 (29.0)	13,5 (32.5)	9.5 (21.5)	12.5 (29.0)	10.0 (34.0)	18.0 (38.5)	6.5 (24.5)	11.0 (29.5)	12.1 (29.4)	3.1 (4.8)
CLIP	12.3 (36.6)	10.4 (33.9)	13.1 (39.0)	16.4 (47.0)	8.0 (26.9)	14.1 (40.6)	15.2 (42.1)	20.0 (49.9)	13.3 (37.1)	14.9 (41.9)	**13.8 (39.5)**	**3.3 (6.5)**

* pre-trained ResNet-50 and ViT-B/16 models.

4.3 Overall Performance

The decoding results are outlined in Table 2, with a sanity check in Appendix A.1. We evaluated the base framework (NICE) alongside its variants enhanced with self-attention (NICE-SA) and graph attention (NICE-GA). The evaluation was on the latest and most extensive image-EEG dataset, with only one state-of-the-art BraVL (Du et al., 2023) for comparison. This dataset posed a 200-way zero-shot task, comprising 200 untrained image concepts, with a chance level of 0.5%. In subject-dependent experiments, NICE achieved a top-1 accuracy of 13.8% and a top-5 accuracy of 39.5%, surpassing BraVL by 8.0% and 22.0%, respectively. The introduction of SA and GA improved the top-1 by 0.9% ($p < 0.05$) and 1.8% ($p < 0.01$), separately. Besides, in subject-independent experiments, our method achieved top-1 of 6.2% and top-5 of 21.4%. NICE-SA improved the top-1 by 0.8% ($p > 0.05$), while NICE-GA only improved for several subjects, with the average top-1 decreasing by 0.3%. We employed base NICE for fair comparison in the following experiments.

4.4 Encoder Comparison

The comparison of EEG encoders and image encoders is present in Table 3. We carefully selected several representative methods, including ShallowNet, DeepNet (Schirrmeister et al., 2017), Conformer (Song et al., 2023), and EEGNet (Lawhern et al., 2018). Our custom-designed TSConv, which utilizes temporal and spatial convolution, outperformed these methods. The average top-1 accuracy of TSConv is 5.9% higher than ShallowNet ($p < 0.001$), 3.9% higher than DeepNet ($p < 0.001$), 2.7% higher than Conformer ($p < 0.001$), and 0.7% higher than EEGNet ($p > 0.05$). The robustness of TSConv reflected by standard deviation was not superior.

We utilized image encoders, including ResNet-50 pre-trained on ImageNet-1k, ViT-B/16 pre-trained on ImageNet-21k and fine-tuned on ImageNet-1k, and CLIP-ViT-L/14 pre-trained on 400 million image-text pairs. CLIP pre-trained on huge amounts of data helped us achieve the highest results. Interestingly, ViT also worked well as the image encoder, with an average top-1 accuracy 1.7% lower than CLIP ($p > 0.05$). The results of ResNet were 6.9% lower ($p < 0.001$).

We also showed the impact of pre-training compared with non-pre-trained ResNet-50 and ViT-B/16. Pre-trained ViT outperformed the non-pre-trained by 4.2% ($p < 0.01$). Surprisingly, ResNet pre-trained with smaller ImageNet-1k decreased by 1.9% ($p < 0.05$). Note that non-pre-trained models performed well but incurred substantially higher computational costs, with details in Appendix A.2.

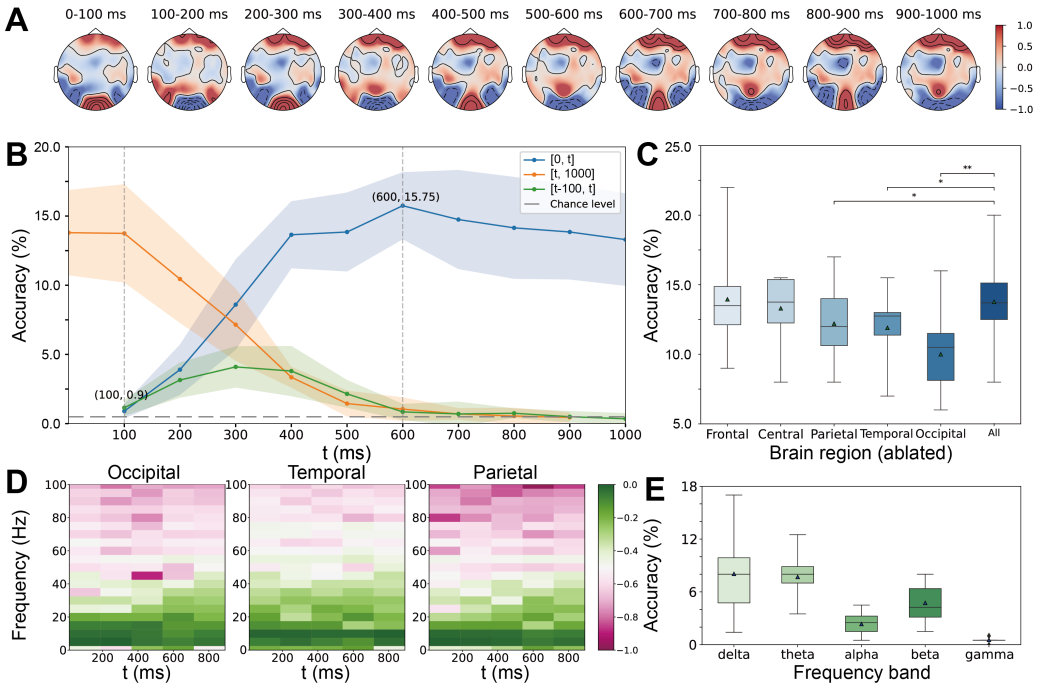

Figure 2: Temporal, spatial, spectral analysis. (A) Topographies of each 100 ms by averaging all training trials. The temporal lobe has a clear response between 100-600 ms. (B) Averaged accuracy of all subjects with different time lengths. The region of interest is 100-600 ms, after hysteresis in visual systems. (C) Ablate electrodes of different brain regions. The occipital, temporal, and parietal lobes contribute significantly to image decoding. (D) Time-frequency maps of the occipital, temporal, and parietal lobes data from one subject. The main components are below 30 Hz, and high-frequency components can be observed on the temporal lobe. (E) Averaged accuracy of different rhythms. Theta (∼4 Hz) and beta (∼14-18 Hz) bands show effective performance.

4.5 Temporal, Spatial, and Spectral Dynamics

Beyond the self-supervised framework, we try to demonstrate the biological plausibility by resolving the visual processing of EEG signals. We conducted a detailed analysis from the temporal, spatial, and spectral dynamics perspective in Fig. 2. The topographies were first plotted with each 100 ms in Fig. 2(A) by averaging all training trials from the first subject. Visual masking (Keysers & Perrett, 2002) would be alleviated because the data used were collected in many rapid series sequences. A clear response could be observed on the temporal cortex in 100-600 ms after the stimulus, although the 200 ms stimulus onset asynchrony still caused periodic responses on the occipital cortex. Besides, the parietal cortex also had a response after 100 ms. The phenomenon is consistent with the bottom-up hierarchy of visual system (DiCarlo & Cox, 2007), that the visual stimulus is processed sequentially by the V1, V2, V4 on the occipital cortex, and inferotemporal (IT) on the temporal cortex along the ventral stream for object recognition (Bao et al., 2020).

We explored the active time range in three ways: incrementing forward, decrementing backward, and segmentation, as in Fig. 2(B). It could be seen that the region of interest was in 100 to 600 ms. The average top-1 accuracy of [0, 1000] ms was 0.1% lower than that of [100, 1000] ms without significance ($p > 0.05$). These findings suggest that the initial 100 ms following stimulus onset contained limited information, possibly due to the hysteresis effect in the visual pathway (Sayal et al., 2020). Besides, the signal after 600 ms brought a negative effect, probably due to the noise from other stimuli and cognitive processing, emerging a response increasing on the frontal lobe as in Fig. 2(A).

We conducted an ablation study involving electrodes from different cortices as in Fig. 2(C). The average accuracy sharply declined 3.8% ($p < 0.01$) without occipital electrodes. Ablation of temporal,

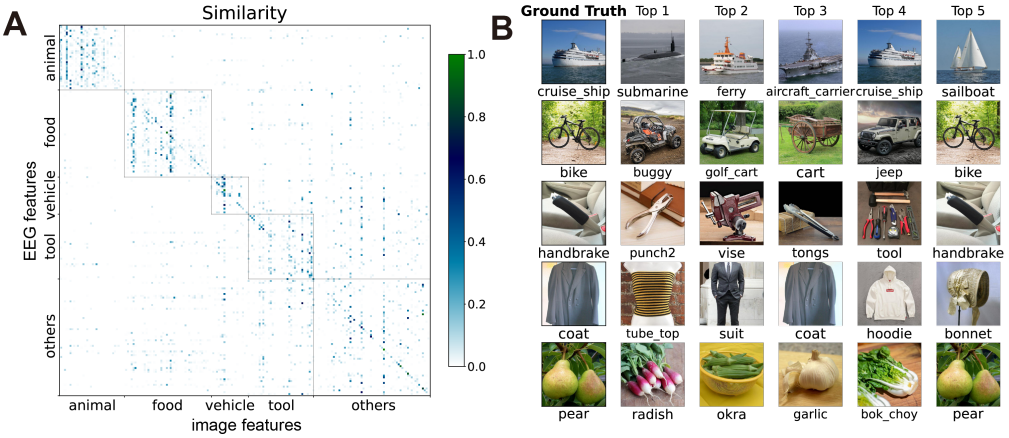

Figure 3: Semantic similarity analysis and visualization. (A) Cosine similarity of feature pairs of 200 concepts in the test set. The results calculated by the trained models of 10 subjects were averaged, and all the concepts were rearranged into five categories: animal, food, vehicle, tool, and others. (B) Classification results visualized with ground truth (first column) and the top-5 predicted.

parietal electrodes and the central area near the motor cortex decreased the results by 1.9% ($p < 0.05$), 1.6% ($p < 0.05$), and 0.5% ($p > 0.05$) separately. The accuracy improved by 0.2% ($p > 0.05$) to discard frontal electrodes. The results are still reasonable with the low spatial resolution of EEG.

We plotted time-frequency maps in Fig. 2(D) by averaging all training trials from the first subject. The main components were below 30 Hz from the occipital, temporal, and parietal region data. High-frequency responses could be observed from electrodes on the temporal cortex. Interestingly, the frequency responses had an upward trend along with visual processing.

We further conducted classification tests with different frequency bands in Fig. 2(E). Theta and delta bands near 4 Hz performed well, where theta was more stable for different subjects. The beta and alpha bands also demonstrated performance above the chance level, coarsely aligning with previous findings (Bastos et al., 2015; Michalareas et al., 2016). However, the gamma band could hardly help us in decoding, for which we have two speculations. Capturing pure gamma oscillations from EEG can be challenging due to the susceptibility of high-frequency artifacts (Fries et al., 2008; Yuval-Greenberg et al., 2008). Besides, the amplitude of the gamma band is quite low and easily modulated by other cognitive processing, such as attention and memory (Herrmann & Demiralp, 2005). This may explain why some studies with MEG have also disregarded the analysis of the gamma band (Hebart et al., 2023). Inspired by the latest work Benchetrit et al. (2023), we provided preliminary analysis on MEG data for reference in Appendix A.5.

4.6 Semantic Similarity

We show the semantic similarity and classification visualization in Fig. 3. One concern is that the discrimination we obtain from EEG is not due to semantic information but only the basic visual components, such as color, brightness, contrast, etc. We performed representational similarity analysis (RSA) (Cichy & Oliva, 2020), and got the similarity matrices by averaging all subjects, and categorizing the 200 concepts in the test set into animal, food, vehicle, tool, and others. As shown in Fig. 3(A), we could observe distinct intra-category aggregation. This meant the obtained EEG representations were closer to the image representations within the corresponding semantic category.

The images in the test set were used for visualization in Fig. 3(B). We randomly chose several top-5 decoding results from the first subject. It could be noticed that the predicted results were semantically similar to the ground truth, such as the bike was predicted into several vehicles with wheels.

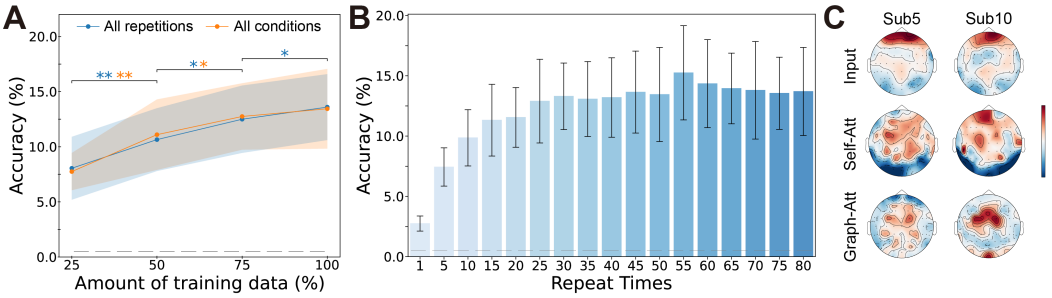

Figure 4: Effect of data size and repetition in training and test set, and visualization of SA and GA. (A) Accuracy with quarters of conditions and all repetitions, with quarters of repetitions and all conditions of training images. Adding more conditions can potentially further improve the performance. (B) Accuracy with different repetitions of the test images. Average ten times to achieve an accuracy of 9.9 (30.1)%. (C) Grad-CAMs of SA and GA show activation on temporal and occipital regions.

4.7 Data Size and Repetition

We compared the impact of the data size and repetition of the training and test sets, which may be crucial for the performance, in Fig. 4. There were 1654 concepts$\times$10 image conditions repeated four times in the training set. We averaged the four repetitions of each condition for training. Two cases were compared in Fig. 4(A), adding conditions by 25% intervals with all repetitions, and adding repetitions by 25% with all conditions. Increasing conditions, i.e., data size, contributed significantly to decoding accuracy from 25% to 50% ($p < 0.01$), from 50% to 75% ($p < 0.05$), and from 75% to 100% ($p < 0.05$). There were also significant contributions by increasing repetitions from 25% to 50% ($p < 0.01$), from 50% to 75% ($p < 0.05$), except from 75% to 100% ($p > 0.05$). The results suggest that increasing repetitions is of limited help, while more data promises further improvements.

From another view, repetitions of test data determine the practice efficiency. We compared the number of repetitions for averaging, from 5 to 80, with an interval of 5 in Fig. 4(B). Ten repetitions achieved a 9.9% (30.1%) accuracy, which tended to stabilize above 13.0% (37.0%) after twenty-five repetitions.

4.8 Effect of Spatial Modules

We tried to provide additional evidence of brain responses facilitated by spatial modules. Gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017) was used to visualize the interest region of SA and GA in Fig. 4(C) and Appendix A.3. The original data was largely affected by the response on the frontal lobe. Surprisingly, self and graph attention focuses on the temporal and occipital lobes, providing implicit evidence that object recognition from EEG is a feasible endeavor.

5 Discussion and Conclusion

There have been some studies achieving commendable results in image reconstruction with fMRI. We turn to more convenient and fast EEG signals and focus on the object recognition tasks, in which the semantic information is the significant gain by natural image decoding compared to visual decoding of contrast, color, etc. We have tried to demonstrate the feasibility and plausibility of EEG-based image decoding from three folds, zero-shot classification performance, detailed resolving of the brain activity, and model interpreting. There are also some limitations of this paper. Firstly, we observed that a stable response required multiple repetitions. This could be attributed to the brief 100 ms stimulus duration, which may make it susceptible to being missed or disrupted by stimuli before and after. We will aim to identify a more optimal window length for stimulus presentation. Secondly, we have yet to capture useful information from the gamma band, which deserves further investigation.

In conclusion, we propose a self-supervised framework to decode natural images from EEG for object recognition. Our framework has achieved remarkable results in zero-shot tasks with rich biological evidence. The results provide new inspiration for practical brain-computer interfaces.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant U2241208, and the Key Research and Development Program of Ningxia under Grant 2023BEG02063.

REFERENCES

Hamad Ahmed, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. Object classification from randomized EEG trials. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3844–3853, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.00384.

Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, January 2022. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x.

Pinglei Bao, Liang She, Mason McGill, and Doris Y. Tsao. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, July 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2350-5.

André Moraes Bastos, Julien Vezoli, Conrado Arturo Bosman, Jan-Mathijs Schoffelen, Robert Oostenveld, Jarrod Robert Dowdall, Peter De Weerd, Henry Kennedy, and Pascal Fries. Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels. *Neuron*, 85(2):390–401, January 2015. ISSN 0896-6273. doi: 10.1016/j.neuron.2014.12.018.

Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: Toward real-time reconstruction of visual perception, October 2023.

Shaked Brody, Uri Alon, and Eran Yahav. How Attentive are Graph Attention Networks?, January 2022.

Xinyu Cheng, Wei Wei, Changde Du, Shuang Qiu, Sanli Tian, Xiaojun Ma, and Huiguang He. VigilanceNet: Decouple Intra- and Inter-Modality Learning for Multimodal Vigilance Estimation in RSVP-Based BCI. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 209–217, Lisboa Portugal, October 2022. ACM. ISBN 978-1-4503-9203-7. doi: 10.1145/3503161.3548367.

Radoslaw M. Cichy and Aude Oliva. A M/EEG-fMRI Fusion Primer: Resolving Human Brain Responses in Space and Time. *Neuron*, 107(5):772–781, September 2020. ISSN 0896-6273. doi: 10.1016/j.neuron.2020.07.001.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature Neuroscience*, 17(3):455–462, March 2014. ISSN 1546-1726. doi: 10.1038/nn.3635.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *International Conference on Learning Representations*, 2016.

Joel Dapello, Kohitij Kar, Martin Schrimpf, Robert Baldwin Geary, Michael Ferguson, David Daniel Cox, and James J. DiCarlo. Aligning Model and Macaque Inferior Temporal Cortex Representations Improves Model-to-Human Behavioral Alignment and Adversarial Robustness. In *The Eleventh International Conference on Learning Representations*, February 2023.

James J. DiCarlo and David D. Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341, August 2007. ISSN 1364-6613. doi: 10.1016/j.tics.2007.06.010.

Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhao Zeng, and Cuntai Guan. LGGNet: Learning From Local-Global-Graph Representations for Brain–Computer Interface. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2023. ISSN 2162-2388. doi: 10.1109/TNNLS.2023.3236635.

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. In *International Conference on Learning Representations*, April 2021.

Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. Decoding Visual Neural Representations by Multimodal Learning of Brain-Visual-Linguistic Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–17, 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2023.3263181.

Pascal Fries, René Scheeringa, and Robert Oostenveld. Finding Gamma. *Neuron*, 58(3):303–305, May 2008. ISSN 0896-6273. doi: 10.1016/j.neuron.2008.04.020.

Kaicheng Fu, Changde Du, Shengpei Wang, and Huiguang He. Multi-View Multi-Label Fine-Grained Emotion Decoding From Human Brain Activity. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022. ISSN 2162-2388. doi: 10.1109/TNNLS.2022.3217767.

Xiaorong Gao, Yijun Wang, Xiaogang Chen, and Shangkai Gao. Interface, interaction, and intelligence in generalized brain–computer interfaces. *Trends in Cognitive Sciences*, 25(8):671–684, August 2021. ISSN 1364-6613. doi: 10.1016/j.tics.2021.04.003.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, December 2022. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2022.119754.

Matthias Guggenmos, Philipp Sterzer, and Radoslaw Martin Cichy. Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage*, 173:434–447, June 2018. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2018.02.044.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife*, 12:e82580, February 2023. ISSN 2050-084X. doi: 10.7554/eLife.82580.

C. S. Herrmann and T. Demiralp. Human EEG gamma oscillations in neuropsychiatric disorders. *Clinical Neurophysiology*, 116(12):2719–2733, December 2005. ISSN 1388-2457. doi: 10.1016/j.clinph.2005.07.007.

Jun Kai Ho, Tomoyasu Horikawa, Kei Majima, Fan Cheng, and Yukiyasu Kamitani. Inter-individual deep image reconstruction via hierarchical neural code conversion. *NeuroImage*, 271:120007, May 2023. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2023.120007.

Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(1):15037, May 2017. ISSN 2041-1723. doi: 10.1038/ncomms15037.

Xiaoxuan Jia, Ha Hong, and James J DiCarlo. Unsupervised changes in core object recognition behavior are predicted by neural plasticity in inferior temporal cortex. *eLife*, 10:e60830, June 2021. ISSN 2050-084X. doi: 10.7554/eLife.60830.

Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5):679–685, May 2005. ISSN 1546-1726. doi: 10.1038/nn1444.

Kendrick N. Kay, Thomas Naselaris, Ryan J. Prenger, and Jack L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, March 2008. ISSN 1476-4687. doi: 10.1038/nature06713.

Christian Keysers and David I Perrett. Visual masking and RSVP reveal neural competition. *Trends in Cognitive Sciences*, 6(3):120–125, March 2002. ISSN 1364-6613. doi: 10.1016/S1364-6613(00)01852-0.

Reinmar Kobler, Jun-ichiro Hirayama, Qibin Zhao, and Motoaki Kawanabe. SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG. *Advances in Neural Information Processing Systems*, 35:6219–6235, December 2022.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, July 2018. ISSN 1741-2552. doi: 10.1088/1741-2552/aace8c.

Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. The Perils and Pitfalls of Block Design for EEG Classification Experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):316–333, January 2021. ISSN 1939-3539. doi: 10.1109/TPAMI.2020.2973153.

Sikun Lin, Thomas Sprague, and Ambuj K Singh. Mind Reader: Reconstructing complex images from brain activities. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 29624–29636. Curran Associates, Inc., 2022.

Bingchuan Liu, Xiaogang Chen, Nanlin Shi, Yijun Wang, Shangkai Gao, and Xiaorong Gao. Improving the Performance of Individually Calibrated SSVEP-BCI by Task- Discriminant Component Analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1998–2007, 2021. ISSN 1558-0210. doi: 10.1109/TNSRE.2021.3114340.

Dongjun Liu, Weichen Dai, Hangkui Zhang, Xuanyu Jin, Jianting Cao, and Wanzeng Kong. Brain-Machine Coupled Learning Method for Facial Emotion Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15, 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2023.3257846.

Georgios Michalareas, Julien Vezoli, Stan van Pelt, Jan-Mathijs Schoffelen, Henry Kennedy, and Pascal Fries. Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas. *Neuron*, 89(2):384–397, January 2016. ISSN 0896-6273. doi: 10.1016/j.neuron.2015.12.018.

Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa-aki Sato, Yusuke Morito, Hiroki C. Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. Visual Image Reconstruction from Human Brain Activity using a Combination of Multiscale Local Image Decoders. *Neuron*, 60(5):915–929, December 2008. ISSN 0896-6273. doi: 10.1016/j.neuron.2008.11.004.

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, November 2021. ISSN 0162-8828. doi: 10.1109/TPAMI.2020.2995909.

Yue-Ting Pan, Jing-Lun Chou, and Chun-Shu Wei. MAtt: A Manifold Attention Network for EEG Decoding. *Advances in Neural Information Processing Systems*, 35:31116–31129, December 2022.

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*, pp. 8748–8763. PMLR, 2021.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494. Curran Associates, Inc., 2022.

Alexandre Sayal, Teresa Sousa, João V. Duarte, Gabriel N. Costa, Ricardo Martins, and Miguel Castelo-Branco. Identification of competing neural mechanisms underlying positive and negative perceptual hysteresis in the human visual system. *NeuroImage*, 221:117153, November 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.117153.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. ISSN 1097-0193. doi: 10.1002/hbm.23730.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06031-6.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, October 2017. doi: 10.1109/ICCV.2017.74.

Guohua Shen, Tomoyasu Horikawa, Kei Majima, and Yukiyasu Kamitani. Deep image reconstruction from human brain activity. *PLOS Computational Biology*, 15(1):e1006633, 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006633.

Nanlin Shi, Xiang Li, Bingchuan Liu, Chen Yang, Yijun Wang, and Xiaorong Gao. Representative-Based Cold Start for Adaptive SSVEP-BCI. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1521–1531, 2023. ISSN 1558-0210. doi: 10.1109/TNSRE.2023.3245654.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. ISSN 1558-0210. doi: 10.1109/TNSRE.2022.3230250.

C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, and M. Shah. Deep Learning Human Mind for Automated Visual Classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4503–4511, July 2017. doi: 10.1109/CVPR.2017.479.

Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. *2023 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding, January 2019.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. ISSN 1533-7928.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *International Conference on Learning Representations*, April 2018.

Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. OmniVL: One foundation model for image-language and video-language tasks. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5696–5710. Curran Associates, Inc., 2022.

Francis R. Willett, Donald T. Avansino, Leigh R. Hochberg, Jaimie M. Henderson, and Krishna V. Shenoy. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858): 249–254, May 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03506-2.

Rui Xu, Wenjing Zhou, Xin Xu, Zhipei Ling, Hesheng Liu, and Bo Hong. A temporal hierarchy of object processing in human visual cortex, September 2023.

Shlomit Yuval-Greenberg, Orr Tomer, Alon S. Keren, Israel Nelken, and Leon Y. Deouell. Transient Induced Gamma-Band Response in EEG as a Manifestation of Miniature Saccades. *Neuron*, 58(3): 429–441, May 2008. ISSN 0896-6273. doi: 10.1016/j.neuron.2008.03.027.