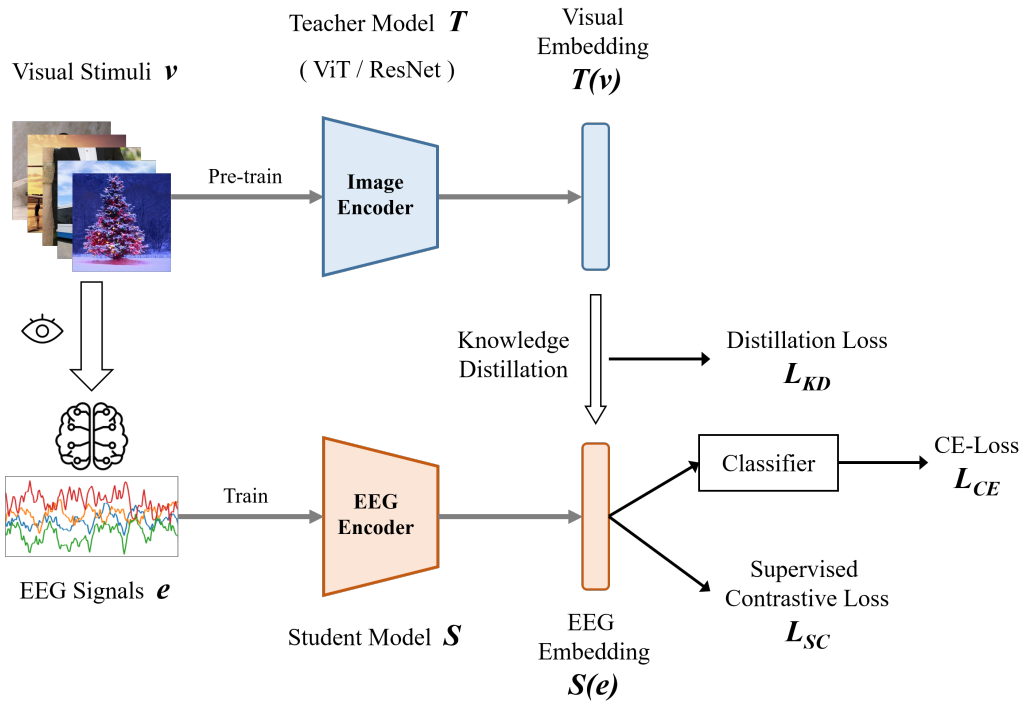
Graphical Abstract

Contrastive Learning with Cross-Modal Information for EEG Visual Classification

Shuning Xue, Jie Jiang, Longteng Guo, Jing Liu



Highlights

Contrastive Learning with Cross-Modal Information for EEG Visual Classification

Shuning Xue, Jie Jiang, Longteng Guo, Jing Liu

- Feature fusion strategy for EEG signals and visual images via contrastive learning.
- Positive samples consist of signals of the same category along with their corresponding cross-modal images.
- First supervised contrastive learning method for EEG visual classification.
- Consistent performance improvements across five EEG encoders.
- Superior performance across three datasets encompassing four classification tasks.

Contrastive Learning with Cross-Modal Information for EEG Visual Classification

Shuning Xue[a,b], Jie Jiang[b], Longteng Guo[b], Jing Liu[a,b,*]

[a]*School of Artificial Intelligence, University of Chinese Academy of Sciences, 100049, Beijing, China*
[b]*Zidongtaichu Foundation Model Research Center, Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China*

Abstract

Visual decoding using EEG signals has attracted considerable attention in both neuroscience and machine learning. Recently, several studies have employed paired training data (stimulus-response) and pre-trained image encoders to facilitate image retrieval and generation based on EEG signals. However, existing algorithms primarily concentrate on aligning EEG features with visual features, often neglecting the significance of semantic constraints on EEG features and classification tasks. To overcome these challenges, this study proposes a novel EEG contrastive learning algorithm that integrates features from pre-trained image encoders with labels from true semantic categories. On one hand, we utilize knowledge distillation losses to guide EEG encoders in learning enhanced EEG features from cross-modal information. On the other hand, we introduce a supervised contrastive learning framework that treats samples from the same category, along with their corresponding cross-modal images, as positive samples. This approach strengthens the semantic discriminative capabilities of EEG encoders. Extensive comparative and generalization experiments reveal that by simultaneously leveraging visual and semantic information in contrastive learning, we achieve improved performance across three EEG datasets and five EEG models. Additionally, we evaluate the impact of eight pre-trained models, three knowledge distil-

*Corresponding author
Email addresses: xueshuning2021@ia.ac.cn (Shuning Xue), jie.jiang@nlpr.ia.ac.cn (Jie Jiang), longteng.guo@nlpr.ia.ac.cn (Longteng Guo), jliu@nlpr.ia.ac.cn (Jing Liu)

lation loss functions, and two types of contrastive loss functions. Our study also investigates the performance of the algorithm with varying data sizes and different training parameters. Overall, this research presents an effective contrastive learning algorithm that leverages cross-modal information for neural decoding and brain-computer interface studies. The source code will be released at https://github.com/xuesn/EEGFusion.

Keywords: Brain-computer interface (BCI), EEG-based visual recognition, contrastive learning, cross-modal learning, supervised contrastive learning

1. Introduction

Analyzing visual responses is essential for understanding the mechanisms of the visual cortex[1, 2]. By decoding visual information from physiological signals of brain activity, we can gain insights into how the brain functions[3]. Besides, this knowledge is instrumental in developing brain-computer interfaces, which can enhance the quality of life for individuals with disabilities[4, 5, 6] and have applications in communication[7], industry[8], and education[9]. Humans can perceive images within just a few hundred milliseconds[10]. To study this rapid perceptual process, researchers are exploring electroencephalography (EEG), a non-invasive technique that provides high temporal resolution of brain signals[11, 12].

Utilizing public EEG datasets, researchers have applied deep-learning algorithms for various tasks, including classification, retrieval, and generation of visual stimuli[13, 14, 15, 16]. However, achieving accurate semantic classification of EEG signals remains a challenge[17]. Recent studies have introduced knowledge distillation algorithms that utilize pre-trained models from computer vision to enhance the learning process of EEG encoders[18, 19, 20]. Knowledge distillation (KD)[21, 22] is an algorithm in which a teacher model guides a student model during the learning process. The teacher model is typically robust and pre-trained on large datasets, while the student model is more lightweight. Due to constraints such as limited sample size or fewer model parameters, the student model has difficulty achieving high classification accuracy. To enhance the student model's learning effectiveness, the teacher model can assist in feature extraction rather than relying solely on label information and cross-entropy classification loss. For visually evoked EEG signals, cross-modal knowledge distillation is a promising method for training EEG encoders. By aligning the features extracted from the EEG

model with the visual features from the pre-trained image encoder, the EEG model can effectively learn the semantic knowledge of visual stimuli during training. This approach is advantageous because the extracted features contain more information than the labels alone.

Although the recent knowledge distillation approaches have shown promising outcomes in the visual decoding of EEG signals, there are still challenges for EEG visual classification. On the one hand, current EEG visual decoding algorithms tend to focus on achieving better alignment with image features while overlooking the inherent semantic feature learning of the EEG signals themselves. Some studies have concentrated on generating images based on EEG signals; in this context, prioritizing image feature alignment can enhance the quality of the generated images[18, 20]. Other methods primarily leverage EEG signals to align image features, ensuring that retrieved images accurately represent real-world images, or they indirectly support the classification task through the images' categories[15, 19, 23]. However, this emphasis on image features neglects the EEG visual classification task based on the characteristics of the EEG signals, and these methods are unsuitable for scenarios where the corresponding image is missing or difficult to obtain. On the other hand, relying solely on self-supervised cross-modal knowledge distillation overlooks the semantic information provided by true labels. A supervised learning framework is crucial for EEG classifiers to develop semantic discriminative abilities. To improve performance in EEG visual classification, it is essential to fully leverage semantic information in a supervised manner. Additionally, beyond merely predicting labels using cross-entropy loss, enforcing constraints on similar samples and related visual samples at the feature level of the EEG encoder, along with managing the distance between features and heterogeneous samples, will assist the classifier in distinguishing features from different semantic categories.

In this study, we propose a feature fusion strategy that combines EEG signals with visual images using contrastive learning to improve performance in EEG visual classification tasks. The foundation of our method is the simultaneous use of visual and semantic information through contrastive learning. Specifically, we utilize a supervised contrastive learning framework that treats samples from the same category, along with their corresponding cross-modal images, as positive samples. On the one hand, we apply knowledge distillation loss to integrate visual information into the EEG features. On the other hand, we introduce supervised contrastive learning to enhance the models' discriminative ability, while fusing features obtained from knowledge

distillation to address gradient collapse issues that commonly arise in contrastive learning. Through this approach, we achieve effective feature fusion from both visual and semantic information, thereby boosting the classification capability of EEG models.

In summary, the main contributions of this study are as follows:

- We proposed a contrastive learning framework with cross-modal information that effectively improves classification accuracy in EEG visual recognition tasks.
- We introduced the supervised contrastive learning method, which further enhanced the accuracy of EEG visual classification based on features learned through the cross-modal knowledge distillation algorithm.
- We conducted comprehensive experiments to assess the performance of our method using three EEG datasets and five EEG models.
- Our investigation focused on the influence of eight pre-trained image encoders, three knowledge distillation loss functions, and various contrastive settings.

The remainder of this paper is organized as follows. Section 2 briefly describes the related works of EEG-based visual classification and cross-modal knowledge distillation algorithms. In Section 3, we introduce our proposed contrastive learning framework in detail, including the EEG models, loss functions, and training strategy. Section 4 illustrates the impressive performance of our method in comparison with baseline models, followed by some discussion experiments in Section 5.

2. Related Work

In this section, we first review current research on EEG-based visual classification. Next, we briefly introduce recent work on cross-modal knowledge distillation for EEG visual decoding.

2.1. EEG-based visual classification

Researchers have explored various paradigms to enhance EEG visual decoding and improve the efficiency of information transmission. Two commonly used paradigms in EEG visual decoding are steady-state visual evoked

potential (SSVEP)[24], which relies on typical brain responses to periodic visual stimuli, and oddball paradigms[25], which detect event-related potentials (ERPs) in response to infrequent target stimuli. These paradigms typically use a limited set of visual stimuli that are controlled during experiments.

Recently, there has been increasing interest in utilizing natural images with randomized stimulus presentations to stimulate EEG signals, which is inherently more complex than using distinct visual stimuli. Kaneshiro et al.[26] were among the first to employ Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to classify EEG signals into semantic categories (6 classes) and individual exemplars (72 classes). Their research highlights the potential of machine learning approaches for EEG visual classification. Spampinato et al.[27] used Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to classify block-design EEG signals. Li et al.[28] and Ahmed et al.[17] evaluated various machine learning models in random-trial designs, drawing attention to the challenges in classifying raw EEG signals. Kalafatovich et al.[29] developed a CNN model specifically tailored for temporal EEG signals, which showed strong performance on the EEG72 dataset[26], a benchmark widely used for assessing deep learning models. Bagchi et al.[30] investigated CNN models with an increased number of convolutional kernels, while Kalafatovich et al.[31] integrated temporal and spatial features through a two-stream convolutional network (TSCNN). Additionally, the transformer architecture, enhanced with attention modules, achieved high accuracy on the EEG72 dataset. Recently, Luo et al.[8] combined time-domain, frequency-domain, and spatial features using a graph convolution and transformer architecture, attaining optimal results on both the EEG72 dataset and another dataset referred to as EEG200. In our previous work[32], we proposed a hybrid model that integrates a temporal CNN for local feature extraction and a transformer model for global feature extraction, achieving optimal performance across five datasets.

For the EEG visual classification task, previous research mainly focuses on developing deep-learning models that adapt to EEG signals while incorporating information from temporal, spatial, and frequency domains. Lately, researchers have begun implementing deep learning algorithms for EEG visual decoding, mainly concentrating on natural image generation and cross-modal knowledge distillation algorithms. Below, we introduce the latest cross-modal knowledge distillation algorithms for EEG visual decoding.

2.2. Cross-modal knowledge distillation for EEG visual decoding

Recent studies have highlighted the effectiveness of cross-modal knowledge distillation, particularly in retrieving pairs of EEG images through feature alignment and generating images based on EEG embeddings. Jiao et al. [33] were the first to propose a visual-guided EEG decoding method using convolutional neural networks (CNNs) with mean squared error (MSE) loss. Their study directly employed pre-trained visual embeddings to supervise the extraction of EEG features from the EEG2000[27] and the EEG72[26] datasets.

Recently, Ye et al. [15] introduced a self-supervised cross-modal retrieval paradigm aimed at reconstructing the exact visual stimulus without relying on image class information from the large visual-evoked EEG dataset, ImageNet-EEG[17]. Their proposed CNN model was designed to maximise the mutual information between the EEG encoding and the associated visual stimulus by utilising InfoNCE loss. Similarly, Song et al. [18] developed a self-supervised framework for learning image representations from EEG signals for object recognition. Their experiments were conducted using the THINGS-EEG-5Hz[34] dataset, where they trained the proposed CNN model by matching image features with corresponding EEG features through contrastive learning. Ferrante et al. [19] also employed a knowledge distillation approach to align EEG models with a pre-trained image classification teacher network using the THINGS-EEG-5Hz dataset. This study differed from previous works in that EEG signals were converted to time-frequency decomposition (TFD), and the class probability distributions of the student model were aligned with those of the teacher, rather than their embeddings. Li et al. [20] adopted a similar self-supervised contrastive learning framework to achieve classification, retrieval, and generation tasks on the training set of the THINGS-EEG-5Hz dataset and the THINGS-MEG[35] datasets. Their study introduced a plug-and-play model called ATM, which is based on a Transformer Encoder and spatiotemporal convolution architecture that surpasses previous models.

In the study mentioned above, EEG signal classification was performed using similarity-calculation methods. This approach classifies EEG signals indirectly by comparing the feature similarity between EEG data and corresponding visual images, utilizing the categories of those images for classification. While this method effectively leverages features from pre-trained visual models, it does have some limitations. It relies on candidate images, making it unsuitable for situations where the corresponding visual stimuli

are not available. Additionally, it does not consider the category labels of the EEG signals themselves, meaning it lacks direct semantic constraints on EEG features.

To address these limitations, we adopt a supervised learning framework and introduce a supervised contrastive learning strategy to enhance the model's ability to discriminate signals semantically. Specifically, we first apply knowledge distillation loss to integrate visual information into EEG features. Then, we implement a supervised contrastive learning strategy that incorporates semantic information, ensuring feature consistency within the same semantic category while increasing differentiation between features from different categories. Ultimately, our classification model combines both visual and semantic information to effectively perform EEG visual classification tasks within a supervised learning framework, resulting in improved performance in semantic classification.

3. Methodology

We comprehensively describe the contrastive learning with cross-modal information framework in this section. Firstly, we introduced the EEG models used in our experiments. Then, the knowledge distillation loss and supervised contrastive loss are discribed in detail. Finally, the training strategy of our contrastive learning and the architecture of EEG visual classification is presented.

3.1. EEG models

To assess the effect of our contrastive learning framework, we evaluated classification accuracies on five models: EEGNet[36], ShallowConvNet[37], Temporal CNN[31], B.D.[38], and a hybrid model[32]. Brief descriptions of the five models are given as follows:

3.1.1. EEGNet[36]:

The EEGNet features a compact CNN architecture comprising a common convolution layer, a depthwise convolution layer, and a separable convolution layer. The EEGNet has performed well across various EEG paradigms and benchmarks[39, 40, 41]. The implementation codes for the EEGNet were sourced from LMDA[42].

3.1.2. ShallowConvNet[37]:

The ShallowConvNet consists of two CNN layers followed by a fully connected layer with a softmax activation function. The ShallowConvNet has performed well across several EEG benchmarks of Motor Imagery paradigm[43, 44]. The implementation codes for the ShallowConvNet were also sourced from LMDA[42].

3.1.3. Temporal CNN:

The Temporal CNN model is reproduced from the temporal stream structure of TSCNN[31]. It includes three temporal 1D convolution layers and one spatial 1D convolution layer. Besides, we add an Electrode Reweight module. The Temporal CNN has been apllied on EEG visual classification and performed well on the EEG72 dataset[26].

3.1.4. B.D. model[38]:

The B.D. model is based on the dilated residual ConvNet architecture from[45], and it consists of a subject block along with five CNN blocks made up of three convolution layers each. The model has demonstrated effective performance in speech decoding tasks using magnetoencephalography (MEG) and EEG, and it has also been utilized in cross-model knowledge distillation algorithms for decoding speech perception. The implementation codes for the B.D. model were sourced from ATM[42].

3.1.5. The hybrid model:

The hybrid model is reproduced from the structure of a EEG visual classification model[32]. We reproduced four modules of this model. The Electrode Reweight module adaptively reweights the input electrodes to pay more attention to vision-related electrodes. The Temporal CNN module extracts local temporal features from the EEG signals of each electrode. The Spatial CNN module then integrates these temporal features across all electrodes. Finally, the Transformer module extracts global EEG embeddings from the extracted features. The implementation codes for the hybrid model will be released at https://github.com/xuesn/EEGFusion. We implemented the models mentioned above, using the publicly available PyTorch 1.6.0 framework.

3.2. Loss functions

As presented in Fig. 1, the total loss function for our framework is a combination of three loss functions, represented as $loss = L_{CE} + \alpha_1 \cdot L_{KD} + \alpha_2 \cdot$

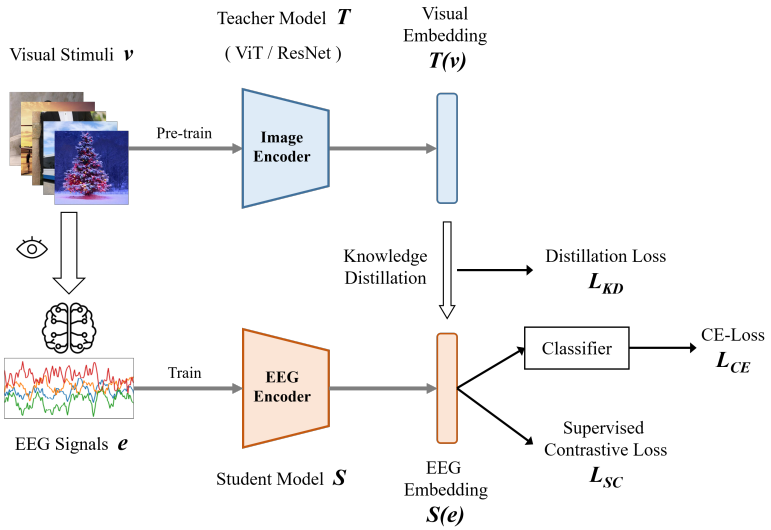


Figure 1: The overview of our contrastive learning framework for EEG visual classification. A pre-trained model based on image data serves as a 'teacher' model. A 'student' model is trained using EEG data and knowledge extracted from the 'teacher' model. Supervised contrastive loss and cross-entropy classification loss are simultaneously optimized with true labels.

L_{SC}. In this equation, L_{CE} represents the cross-entropy loss, L_{KD} represents the knowledge distillation loss, and L_{SC} represents the supervised contrastive loss. The α_1 and α_2 are hyperparameters that balance the contribution of each loss function. In our implementation, we assign different values to α_1 and α_2 as adjustable hyperparameters for three EEG datasets. We have compared different loss functions in knowledge distillation and supervised contrastive learning to evaluate the impact of the loss function in Section 5.2 and 5.3.

3.2.1. Knowledge distillation loss function

For L_{KD}, we compared InfoNCE loss[46], Huber loss[47], and Kullback-Leibler divergence loss. InfoNCE loss is commonly used for contrastive learning[15, 18, 34]. Its training goal is to increase the similarity between EEG embeddings and their corresponding visual embeddings while simultaneously decreasing the similarity with non-paired visual embeddings. Huber loss is the modified loss from mean square error (MSE) loss, which is used for knowledge distillation in [33]. The loss is commonly used in regression

tasks to measure the Euclidean distance between EEG embeddings and their paired visual embeddings. It is effective because of its robustness to outliers, which helps to smooth the training process. The Kullback-Leibler divergence loss is frequently used to ensure that the approximate posterior distribution learned by the encoder closely matches the true prior distribution. Moreover, in the context of knowledge distillation, this loss function treats EEG embeddings and visual embeddings as two distinct probability distributions and can be applied to align the feature distribution of the student model with that of the teacher model. These three loss functions provide different perspectives for making EEG embeddings align with visual embeddings, leading to varying training outcomes. The relevant experimental results can be found in Section 5.2.

InfoNCE loss[46]:. InfoNCE loss is commonly used in contrastive learning to increase similarities between matched pairs and decrease those between unmatched pairs.

$$L_{InfoNCE} = -\mathbb{E}[log\frac{exp(Sim_{S,T}/\tau)}{\sum_{k=1}^{N} exp(Sim_{S,T_k}/\tau)}] \tag{1}$$

where the $Sim_{S,T}$ denotes the similarity score between student embedding S and teacher embedding T pairing data, the τ is the temperature parameter that determines the emphasis on hard samples.

Huber loss[47]:. Huber loss is the combination of mean square error (MSE) loss and mean absolute error (MAE) loss.

$$L_{Huber}(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{for } |y - \hat{y}| \leq \delta \quad (2a) \\ \delta(|y - \hat{y}| - \frac{1}{2}\delta), & \text{for } |y - \hat{y}| > \delta \quad (2b) \end{cases}$$

where the y denotes the teacher embedding, the $\hat{y}$ denotes the student embedding, the δ determines the emphasis on MSE loss and MAE loss.

Kullback-Leibler divergence loss:. Kullback-Leibler divergence loss quantifies the difference between the probability distributions of two features using the Kullback-Leibler divergence. The loss guides the student model to learn the feature distribution of the teacher model.

$$L_{KL}(y, \hat{y}) = y \cdot log\frac{y}{\hat{y}} \tag{3}$$

where the y denotes the teacher embedding, the $\hat{y}$ denotes the student embedding.

3.2.2. Supervised contrastive loss functions

For L_{SC}, we compared SupCon loss[48] and triplet loss[49], both of which are commonly used in contrastive learning. Generally speaking, SupCon loss, similar to InfoNCE loss, utilizes cross-entropy to gauge the similarity between positive and negative samples, with a temperature coefficient that regulates the emphasis placed on negative samples. In contrast, triplet loss measures the similarity of features between positive and negative samples using Euclidean distance and explicitly requires a minimum margin between the distances of positive and negative samples. The experiments comparing these two supervised contrastive loss functions and their corresponding unsupervised contrastive learning losses are presented in Section 5.3.

SupCon loss[48]:. SupCon loss is generalized from InfoNCE loss to incorporate supervision.

$$L_{SupCon} = \sum_{e \in E} \frac{-1}{|P(e)|} \sum_{p \in P(e)} log \frac{exp(z_e \cdot z_p / \tau)}{\sum_{a \in A(e)} exp(z_e \cdot z_a / \tau)} \tag{4}$$

where the z denotes the embedding, the E denotes the set of indices of embeddings, the $A(e)$ denotes the set of indices of all embeddings in the batch distinct from e, the $P(e)$ denotes the set of indices of all positives in the batch distinct from e, and $|P(e)|$ is its cardinality.

Triplet loss[49]:. Triplet loss is implemented using sets of three (anchor, positive, negative) to ensure that the distance between the anchor and the positive is smaller than the distance between the anchor and the negative by at least one margin.

$$L_{triplet}(a, p, n) = max(|z_a - z_p| - |z_a - z_n| + m, 0) \tag{5}$$

where the a is the anchor, the p is the positive, the n is the negative, the z_a denotes the embedding of the anchor, the z_p denotes the embedding of the positive, the z_n denotes the embedding of the negative, the m is the parameter that determines the emphasis on distance between the negative and the anchor relative to the positive.

3.3. Overall framework

This section provides a detailed description of the proposed contrastive framework, introduces feature-based cross-modal knowledge distillation, describes supervised contrastive learning, and explains the classification architecture for EEG signals. The overall framework is depicted in Fig. 1.

3.3.1. Feature-based cross-modal knowledge distillation

During training, pairs of visual stimuli and EEG signals are inputted into the framework. The image encoder and the EEG encoder extract embeddings from their respective modalities. The image encoder acts as the teacher model, while the EEG encoder is the student model. The distillation loss measures the difference in output between the student and teacher models, allowing the EEG encoder to leverage the knowledge of the image encoder to learn the neural patterns associated with visual recognition. Consider a stimulus image v. Let $T(v)$ represent the output visual embedding of the teacher model, and let $S(e;\theta)$ represent the output EEG embedding of the student model, where θ is its parameter and e represents the EEG signal. The student model was trained by minimizing the loss function $L(\theta)$, which combines the cross-entropy loss L_{CE} with the knowledge distillation loss L_{KD} and the supervised contrastive loss L_{SC}. This feature-based cross-modal knowledge distillation approach enables the model to learn the shared information between the two modalities, rather than using supervised learning with predefined labels alone.

There are several pre-trained image encoders available, including including: AlexNet[50], ResNet[51], and Vision Transformer (ViT)[52], and CLIP[53] models. In this study, we compared eight image encoders to evaluate how well their features aligned with EEG embeddings, .

Since the output features from different encoders vary in dimensions, we used a linear projection layer to map the output features of the EEG encoder to a common dimensional space before aligning the two sets of features. For the teacher feature of the image encoder, we used the output of last layer before the classification network in the pre-trained model.

3.3.2. Supervised contrastive learning

The supervised contrastive learning method was proposed in[48], and its process is illustrated in Fig. 2. The supervised contrastive loss involves contrasting all samples from the same class as positives against the negatives

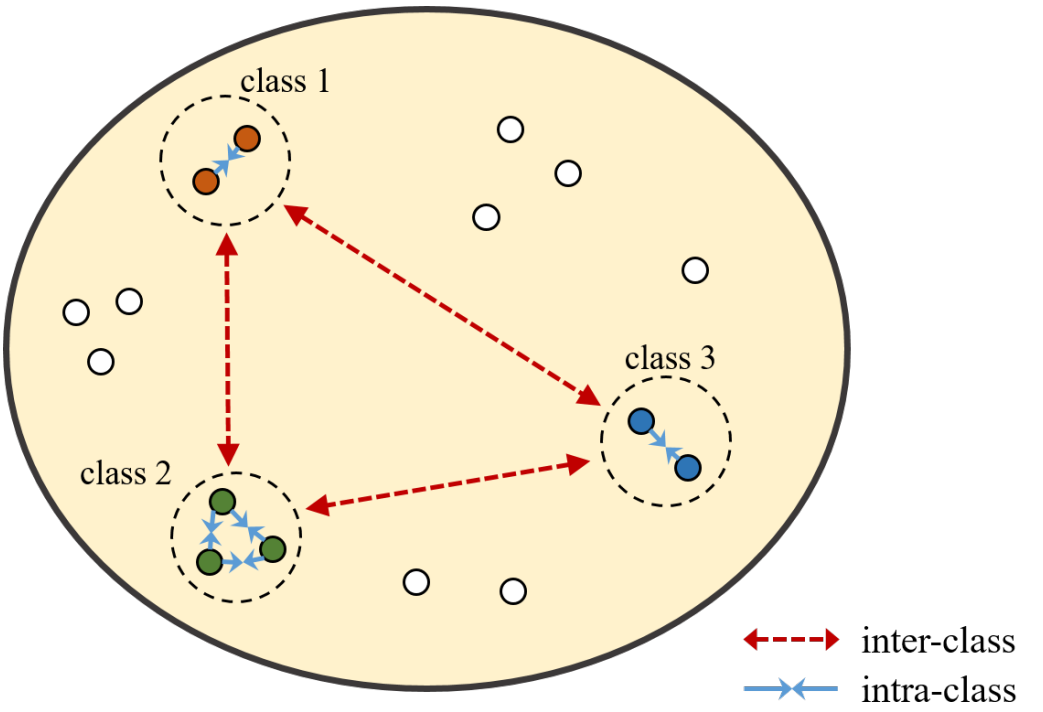


Figure 2: Illustration of the concept of supervised contrastive learning. The supervised contrastive loss considers the set of all samples from the same class as positives and contrasts against the negatives from the remainder of the batch.

from the rest of the batch. As shown by the blue and red arrows, the embeddings from the same class are brought closer together, while the embeddings from different classes are pushed apart. By utilizing label information, supervised contrastive learning ensures that the embeddings of the same class are closer together compared to self-supervised contrastive learning.

For the supervised contrastive learning approach, we first compute the embeddings for all samples in a batch using the EEG encoder. These embeddings are then normalized by removing the mean and dividing by the standard deviation. We also normalize the modulus of the features to prevent extremely large values when calculating the similarity matrix, as this could adversely affect the performance of contrastive learning. Furthermore, when calculating the supervised contrastive loss, we utilize the features output by the EEG encoder before the dimension mapping, rather than the features input into the classification network.

During the training process of the model employing supervised contrastive learning, we simultaneously optimize three types of loss: the cross-entropy loss L_{CE}, the knowledge distillation loss L_{KD}, and the supervised contrastive loss L_{SC}. Different weights for these losses control the focus of model parameter updates. The parameters of the EEG encoder are updated through gradient backpropagation from all three loss types. The parameters of the linear projection layer is updated using the gradients from L_{CE} and L_{KD}. The parameters of the classification network is only updated through the gradient backpropagation of L_{CE}. In this study, we have observed that incorporating supervised contrastive learning can enhance the effectiveness of knowledge distillation rather than solely relying on cross-entropy loss to leverage label information.

3.3.3. Classification of EEG embedding

As shown in Fig. 1, the EEG embeddings are sent to classifier for computing the cross-entropy loss L_{CE}. We use the fully connected neural network as our classifier, adjusting the number of output units to match the number of categories in each dataset. We compute the cross-entropy loss L_{CE} based on the true labels and the probabilities assigned to all categories by the classifier. Our training architecture follows an end-to-end framework, allowing us to train the classification tasks simultaneously with feature-based cross-modal knowledge distillation and supervised contrastive learning. Throughout the training process, the EEG embeddings transfer knowledge from the visual embeddings and are guided by the supervised contrastive loss, enabling the classifier to accurately predict the correct class for each sample.

4. Experiments and Results

To thoroughly evaluate the proposed algorithm, we conducted extensive experiments on three EEG datasets. We implemented our method using five EEG models described in the previous section and applied 10-fold cross-validation to assess the classification performance of both the baseline and our algorithm. The overall results demonstrate the effectiveness of our algorithm. Additionally, we carried out ablation studies using the hybrid model to assess the contributions of different information, including the visual information obtained from knowledge distillation and the semantic information derived from supervised contrastive learning. Further experiments involving different pre-trained models, loss functions, and data sizes are presented in Section 3.2.

Table 1: Description of the Three Public EEG Datasets.

Dataset	Exp. paradigm	# of channels	# of class	# of subjects	# of trials per class	Epoch length (ms)
EEG72[26]	low-speed	124	6 or 72	10	864 or 72	500
ImageNet-EEG[17]	low-speed	96	40	1	1,000	2,100
THINGS-EEG-5Hz[34]	RSVP	64	1,654	10	40	500

*Exp.: Experimental, low-speed: low-speed serial visual presentation, RSVP: rapid serial visual presentation

4.1. Datasets

Table 1 provides the details of three public EEG datasets for visual recognition.

4.1.1. EEG72[26]

The first dataset includes EEG data collected from 10 subjects. During the experiment, participants were shown color images from six different semantic categories. A total of 72 images were presented against a mid-grey background, with each image displayed for 500 milliseconds. The EEG data was recorded using a 128-channel EGI HCGSN 110 net.

4.1.2. ImageNet-EEG[17]

The second dataset consists of EEG data collected from a single participant. This study involved presenting natural images from 40 different classes of the ImageNet database[54]. In total, 40,000 images were shown, with each image displayed for a duration of 2 seconds. The EEG data was recorded using a BioSemi ActiveTwo system, which includes 96 channels.

4.1.3. THINGS-EEG-5Hz[34]

The third dataset consists of EEG data collected from ten participants. The data was gathered using a Rapid Serial Visual Presentation (RSVP) paradigm, which involved displaying natural images cropped into a square shape. These images were selected from 1,854 object concepts in the THINGS database[55]. In this dataset, each image was presented for 100 milliseconds, followed by a 100-millisecond blank screen. The EEG data was collected using a BrainVision actiCHamp amplifier and 64-channel EASYCAP. The dataset is divided into two parts: 1,654 concepts $\times$ 10 images $\times$ 4 repetitions and 200 concepts $\times$ 1 images $\times$ 80 repetitions. We only used the 1,654 concepts to ensure that the training and test sets have the same distribution.

4.2. Data preprocessing

The EEG signals from EEG72 has already been preprocessed when it was released. The signals were filtered within the frequency range of 1 to 25 Hz. After filtering, the signals were downsampled to a rate of 62.5 Hz and segmented into trials consisting of 32 time samples each.

For other datasets, we utilized the MNE package[56] for preprocessing. The sampling rates were adjusted to 250 Hz. We filtered the signals within a frequency range of 0.1 to 100 Hz and applied a notch filter to eliminate 50 Hz power line interference. The epoching schemes are presented below:

- For the ImageNet-EEG dataset, epochs were created ranging from -100 to 2,000 ms relative to stimulus onset. Baseline correction was performed by subtracting the mean of the pre-stimulus interval for each trial.

- For THINGS-EEG-5Hz datasets, epochs were created ranging from 0 to 500 ms relative to stimulus onset. As there were 1,654 concepts $\times$ 10 image conditions repeated four times, we averaged the four repetitions of each condition to reduce noise.

Besides, we applied z-score normalization and clamped large voltages. No further preprocessing or artifact correction methods were applied.

4.3. Evaluation metrics

In our experiments, we implemented 10-fold cross-validation to evaluate the methods. The subject-wise accuracies obtained from the test folds were averaged and presented in following sections. When splitting the training set and the test set, we performed the 10-fold cross-validation on the samples from each class before merging the training and test sets from different classes. This approach ensured that each set contained the same proportion of samples from each class, helping to prevent bias towards any particular class during model training. It's important to note that we used a portion of the THINGS-EEG-5Hz dataset, which consists of 1,654 classes, with 10 images per class. As a result, for each test fold of the THINGS-EEG-5Hz dataset, we selected only one image from each class. For the optimization algorithm, we employed Adam[57].

Table 2: Experiments of Five EEG Models for Three EEG Dataset.

Method	EEG72		ImageNet-EEG	THINGS-EEG-5Hz
	6-class	72-class	40-class	1,654-class
EEGNet[36]	46.15 ± 0.38	17.79 ± 0.39	7.51 ± 0.31	0.85 ± 0.04
EEGNet + CL	50.61 ± 0.47*	21.36 ± 0.27*	6.90 ± 0.19	1.48 ± 0.04*
ShallowConvNet[37]	40.95 ± 0.54	13.51 ± 0.38	4.22 ± 0.18	1.20 ± 0.06
ShallowConvNet + CL	42.84 ± 0.59*	14.06 ± 0.32*	4.02 ± 0.27	0.73 ± 0.04
Temporal CNN[31]	50.42 ± 0.57	23.51 ± 0.32	10.03 ± 0.32	2.13 ± 0.08
Temporal CNN + CL	52.84 ± 0.58*	22.84 ± 0.31	11.29 ± 0.39*	2.90 ± 0.09*
B.D.[38]	43.51 ± 0.48	11.85 ± 0.40	3.00 ± 1.42	0.08 ± 0.03
B.D. + CL	47.19 ± 0.41*	12.64 ± 0.29*	11.05 ± 0.24*	1.75 ± 0.04*
Hybrid[32]	55.13 ± 1.49	26.58 ± 1.23	10.89 ± 0.77	1.29 ± 0.31
Hybrid + CL	58.59 ± 1.39*	28.64 ± 1.44*	15.91 ± 0.28*	3.23 ± 0.25*

where $\star$ denotes the classification accuracy of the model with our contrastive learning algorithm is significantly better than the baseline (paired t-test $p<5e^{-4}$).

4.4. Overall performance

In this section, we evaluated the classification accuracies on three EEG datasets for five models. Table 2 demonstrates that most of the models utilizing our contrastive learning algorithm could achieve higher decoding accuracy.

The hybrid model achieves the following accuracies: 55.13% in the 6-class task, 26.58% in the 72-class task for EEG72, 10.89% in the 40-class task for ImageNet-EEG, and 1.29% in the 1,654-class task for THINGS-EEG-5Hz. With the implementation of our contrastive learning algorithm, classification accuracy improves by 3.47% in the 6-class task ($p = 2.574e^{-25}$), 2.06% in the 72-class task ($p = 6.800e^{-14}$), 5.02% in the 40-class task ($p = 5.713e^{-9}$), and 1.94% in the 1,654-class task ($p = 3.103e^{-54}$). The experimental results indicate that our contrastive learning algorithm significantly benefits the hybrid model in the more challenging tasks associated with ImageNet-EEG and THINGS-EEG-5Hz compared to the EEG72 dataset. With our contrastive learning algorithm method, the performance of the hybrid model improved by approximately 50% on the ImageNet-EEG dataset and around 150% on the THINGS-EEG-5Hz dataset. In contrast, while the classification accuracy for the EEG72 dataset is relatively higher, there is less potential for further improvement.

Other models demonstrate somewhat different performance outcomes. For the EEG72 dataset, most models benefit from our contrastive learning algorithm, particularly EEGNet, which achieves an accuracy increase of

4.47% in the 6-class task ($p = 3.698e^{-30}$) and 3.57% in the 72-class task ($p = 2.789e^{-29}$). However, the impact of our contrastive learning algorithm on the Temporal CNN is relatively limited; while its classification accuracy improves by 2.41% in the 6-class task ($p = 2.468e^{-26}$), it declines by 0.67% in the 72-class task ($p = 7.305e^{-5}$). In the case of the ImageNet-EEG and THINGS-EEG-5Hz datasets, both EEGNet and ShallowConvNet show relatively weak performance, and the application of our contrastive learning algorithm even leads to a decrease in performance for both models. Specifically, EEGNet's classification accuracy increases by 0.63% for the THINGS-EEG-5Hz dataset ($p = 5.968e^{-34}$) but decreases by 0.61% for the ImageNet-EEG dataset ($p = 6.673e^{-4}$). Similarly, ShallowConvNet's classification accuracy drops by 0.20% for the ImageNet-EEG dataset ($p = 0.0801$) and by 0.47% for the THINGS-EEG-5Hz dataset ($p = 2.401e^{-30}$). We hypothesize that the negative effect of our contrastive learning algorithm on these models may stem from their relatively simple architectures and small number of parameters, making it difficult for them to effectively integrate both true labels and visual features during training on the challenging tasks of the ImageNet-EEG and THINGS-EEG-5Hz datasets. The Temporal CNN achieves an accuracy increase of 1.26% for the ImageNet-EEG dataset ($p = 7.305e^{-5}$) and 0.77% for the THINGS-EEG-5Hz dataset ($p = 3.130e^{-23}$) with the assistance of our contrastive learning algorithm. It is noteworthy that the B.D. model encounters gradient explosion issues during training on the ImageNet-EEG and THINGS-EEG-5Hz datasets. However, with our contrastive learning algorithm, B.D. achieves higher accuracy than both EEGNet and ShallowConvNet at 1.75%.

Given the strong performance of the hybrid model across all three datasets, we use its results as a baseline for our subsequent experiments. In addition to the EEG models and datasets, the variation in the effects of our contrastive learning algorithm is also related to the hyperparameters of the training strategies. To further investigate the influence of these hyperparameters, we conducted extensive experiments with the hybrid model, and the results are discussed in Section 5.

Overall, the quantitative results indicate that our contrastive learning algorithm is an effective method for EEG visual classification. The visual embeddings extracted by the image encoder can efficiently assist the EEG encoder in extracting EEG embeddings relevant to visual classification tasks. This approach enhances the classification performance of all five models on specific datasets, demonstrating its ability to derive more discriminative fea-

tures from EEG signals.

Table 3: Alation Experiments of the Hybrid Model for Three EEG Datasets.

Method	EEG72		ImageNet-EEG	THINGS-EEG-5Hz
	6-class	72-class	40-class	1,654-class
Baseline	55.13 ± 1.49	26.58 ± 1.23	10.89 ± 0.77	1.29 ± 0.31
+ KD	<u>58.59</u> ± 1.39*	28.64 ± 1.44*	<u>15.91</u> ± 0.28*	<u>3.23</u> ± 0.25*
+ SCL	55.32 ± 1.49	<u>28.69</u> ± 1.33*	13.70 ± 0.36*	0.74 ± 0.56
+ KD + SCL	**59.08** ± 1.39*	**29.60** ± 1.47*	**16.15** ± 0.31*	**3.83** ± 0.58*

where **bold** fonts indicate the best results, and <u>underlined</u> fonts are the second best results.
$\star$ denotes the method are significantly better than the baseline (paired t-test $p<5e^{-4}$).
gray fonts indicate the training suffered a gradient collapse during contrastive learning.

4.5. Ablation studies

In this section, we evaluated the proposed method on the hybrid model with supervised contrastive learning based on features learned through the cross-modal knowledge distillation algorithm. Table 3 presents the classification performance of the hybrid model across four conditions for EEG72, ImageNet-EEG, and THINGS-EEG-5Hz datasets.

For the EEG72 dataset, using only cross-modal knowledge distillation methods improved performance by 3.47% in the 6-class task ($p = 2.574e^{-25}$) and 2.06% in the 72-class task ($p = 6.800e^{-14}$). In the case of the ImageNet-EEG dataset, this approach resulted in a 5.02% improvement in the 40-class task ($p = 5.713e^{-9}$). For the THINGS-EEG-5Hz dataset, it improved performance by 1.94% in the 1,654-class task ($p = 3.103e^{-54}$). Notably, considering the number of categories for classification tasks, the cross-modal knowledge distillation methods produced greater improvements in the ImageNet-EEG and THINGS-EEG-5Hz datasets compared to the 6-class task of the EEG72 dataset. This suggests that as task complexity increases, the model benefits more from the guidance provided by image encoders, enabling it to tackle difficult tasks effectively. Since EEG signals are more challenging to collect than images, training data is often limited for EEG visual classification tasks. Therefore, it is more beneficial to learn information from teacher models for multi-category EEG visual classification.

Moreover, when incorporating supervised contrastive learning methods, we observed further improvements across all three datasets. For the EEG72

dataset, the proposed supervised contrastive learning method further improved performance by 0.49% in the 6-class task ($p = 6.348e^{-5}$) and 0.96% in the 72-class task ($p = 2.396e^{-7}$). In the case of the ImageNet-EEG dataset, the method resulted in a 0.24% improvement ($p = 3.526e^{-2}$). For the THINGS-EEG-5Hz dataset, it improved performance by 0.60% ($p = 1.411e^{-38}$). However, using only supervised contrastive learning without cross-modal knowledge distillation led to mixed results. Specifically, for the EEG72 dataset, supervised contrastive learning achieved a minor improvement of 0.19% in the 6-class task ($p = 0.1543$) and a major improvement of 2.11% in the 72-class task ($p = 7.225e^{-17}$). For the ImageNet-EEG dataset, it resulted in a significant improvement of 2.81% in the 40-class task ($p = 4.635e^{-6}$). Conversely, for the THINGS-EEG-5Hz dataset, it decreased performance by 0.55% in the 1,654-class task ($p = 2.328e^{-24}$).

Overall, the combination of cross-modal knowledge distillation and supervised contrastive learning enhanced the classification performance of the model across the three EEG datasets. In contrast, using only supervised contrastive learning methods produced uncertain effects The results highlight the potential for additional deep-learning algorithms to extract more discriminative features from EEG signals based on the embeddings learned from the teacher model.

Table 4: Accuracy (%) of the Hybrid Model for Eight Image Encoders.

Image Encoder	EEG72		ImageNet-EEG	THINGS-EEG-5Hz
	6-class	72-class	40-class	1,654-class
AlexNet	58.78 ± 1.45	29.52 ± 1.50	16.57 ± 0.28	**4.17** ± 0.49
ResNet18	<u>59.03</u> ± 1.34	<u>29.68</u> ± 1.18	<u>16.74</u> ± 0.45	3.20 ± 0.27
ResNet34	58.79 ± 1.37	29.11 ± 1.23	16.51 ± 0.21	3.38 ± 0.28
ResNet50	57.72 ± 1.31	29.64 ± 1.24	15.97 ± 0.34	3.20 ± 1.47
ResNet101	57.85 ± 1.39	29.44 ± 1.24	16.25 ± 0.60	2.86 ± 1.41
ViT	**59.08** ± 1.39	29.64 ± 1.47	16.15 ± 0.31	3.83 ± 0.58
CLIP-ViT-base	58.72 ± 1.47	29.59 ± 1.11	**16.88** ± 0.48	<u>4.13</u> ± 0.32
CLIP-ViT-large	58.83 ± 1.52	**30.10** ± 1.37	16.66 ± 0.53	4.04 ± 0.26

where **bold** fonts indicate the best results, and <u>underlined</u> fonts are the second best results.

5. Discussion

5.1. Impact of image encoder

In this section, we evaluated the hybrid model on three EEG datasets to assess the impact of different image encoders. Table 4 presents the classification performance across eight image encoders. The AlexNet, ResNet series,

and the ViT are pre-trained on the ImageNet dataset, while the CLIP models are pre-trained on a dataset of 400 million (image, text) pairs collected from the internet.

Different image encoders extract various visual features, which affects the guidance direction for EEG encoders when using the knowledge distillation algorithm. As shown in Table 4, there are significant differences in the effectiveness of different image encoders. The most notable difference is observed in the THINGS-EEG-5Hz dataset. When employing ResNet34, the performance is 3.38%. This improves to 4.17% with features from AlexNet ($p = 1.655e^{-18}$), 3.83% with features from ViT ($p = 9.897e^{-21}$), and 4.13% with features from CLIP-ViT-base ($p = 1.216e^{-23}$). For the ImageNet-EEG dataset, CLIP-ViT-base achieves the best performance, being 0.31% higher than AlexNet ($p = 3.338e^{-2}$), 0.14% higher than ResNet18 ($p = 0.5512$), and 0.73% higher than ViT ($p = 8.495e^{-3}$). In the 6-class task of EEG72, ResNet18, ViT, and CLIP-ViT-large demonstrate similar performance levels, with the best performance reaching 59.08% using ViT features, which is 0.30% better than AlexNet ($p = 2.686e^{-2}$). For the more challenging 72-class task of EEG72, CLIP-ViT-large significantly outperforms other models, achieving 30.10%, which is 0.58% higher than AlexNet ($p = 1.902e^{-2}$), 0.42% higher than ResNet18 ($p = 3.393e^{-2}$), and 0.46% higher than ViT ($p = 2.651e^{-2}$).

Moreover, within the ResNet series, using larger ResNet models does not enhance the accuracy of EEG visual classification. This may be due to larger image encoders with more parameters tending to overfit the training data, whereas the generalizability of the images used in EEG acquisition is lower. Such overfitting can negatively impact the effectiveness of the knowledge distillation algorithm in guiding EEG encoder learning. In contrast, CLIP models generally exhibit better stability and higher accuracy, making them more suitable than the ResNet series. These models consistently achieve the best or second-best performance on more difficult tasks, except for the 6-class task of EEG72.

In conclusion, based on our experimental results, we recommend using CLIP models as image encoders for knowledge distillation in most cases. However, for data from different sources, ResNet18, AlexNet, or ViT may sometimes yield better results.

Table 5: Accuracy (%) of the Hybrid Model for Three Knowledge Distillation Loss Functions.

Knowledge Distillation Loss	EEG72		ImageNet-EEG	THINGS-EEG-5Hz
	6-class	72-class	40-class	1,654-class
Huber	59.08 ± 1.39	29.64 ± 1.47	16.15 ± 0.31	**3.83** ± 0.58*
InfoNCE	**59.86** ± 1.44*	**31.88** ± 1.41*	16.07 ± 0.57	2.78 ± 0.21
KL	58.96 ± 1.41	29.40 ± 1.42	**16.35** ± 0.29	3.63 ± 0.30

where **bold** fonts indicate the best results. $\star$ denotes the classification accuracy of the model trained with this specific knowledge distillation loss function are significantly better than other loss functions (paired t-test $p<5e^{-4}$).

5.2. Impact of knowledge distillation loss function

In this section, we evaluated the hybrid model on three EEG datasets to assess the impact of different knowledge distillation loss functions. Table 5 presents the classification performance across three knowledge distillation loss functions.

During the training of model, different knowledge distillation loss functions employ various methods to assess the similarity between visual embeddings and EEG embeddings, leading to distinct directions in gradient descent. The details and equations of these loss functions can be found in Section 3.2.

The experimental results indicate that the optimal loss function varies across different datasets. As illustrated in Table 5, among the three knowledge distillation loss functions, Huber loss outperforms the others on the THINGS-EEG-5Hz dataset, surpassing InfoNCE loss ($p = 2.033e^{-52}$) and showing a slight improvement compared to KL loss ($p = 8.744e^{-17}$). In contrast, on the EEG72 dataset, InfoNCE loss is more effective than both Huber loss ($p = 4.532e^{-19}$) and KL loss ($p = 6.455e^{-24}$). For the ImageNet-EEG dataset, the performances of all three loss functions are quite similar.

In summary, when assessing the effectiveness of knowledge distillation loss functions across different EEG datasets, Huber loss demonstrates performance comparable to KL loss, while it exhibits a notable difference from InfoNCE loss. Future research could explore the design of new knowledge distillation loss functions to achieve more robust performance.

5.3. Impact of contrastive loss function

In this section, we evaluated the hybrid model on three EEG datasets to assess the impact of different contrastive loss functions. Table 6 presents the classification performance associated with two contrastive loss functions.

Table 6: Accuracy (%) of the Hybrid Model for Two Contrastive Loss Functions.

Contrastive Loss		EEG72		ImageNet-EEG	THINGS-EEG-5Hz
		6-class	72-class	40-class	1,654-class
Unsupervised	triplet	**59.54** ± 1.64*	**31.06** ± 1.38*	14.39 ± 0.62	1.80 ± 0.43
	SupCon	58.64 ± 1.69	29.05 ± 1.88	13.89 ± 0.47	2.29 ± 0.23*
Supervised	triplet	59.08 ± 1.39	29.64 ± 1.47	**16.15** ± 0.31*	**3.83** ± 0.58*
	SupCon	58.80 ± 1.30	30.27 ± 1.47*	14.43 ± 0.59	3.03 ± 0.26

where **bold** fonts indicate the best results. $\star$ denotes the classification accuracy of the model trained with this specific contrastive loss function are significantly better than the other loss function (paired t-test $p<5e^{-4}$).

The lower half of the Table 6 shows that the choice of loss function in supervised contrastive learning affects performance across different datasets. Specifically, the triplet loss outperforms the SupCon loss on the ImageNet-EEG ($p = 2.940e^{-5}$) and THINGS-EEG-5Hz ($p = 6.202e^{-39}$) datasets. However, the performance difference between the triplet loss and SupCon loss is minimal for the 6-class task of EEG72 ($p = 1.825e^{-2}$). Additionally, for the 72-class task of EEG72, the triplet loss performs worse than the SupCon loss ($p = 7.811e^{-5}$).

The upper half of the Table 6 demonstrates that the choice of loss function in unsupervised contrastive learning affects performance across different datasets. Notably, on simpler tasks in EEG72, unsupervised learning outperformed supervised learning. Specifically, for the 6-class task of EEG72, unsupervised contrastive learning using triplet loss achieves 0.46% higher accuracy than supervised contrastive learning using triplet loss ($p = 1.084e^{-5}$). For the 72-class task of EEG72, unsupervised contrastive learning using triplet loss achieves 0.79% higher accuracy than supervised contrastive learning using SupCon loss ($p = 1.449e^{-59}$).

However, in the relatively more challenging tasks of ImageNet-EEG and THINGS-EEG-5Hz, supervised learning significantly outperformed unsupervised learning. For the 40-class task of ImageNet-EEG, supervised contrastive learning using triplet loss achieves 1.76% higher accuracy than unsupervised contrastive learning using triplet loss ($p = 3.520e^{-2}$). For the 1,654-class task of THINGS-EEG-5Hz, supervised contrastive learning using triplet loss achieves 1.54% higher accuracy than unsupervised contrastive learning using SupCon loss ($p = 7.528e^{-10}$).

This discrepancy may arise from that unsupervised contrastive learning treats each sample as a positive while considering all others as negatives. On relatively simple tasks of EEG72, this allows the model to better distinguish between individual samples, improving classification accuracy. Conversely,

distinguishing each sample from all others becomes more challenging on more complex tasks of ImageNet-EEG and THINGS-EEG-5Hz, making supervised contrastive learning a more effective approach for capturing catogery similarities with label information.

In summary, when selecting a supervised contrastive loss function, much like the knowledge distillation loss function, the suitable loss can change depending on the dataset. It is also worth noted that the settings of hyperparameters and the weights assigned to the loss functions also influence the model's training performance. We conducted experiments on the effects of loss weight and hyperparameter settings, which are discussed in Section 5.4.

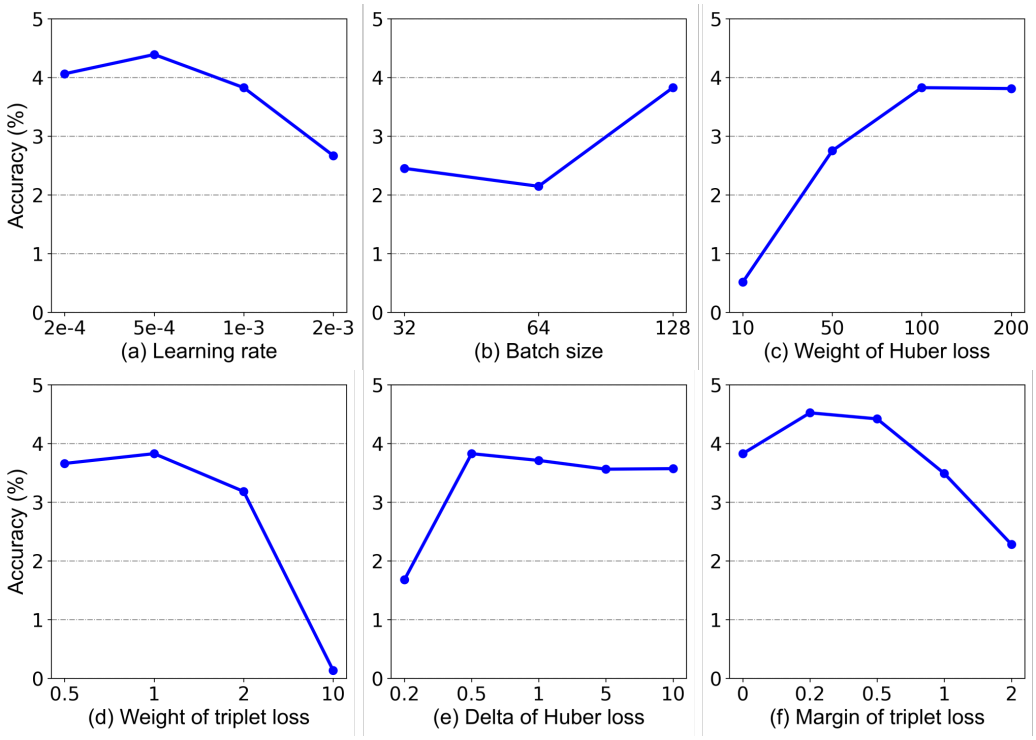


Figure 3: Performance comparison with different hyperparameters on THINGS-EEG-5Hz dataset. (a) Learning rate. (b) Batch size. (c) Weight of Huber loss. (d) Weight of triplet loss. (e) Delta of Huber loss. (f) Margin of triplet loss.

5.4. Parameter sensitivity

Deep-learning models are sensitive to hyperparameter settings, making it essential to optimize these parameters for achieving a high-performance

model. In this section, we experimented on THINGS-EEG-5Hz dataset and made detailed adjustments to the hyperparameters related to training parameters and loss functions.

Learning rate and batch size are the basic parameters that affect the performance of deep learning algorithms. From Fig. 3 (a), it can be observed that setting the learning rate to $5e^{-4}$ resulted in the highest accuracy. A learning rate of $2e^{-3}$ may cause the model to oscillate around the global optimum, resulting in lower accuracy. Meanwhile, a learning rate of $2e^{-4}$ may lead to the model falling into a local optimum, also resulting in lower accuracy.

From Fig. 3 (b), it can be observed that setting batch size to 128 leads to the highest accuracy. A small batch size of may cause the gradient descent process unstable, resulting in lower accuracy. Beside, in contrastive learning, a larger batch size allows for more positive and negative sample pairs to be contrasted, which enhances the training. Due to the memory size of our GPU, we didn't experiment with larger batch sizes.

We also conducted experiments on the hyperparameters of loss functions, including the weight of the Huber loss, the weight of the triplet loss, the delta of the Huber loss, and the margin of the triplet loss. Fig. 3 (c) and (d) show that using inappropriate weights for these losses significantly decreased the model's performance. Specifically, when the weight of the Huber loss was set too low at 10 and the weight of the triplet loss was set too high at 10, supervised contrastive learning became dominant. This imbalance led to an unstable training process and a sharp decline in accuracy.

Furthermore, as shown in Fig. 3 (e), the delta of the Huber loss should not be set too low, as in the case of 0.2. When the absolute difference between the actual and predicted values exceeds the delta (i.e., $|y - \hat{y}| > \delta$), the gradient is approximately equal to the delta. If this delta is set too small, it can adversely affect the model's convergence.

Additionally, Fig. 3 (f) shows that optimizing the margin of the triplet loss further improved the model's performance, achieving an accuracy increase of 4.52% when set to a value of 0.2. However, when the margin was set too high at 2, the accuracy declined.

5.5. Contrastive learning on different data size

In this section, we examine the performance of our contrastive learning algorithm across various data sizes. Given the limited number of samples per category in the 72-class task for EEG72 and the 1,654-class task for

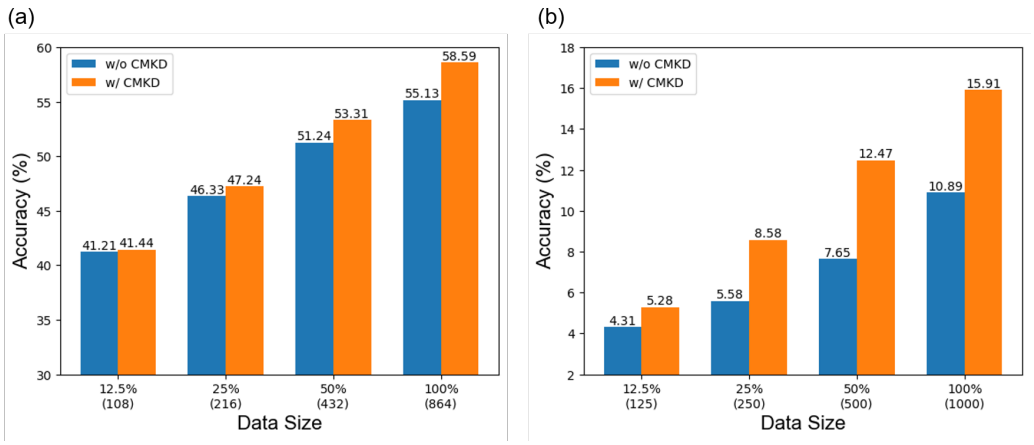


Figure 4: Performance comparison with different data size on EEG72 and ImageNet-EEG dataset. (a) EEG72. (b) ImageNet-EEG. The numbers in parentheses are the number of samples for each class.

THINGS-EEG, we decided to conduct experiments on the 6-class task for EEG72 and the 40-class task for ImageNet-EEG. We utilized 12.5%, 25%, and 50% of the training samples from both datasets while keeping the test samples consistent across all experiments.

As illustrated in Fig. 4 (a), the results of the EEG72 dataset reveals that the performance of our contrastive learning algorithm improves significantly with the increasing number of training samples. However, when the training sample size is as low as 12.5%, there is barely any noticeable performance improvement. This suggests that the method has the potential for further performance enhancements as the dataset size increases.

In Fig. 4 (b), the results for the ImageNet-EEG dataset indicate that at a training sample size of 12.5%, the knowledge distillation method achieves a modest performance improvement of approximately 23%. As the training samples increase to 25%, 50%, and 100%, the performance enhancements rise to about 54%, 63%, and 48%, respectively. This indicates that with 1,000 samples per category, our contrastive learning algorithm consistently demonstrates the ability to improve performance and shows promise for achieving even better results with larger sample sizes.

In summary, our contrastive learning algorithm for EEG deep learning models can improve performance across different data sizes, but it is more effective with larger data sizes, as deep learning models and algorithms are

designed for substantial volumes of data.

5.6. Limitations and future work

Although the prposed method shows promising results in the visual classification of EEG signals, there are still some limitations that need to be addressed. Firstly, in the cross-modal knowledge distillation algorithm, there is a distinction between the visual mode and the EEG mode. Knowledge distillation in the EEG domain can be performed if a robust, large EEG model is available, potentially improving performance. Secondly, the supervised contrastive learning method we introduced relies on label information. However, contrastive learning can also be effectively applied in unsupervised scenarios. Future work could investigate pre-trained algorithms based on unsupervised contrastive learning. In our future work, we plan to explore pre-trained EEG models to improve the accuracy of EEG-based visual classification. Additionally, we intend to investigate various deep learning algorithms. This may include employing data augmentation and domain adaptation techniques to enhance the model's generalization performance.

6. Conclusion

In this study, we present a contrastive learning algorithm that utilizes cross-modal information for EEG visual classification. Our method combines features obtained from knowledge distillation and supervised contrastive learning, leading to a significant improvement in the accuracy of five EEG models across three datasets. We also examine the effects of eight pre-trained image encoders, three knowledge distillation loss functions, two contrastive learning loss functions, and various hyperparameters on the performance of our algorithm. The results indicate that image encoders can effectively facilitate knowledge distillation without the need for excessively large models. Furthermore, the optimal loss function varies depending on the dataset, and both the weights and hyperparameters of the loss have a substantial impact on performance. Overall, our research enhances the classification performance of deep learning algorithms for EEG visual recognition tasks and supports EEG research for practical applications in brain-computer interfaces.

CRediT authorship contribution statement

Shuning Xue: Conceptualization, Methodology, Experiment, Writing. **Jie Jiang:** Conceptualization, Methodology, Revised the manuscript. **Longteng**

Guo: Conceptualization, Methodology, Revised the manuscript. **Jing Liu:** Conceptualization, Methodology, Revised the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research is supported by Artificial Intelligence-National Science and Technology Major Project (2023ZD0121200) and the National Natural Science Foundation of China (U21B2043, 6243000159, 62102416), and the Key Research and Development Program of Jiangsu Province under Grant BE2023016-3.

References

[1] E. Colombari, G. Parisi, A. Tafuro, S. Mele, C. Mazzi, S. Savazzi, Beyond primary visual cortex: The leading role of lateral occipital complex in early conscious visual processing, NeuroImage 298 (2024) 120805.

[2] J. Wang, J. Wang, J. Hu, S. Tong, X. Hong, J. Sun, Willed attentional selection of visual features: an eeg study, IEEE Transactions on Neural Systems and Rehabilitation Engineering (2024).

[3] F. Mushtaq, D. Welke, A. Gallagher, Y. G. Pavlov, L. Kouara, J. Bosch-Bayard, J. J. van den Bosch, M. Arvaneh, A. R. Bland, M. Chaumon, et al., One hundred years of eeg for brain and behaviour research, Nature Human Behaviour 8 (8) (2024) 1437–1443.

[4] A. Pirasteh, M. Shamseini Ghiyasvand, M. Pouladian, Eeg-based brain-computer interface methods with the aim of rehabilitating advanced stage als patients, Disability and Rehabilitation: Assistive Technology (2024) 1–11.

[5] Z. Yi, J. Pan, Z. Chen, D. Lu, H. Cai, J. Li, Q. Xie, A hybrid bci integrating eeg and eye-tracking for assisting clinical communication in patients with disorders of consciousness, IEEE Transactions on Neural Systems and Rehabilitation Engineering (2024).

[6] J. Su, J. Wang, W. Wang, Y. Wang, C. Bunterngchit, P. Zhang, Z.-G. Hou, An adaptive hybrid brain computer interface for hand function rehabilitation of stroke patients, IEEE Transactions on Neural Systems and Rehabilitation Engineering (2024).

[7] J. Mathew, E. Kanaga, M. Bhuvaneshwari, C. Stephen, P. S. Sarkar, Development of a web application for audio communication using eeg signals., Grenze International Journal of Engineering & Technology (GIJET) 10 (1) (2024).

[8] J. Luo, W. Cui, S. Xu, L. Wang, X. Li, X. Liao, Y. Li, A dual-branch spatio-temporal-spectral transformer feature fusion network for eeg-based visual recognition, IEEE Transactions on Industrial Informatics 20 (2) (2023) 1721–1731.

[9] V. Gashaj, D. Trninić, C. Formaz, S. Tobler, J. S. Gómez-Cañón, H. Poikonen, M. Kapur, Bridging cognitive neuroscience and education: Insights from eeg recording during mathematical proof evaluation, Trends in Neuroscience and Education (2024) 100226.

[10] T. Grootswagers, I. Zhou, A. K. Robinson, M. N. Hebart, T. A. Carlson, Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams, Scientific Data 9 (1) (2022) 3.

[11] M. A. Pfeffer, S. S. H. Ling, J. K. W. Wong, Exploring the frontier: Transformer-based models in eeg signal analysis for brain-computer interfaces, Computers in Biology and Medicine (2024) 108705.

[12] C. Du, K. Fu, J. Li, H. He, Decoding visual neural representations by multimodal learning of brain-visual-linguistic features, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

[13] H. Chen, D. Wang, M. Xu, Y. Chen, Cre-tscae: A novel classification model based on stacked convolutional autoencoder for dual-target rsvp-bci tasks, IEEE Transactions on Biomedical Engineering (2024).

[14] Z. Zhao, Y. Lin, Y. Wang, X. Gao, Single-trial eeg classification using spatio-temporal weighting and correlation analysis for rsvp-based collaborative brain computer interface, IEEE Transactions on Biomedical Engineering (2023).

[15] Z. Ye, L. Yao, Y. Zhang, S. Gustin, Self-supervised cross-modal visual retrieval from brain activities, Pattern Recognition 145 (2024) 109915.

[16] H. Ahmadieh, F. Gassemi, M. H. Moradi, Visual image reconstruction based on eeg signals using a generative adversarial and deep fuzzy neural network, Biomedical Signal Processing and Control 87 (2024) 105497.

[17] H. Ahmed, R. B. Wilbur, H. M. Bharadwaj, J. M. Siskind, Object classification from randomized eeg trials, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3845–3854.

[18] Y. Song, B. Liu, X. Li, N. Shi, Y. Wang, X. Gao, Decoding natural images from eeg for object recognition, arXiv preprint arXiv:2308.13234 (2023).

[19] M. Ferrante, T. Boccato, S. Bargione, N. Toschi, Decoding eeg signals of visual brain representations with a clip based knowledge distillation, in: ICLR 2024 Workshop on Learning from Time Series For Health, 2024.

[20] D. Li, C. Wei, S. Li, J. Zou, Q. Liu, Visual decoding and reconstruction via eeg embeddings with guided diffusion, arXiv preprint arXiv:2403.07721 (2024).

[21] G. Hinton, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).

[22] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, Y. Bengio, Fitnets: Hints for thin deep nets, arXiv preprint arXiv:1412.6550 (2014).

[23] C.-S. Chen, C.-S. Wei, Mind's eye: Image recognition by eeg via multimodal similarity-keeping contrastive learning, arXiv preprint arXiv:2406.16910 (2024).

[24] P. O. de Paula, T. B. da Silva Costa, R. R. de Faissol Attux, D. G. Fantinato, Classification of image encoded ssvep-based eeg signals using convolutional neural networks, Expert Systems with Applications 214 (2023) 119096.

[25] H. Zhang, Z. Wang, Y. Yu, H. Yin, C. Chen, H. Wang, An improved eegnet for single-trial eeg classification in rapid serial visual presentation task, Brain Science Advances 8 (2) (2022) 111–126.

[26] B. Kaneshiro, M. Perreau Guimaraes, H.-S. Kim, A. M. Norcia, P. Suppes, A representational similarity analysis of the dynamics of object processing using single-trial eeg classification, Plos one 10 (8) (2015) e0135697.

[27] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, N. Souly, M. Shah, Deep learning human mind for automated visual classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6809–6817.

[28] R. Li, J. S. Johansen, H. Ahmed, T. V. Ilyevsky, R. B. Wilbur, H. M. Bharadwaj, J. M. Siskind, The perils and pitfalls of block design for eeg classification experiments, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (1) (2020) 316–333.

[29] J. Kalafatovich, M. Lee, S.-W. Lee, Decoding visual recognition of objects from eeg signals based on attention-driven convolutional neural network, in: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2020, pp. 2985–2990.

[30] S. Bagchi, D. R. Bathula, Adequately wide 1d cnn facilitates improved eeg based visual object recognition, in: 2021 29th European Signal Processing Conference (EUSIPCO), IEEE, 2021, pp. 1276–1280.

[31] J. Kalafatovich, M. Lee, S.-W. Lee, Learning spatiotemporal graph representations for visual perception using eeg signals, IEEE Transactions on Neural Systems and Rehabilitation Engineering 31 (2022) 97–108.

[32] S. Xue, B. Jin, J. Jiang, L. Guo, J. Liu, A hybrid local-global neural network for visual classification using raw eeg signals, Scientific Reports 14 (1) (2024) 27170.

[33] Z. Jiao, H. You, F. Yang, X. Li, H. Zhang, D. Shen, Decoding eeg by visual-guided deep neural networks., in: IJCAI, Vol. 28, Macao, 2019, pp. 1387–1393.

[34] A. T. Gifford, K. Dwivedi, G. Roig, R. M. Cichy, A large and rich eeg dataset for modeling human visual object recognition, NeuroImage 264 (2022) 119754.

[35] M. N. Hebart, O. Contier, L. Teichmann, A. H. Rockter, C. Y. Zheng, A. Kidder, A. Corriveau, M. Vaziri-Pashkam, C. I. Baker, Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior, Elife 12 (2023) e82580.

[36] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, B. J. Lance, Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces, Journal of neural engineering 15 (5) (2018) 056013.

[37] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for eeg decoding and visualization, Human brain mapping 38 (11) (2017) 5391–5420.

[38] Y. Benchetrit, H. Banville, J.-R. King, Brain decoding: toward real-time reconstruction of visual perception, arXiv preprint arXiv:2310.19812 (2023).

[39] H. Raza, A. Chowdhury, S. Bhattacharyya, S. Samothrakis, Single-trial eeg classification with eegnet and neural structured learning for improving bci performance, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–8.

[40] X. Wang, M. Hersche, B. Tömekce, B. Kaya, M. Magno, L. Benini, An accurate eegnet-based motor-imagery brain–computer interface for low-power edge computing, in: 2020 IEEE international symposium on medical measurements and applications (MeMeA), IEEE, 2020, pp. 1–6.

[41] Y. Zhu, Y. Li, J. Lu, P. Li, Eegnet with ensemble learning to improve the cross-session classification of ssvep based bci from ear-eeg, IEEE Access 9 (2021) 15295–15303.

[42] Z. Miao, M. Zhao, X. Zhang, D. Ming, Lmda-net: A lightweight multi-dimensional attention network for general eeg-based brain-computer interfaces and interpretability, NeuroImage (2023) 120209.

[43] M. Riyad, M. Khalil, A. Adib, Incep-eegnet: a convnet for motor imagery decoding, in: Image and Signal Processing: 9th International Conference, ICISP 2020, Marrakesh, Morocco, June 4–6, 2020, Proceedings 9, Springer, 2020, pp. 103–111.

[44] S.-J. Kim, D.-H. Lee, S.-W. Lee, Rethinking cnn architecture for enhancing decoding performance of motor imagery-based eeg signals, IEEE Access 10 (2022) 96984–96996.

[45] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, J.-R. King, Decoding speech perception from non-invasive brain recordings, Nature Machine Intelligence 5 (10) (2023) 1097–1107.

[46] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[47] P. J. Huber, Robust estimation of a location parameter, in: Breakthroughs in statistics: Methodology and distribution, Springer, 1992, pp. 492–518.

[48] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, Advances in neural information processing systems 33 (2020) 18661–18673.

[49] K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification., Journal of machine learning research 10 (2) (2009).

[50] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems 25 (2012).

[51] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[52] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[55] M. N. Hebart, A. H. Dickter, A. Kidder, W. Y. Kwok, A. Corriveau, C. Van Wicklin, C. I. Baker, Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images, PloS one 14 (10) (2019) e0223792.

[56] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, M. S. Hämäläinen, Mne software for processing meg and eeg data, neuroimage 86 (2014) 446–460.

[57] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).