Self-supervised EEG Representation Learning based on Temporal Prediction and Spatial Reconstruction for Emotion Recognition

Ren-Jie Dai (renjiedai@sjtu.edu.cn)
Keya Hu (hu_keya@sjtu.edu.cn)
Hao-Long Yin (yinhaolong@sjtu.edu.cn)
Bao-Liang Lu (bllu@sjtu.edu.cn)
Wei-Long Zheng* (weilong@sjtu.edu.cn)
School of Computer Science, Shanghai Jiao Tong University, Shanghai, China

Abstract

Affective Brain-Computer Interfaces has achieved remarkable advancements, enabling researchers to interpret labeled EEG data accurately. However, the annotation of EEG data is time-consuming and requires substantial effort, which limits the application in practical scenarios. In this paper, we propose a self-supervised EEG representation learning framework based on temporal prediction and spatial reconstruction (EEG-TPSR) to learn EEG representations from a large amount of unlabeled data. Our model consists of two stages: 1) In the pre-training stage, we use contrastive temporal prediction and spatial reconstruction as proxy tasks, which utilize the spatio-temporal information to learn the generic representations from EEG data; 2) In the fine-tuning stage, few data is used to calibrate the pre-trained model. We conduct extensive experiments on three emotion EEG datasets. The results demonstrate that our proposed model achieves excellent performance, with over 20% relative accuracy improvement and more than 15% improvement using only 1% labeled data.

Keywords: Emotion Recognition; EEG; Self-supervised Learning; Affective Computing

Introduction

Emotion is a fundamental aspect of human life, playing a crucial role in communication, decision-making, and learning processes (Dolan, 2002). As EEG signals are directly linked to brain activity, they serve as a more trustworthy and objective source of information for healthcare applications (Y. Wang, Zhang, & Di, 2024; Z. He, Cai, Li, Tian, & Dai, 2024). Affective Brain-Computer Interfaces (aBCIs) offer a technological approach to detecting human emotions through electroencephalogram (EEG) signals (Wu, Lu, Hu, & Zeng, 2023). In particular, aBCIs hold great promise in mental health contexts by enabling objective assessments of emotional disorders. Thus, EEG-based emotion recognition has evolved rapidly in recent years with many excellent studies exploiting intact EEG data in a supervised manner (Jia et al., 2020; R. Li, Wang, & Lu, 2021). However, these approaches typically rely on extensively annotated EEG data, which requires a lot of effort.

Nowadays, the volume of data utilized for training plays an increasingly crucial role in the performance of models. Self-supervised learning (SSL) is a subset of unsupervised learning. It uses the unlabeled data as a signal to extract feature representations for downstream tasks and is beneficial to many types of downstream tasks (X. Liu et al., 2021). Based on vast amounts of unlabeled data, SSL has shown remarkable advantages in fields such as speech recognition (Baevski, Zhou, Mohamed, & Auli, 2020; Shi, Hsu, Lakhotia, & Mohamed, 2022) and natural language processing (Devlin, Chang, Lee, & Toutanova, 2019). Previous research on SSL for temporal physiological signals has also yielded promising results. For instance, in datasets related to sleep stage detection, epilepsy detection, and human activity recognition (Eldele et al., 2021, 2023), the SSL approaches have been demonstrated promising results.

In the task of EEG-based emotion classification, previous studies with various manually extracted EEG features have achieved good results (Duan, Zhu, & Lu, 2013; X.-W. Wang, Nie, & Lu, 2014; Y. Liu & Sourina, 2013). Emotion recognition with SSL has also yielded excellent results on these features (R. Li, Wang, Zheng, & Lu, 2022; Y. Li et al., 2023; Zhang, Liu, & Zhong, 2023). However, such manually extracted features are often tailored to specific datasets and downstream tasks, which constrains the model's scalability and undermines its generalization capability. To better utilize the vast amount of EEG data, especially unlabeled data, and to learn more robust representations of EEG data, it is ultimately necessary to perform SSL on raw EEG signals.

To cope with these issues, we propose a self-supervised framework based on temporal prediction and spatial reconstruction (EEG-TPSR), which fully exploits the spatio-temporal information in unlabeled EEG data. Unlike previous studies that rely on differential entropy features, we train our model directly on raw EEG signals. Our framework pre-trains a generalized encoder by predicting future timesteps and reconstructing masked data. Additionally, we employ contextual contrast to enhance intra-sample similarity while reducing inter-sample similarity. This enables the model to learn robust representations by utilizing spatio-temporal domain information from EEG data, and it requires only a few labeled samples for fine-tuning. Extensive experiments validate that our method learns generalized EEG representations, achieving superior performance in emotion recognition tasks.

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

- We propose a self-supervised EEG representation learning framework based on Temporal Prediction and Spatial Reconstruction (EEG-TPSR) for solving the problems of decoding emotions from few labeled EEG data.

*Corresponding author

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

- Our model consists of an encoder, a temporal predictor, and a spatial reconstruction module, leveraging the spatio-temporal characteristics of EEG signals to enhance the performance of EEG-based emotion recognition.
- Extensive experiments demonstrate that our proposed method effectively learns generalized EEG representations and achieves excellent performance in recognizing emotional states, requiring only a few labeled samples for fine-tuning. Moreover, cross-dataset pre-training highlights the transferability of our approach.

Related Work

Self-supervised Learning

With the continuous improvement of computing and data resources, research based on the self-supervised has gradually developed in recent years. Unlike supervised learning, which is limited by the availability of labeled data, self-supervised methods can learn general representations from a large amount of unlabeled data, and have broad application prospects in fields such as medicine where data labels are difficult to obtain (Balestriero et al., 2023). Considerable models based on self-supervised learning emerge during the past decades, including the models with contrastive learning and generative learning, and so on (X. Liu et al., 2021).

Contrastive learning is based on the principle that different transformations of the same sample should exhibit similarity. It constructs positive and negative pairs through data augmentation or context sampling, and trains the model by maximizing the mutual information between positive pairs while minimizing the mutual information between negative pairs. As an example of this method, Oord, Li, and Vinyals (2018) proposed a contrastive predictive coding model (CPC) to predict future coding features in the latent space and extract useful representation information from high-dimensional data, achieving excellent results in multiple fields. Chen, Kornblith, Norouzi, and Hinton (2020) proposed a contrastive learning model for visual representation (SimCLR), which outperformed supervised training on the ImageNet dataset.

Generative learning with auto-encoder can reconstruct inputs from the corrupted original inputs and construct the representation distribution. The training goal is to minimize the reconstruction error between the original input and the reconstructed input. K. He et al. (2022) proposed masked autoencoders (MAE) as scalable self-supervised learners for computer vision to reconstruct the missing patches in images, which has shown the effectiveness of generative learning in the field of computer vision.

EEG-based Emotion Recognition

Emotion is a high-level cognitive function of human beings and can be detected through physiological signals such as EEG. Compared to other signals like facial expressions and speech, EEG has the advantage of being more difficult to manipulate or disguise, providing a more reliable and objective measure of emotional states. Currently, emotion recognition based on EEG signals is receiving more and more attention. Due to the complexity of EEG signals, one important stage of EEG-based emotion recognition is extracting EEG spectral features, such as differential entropy (DE) (Duan et al., 2013), power spectral density (PSD) (X.-W. Wang et al., 2014) and differential asymmetry (DASM) (Y. Liu & Sourina, 2013).

With the application of deep learning, many emotion recognition algorithms based on EEG have been proposed over the years (X. Li et al., 2022). Zheng and Lu (2015) employed a deep belief network (DBN) to investigate critical frequency bands and channels of EEG signals for emotion recognition. Jiang, Zhao, Guo, and Lu (2021) proposed a graph convolutional network with channel attention (GCNCA) to classify anger and surprise emotions from EEG. Zhong, Wang, and Miao (2020) proposed a regularized graph neural network (RGNN) capturing both local and global inter-channel relations to learn the topological structure of EEG channels. Although these models have achieved superior performance in emotion recognition task, they rely on manually extracted spectral features, which may result in the loss of critical temporal information inherent in the original EEG signals. Furthermore, these approaches are grounded in supervised learning paradigms, which inherently constrain the scalability of the model size.

Recent studies have highlighted the growing use of self-supervised models for EEG emotion decoding. Hu et al. (2024) applied contrastive self-supervised learning to emotion classification with raw EEG signals, improving accuracy and feature transferability in low-data scenarios. Y. Li et al. (2023) introduced a graph-based multi-task self-supervised learning model (GMSS) for EEG emotion recognition, which integrates multiple tasks from space and frequency to learn general representations. Zhang et al. (2023) proposed GANSER, a generative adversarial network-based self-supervised framework that generates diverse and high-quality EEG samples. To better utilize unlabeled and damaged EEG data, R. Li et al. (2022) developed a multi-view spectral-spatial-temporal masked autoencoder (MV-SSTMA) for emotion recognition, using generative self-supervised learning to reconstruct masked EEG channels. These studies demonstrate that self-supervised methods can effectively learn the representational information from EEG features. They also highlight the potential of self-supervised learning on raw EEG signals to extract their spatio-temporal representational information.

Proposed Methods

Overview

We propose a self-supervised EEG representation learning framework based on Temporal Prediction and Spatial Reconstruction (EEG-TPSR). The architecture of EEG-TPSR is shown in Figure 1. The whole model can be divided into two stages: pre-training and fine-tuning. In the pre-training stage, we use temporal prediction with contrasting and spatial reconstruction as proxy tasks. The main goal of this stage is

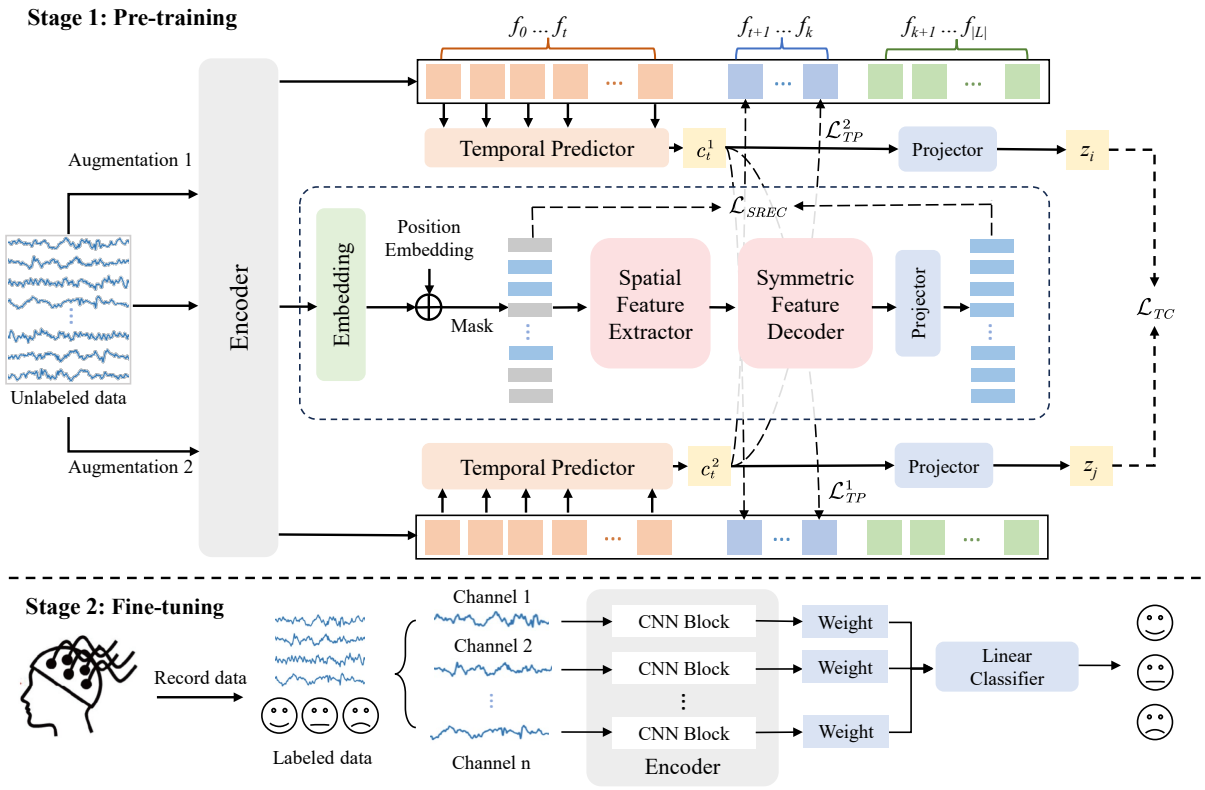

Figure 1: Overall architecture of proposed EEG-TPSR, a self-supervised EEG representation learning framework based on Temporal Prediction and Spatial Reconstruction. **Stage 1: Pre-training** involves temporal prediction and spatial reconstruction tasks to learn spatio-temporal representations from unlabeled EEG data. **Stage 2: Fine-tuning** uses labeled data to calibrate the pre-trained encoder (CNN architecture) followed by a set of learnable weights and a linear classifier for emotion recognition task.

to fully learn the spatial-temporal features of emotion EEG data and complete the pre-training of the encoder. In the fine-tuning stage, the pre-trained encoder is used to extract the features of the raw EEG and then connected to a linear classifier for classification tasks.

Pre-training

In the pre-training stage, we use the unlabeled EEG data to train a generalized encoder with pre-training tasks. Considering the spatio-temporal characteristics of EEG signals, the model consists of an encoder, a temporal predictor, a spatial feature extractor, and a symmetric feature decoder. Next, we will introduce each component in the following subsections.

Encoder We formulate the EEG input as $X \in \mathbb{R}^{C\times T}$, where C is the number of channels and T is the timesteps. For encoder $h_E(\cdot)$, we employ $C\times$ independent CNN blocks for $C\times$ EEG channels, with each block operating without shared parameters. This architecture preserves the unique characteristics of each channel, facilitating tasks such as channel-independent reconstruction. Each CNN block incorporates a set of CNN-1D modules designed to capture temporal information, and the outputs are then normalized and passed through an activation function, followed by a pooling layer that aggregates higher-dimensional information. Followed by Eldele et al. (2021), we randomly select jitter, scale, and permutation as data augmentation strategies, where permutation includes splitting the signal into a random number of segments and randomly shuffling them. Specifically, we apply two different augmentations to raw EEG data x and get $\hat{x}_1$ and $\hat{x}_2$. Then we gain two representation vectors $f^1 = h_E(\hat{x}_1)$ and $f^2 = h_E(\hat{x}_2)$ by passing the augmented data through the encoder.

Temporal Predictor The temporal predictor uses the structure based on the Transformer encoder (Vaswani et al., 2017), with the attention layer performed in the temporal dimension. We stack $L\times$ Transformer blocks to get the temporal predictor module. Inspired by BERT (Devlin et al., 2019), a learnable token $c \in \mathbb{R}^h$ is added for the prediction task and the contrasting learning task, where h is the dimension of an embedding layer.

In the temporal prediction task, we randomly selected a time point t and each pair of vectors $f^1_{\leq t}$ and $f^2_{\leq t}$ will be processed by the temporal predictor $g_{TP}(\cdot)$ to generate $c^1_t = g_{TP}(f^1_{\leq t})$ and $c^2_t = g_{TP}(f^2_{\leq t})$. Then, we conduct the temporal

prediction task by using the representation c_t^1 to predict the future K timesteps of the other augmentation representation $f_{t+i}^2, i \in [1,K]$ and vice versa. To minimize the dot product between the predicted representation and the true one of the same sample while maximizing the dot product with the other samples $N_{t,i}$ within the same batch. The temporal prediction loss is calculated as follows:

$$\mathcal{L}_{TP}^1 = -\frac{1}{K}\sum_{i=1}^{K}\log\left(\frac{\exp\left((W_i(c_t^1)^T)f_{t+i}^2\right)}{\sum_{n\in N_{t,i}}\exp\left((W_i(c_t^1)^T)f_n^2\right)}\right), \quad (1)$$

$$\mathcal{L}_{TP}^2 = -\frac{1}{K}\sum_{i=1}^{K}\log\left(\frac{\exp\left((W_i(c_t^2)^T)f_{t+i}^1\right)}{\sum_{n\in N_{t,i}}\exp\left((W_i(c_t^2)^T)f_n^1\right)}\right), \quad (2)$$

where W_i is a linear function that maps c_t back into the same dimension as f, and L_{TP}^1 and L_{TP}^2 are symmetric loss functions.

In the temporal contrasting task, we apply a non-linear projection head to contexts c and map it to the space where the contrasting task is applied. For the contrasting loss, we use a temporal contrasting loss to maximize the similarity between the positive pair and minimize the similarity between negative pairs. In detail, we will have two augmented views for each sample and thus have $2N$ samples for a batch of N samples. The positive pair (z_i, z_j) is defined as two views derived from the same augmented sample, while the remaining $2N-2$ views in the batch form negative pairs, denoted as (z_i, z_m), where $m \neq i$ and m represents indices of views from different samples. The temporal contrasting loss is calculated as follows:

$$\mathcal{L}_{TC} = -\sum_{i=1}^{N}\log\left(\frac{\exp\left(sim(z_i,z_j)/\tau\right)}{\sum_{m=1}^{2N}\mathbb{1}_{[m\neq i]}\exp\left(sim(z_i,z_m)/\tau\right)}\right), \quad (3)$$

where $sim(\cdot)$ denotes the dot product, $\mathbb{1}_{[m\neq i]}$ is an indicator function, evaluating to 1 iff $m \neq i$, and τ is a temperature parameter.

Spatial Feature Extractor and Symmetric Feature Decoder The spatial feature extractor is based on the CNN-Transformer structure, where the attention layer operates along the spatial dimension. Following the spatial feature extractor, a symmetric feature decoder is employed to reconstruct the masked EEG channels. We stack $H\times$ CNN-Transformer blocks for the spatial feature extractor and symmetric feature decoder. This symmetric architecture ensures a robust decoding process for reconstructing complicated EEG data.

In the spatial reconstruction task, the EEG features will be segmented into patches corresponding to the EEG channels along channel dimension C, with each patch representing a single EEG channel. To retain the spatial information of each EEG channel, sine-cosine positional encoding is applied along the spatial dimension. For the masking step, we randomly sample a visible subset $e_v \in \mathbb{R}^{C_v\times T\times D}$ and mask subset $e_m \in \mathbb{R}^{C_m\times T\times D}$, where D is the feature dimension, $C_v + C_m = C$. Only the e_v is employed as the input of the spatial feature extractor and gets the further extracted feature $\hat{e}_v$. Then the input to the feature decoder consists of further extracted visible features $\hat{e}_v$ and the masked features $\hat{e}_m$, where $\hat{e}_m$ is initialized with randomly generated parameters and concatenated with the $\hat{e}_v$. The symmetric feature decoder outputs the reconstructed EEG feature $e_m^{'}$ for each channel. We use the Mean Squared Error (MSE) to calculate the spatial reconstruction loss:

$$\mathcal{L}_{SREC} = \frac{1}{n}\sum_{i=1}^{n}(e_{m_i} - e_{m_i}^{'})^2, \quad (4)$$

where e_{m_i} and $e_{m_i}^{'}$ denote the original masked features and the reconstructed masked features, respectively.

The overall pre-training loss is formulated as a composite of three distinct components: the prediction loss, the contrasting loss, and the reconstruction loss.

$$\mathcal{L} = \lambda_1\cdot(\mathcal{L}_{TP}^1 + \mathcal{L}_{TP}^2) + \lambda_2\cdot\mathcal{L}_{TC} + \lambda_3\cdot\mathcal{L}_{SREC}, \quad (5)$$

where λ_1, λ_2 and λ_3 are fixed scalar hyperparameters denoting the relative weight of each loss.

Fine-tuning

During the fine-tuning stage, we utilize the labeled training dataset of a specific subject s to fine-tune both the pre-trained encoder and the linear classifier. The fine-tuned model is then evaluated on the validation dataset of subject s to verify its performance and select the best model for subsequent testing. Cross-entropy loss is employed to measure the classification accuracy and guide the optimization process. Additionally, a set of learnable weight, connected after the encoder, are used to visualize the key EEG channels involved in emotion recognition.

Experiments

Datasets and Implementation Details

We use the SEED (Zheng & Lu, 2015), SEED-IV (Zheng, Liu, Lu, Lu, & Cichocki, 2019), and SEED-V (W. Liu, Qiu, Zheng, & Lu, 2021) datasets for both pre-training and fine-tuning. The SEED dataset includes EEG data from 15 subjects, covering three emotional states: happy, sad, and neutral. The SEED-IV and SEED-V datasets are extensions of the original SEED dataset. The SEED-IV contains EEG data from 15 subjects with four emotional states, while SEED-V includes data from 16 subjects and expands the categories to five, adding fear and disgust. For all three datasets, we downsample the continuous raw EEG signals to 200 Hz. Each sample is taken as a 1-second window, with no overlap between consecutive samples, forming the datasets. For SEED, which consists of 15 trials per session, the trials are split into training, validation, and testing sets in a 3:1:1 ratio. Similarly, SEED-IV and SEED-V follow 4:1:1 and 5:2:2 ratios, respectively.

Table 1: The fine-tuning and linear fine-tuning accuracies and standard deviations (acc/std %) of different methods. Bold indicates the best accuracy, and underline indicates the second-best accuracy.

Method	Fine-tuning						Linear Fine-tuning					
	SEED		SEED-IV		SEED-V		SEED		SEED-IV		SEED-V	
	Acc.	Std.	Acc.	Std.	Acc.	Std.	Acc.	Std.	Acc.	Std.	Acc.	Std.
Supervised	0.6825	0.0626	0.4615	0.1336	0.4174	0.0663	0.5077	0.0987	0.3517	0.0875	0.2820	0.0520
CPC	0.6820	0.1140	0.4218	0.1445	0.4288	0.1012	0.6150	0.0756	0.3556	0.0873	0.3744	0.1145
BYOL	0.6505	0.1115	0.4383	0.1423	0.4342	0.1056	0.5843	0.0928	0.3432	0.0956	0.3323	0.0861
MoCo	0.6857	0.0989	0.4444	0.1287	0.3882	0.1024	0.6018	0.0889	0.3887	0.0994	0.3604	0.1047
ContraWR	0.6680	0.1064	0.4530	0.1300	0.4354	0.1070	0.6238	0.0963	0.3753	0.0945	0.3778	0.1186
SimCLR	0.6767	0.0923	0.4371	0.1336	0.4064	0.1193	0.6432	0.0969	0.4031	0.1090	0.4014	0.1138
TS-TCC	0.6875	0.0968	0.4791	0.1473	**0.4652**	0.1122	0.6282	0.0790	0.4154	0.1218	0.3798	0.1088
EEG-TPSR	**0.6967**	0.0799	**0.4918**	0.1378	0.4542	0.1142	**0.6444**	0.0903	**0.4235**	0.1083	**0.4136**	0.1256

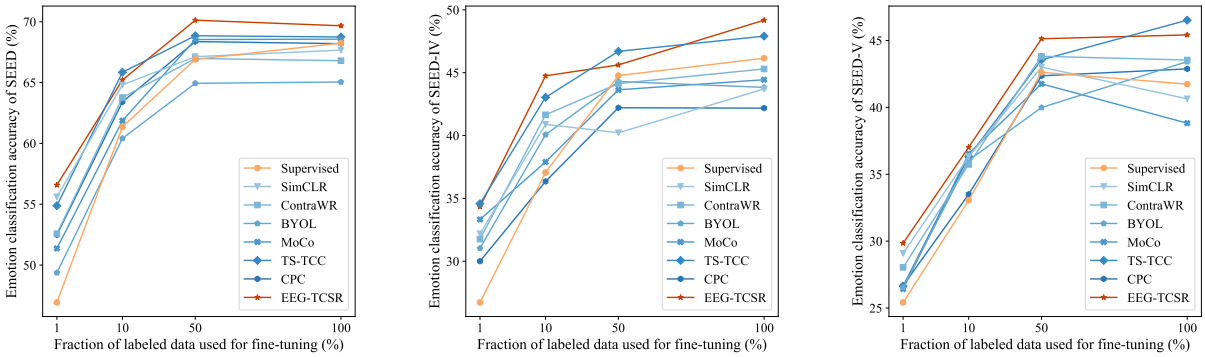

Figure 2: The accuracy of emotion classification with different fractions of fine-tuning data on SEED (left), SEED-IV (middle) and SEED-V (right) datasets.

To ensure comparability, we adopt the same experimental settings for EEG-TPSR and all baseline SSL models. Both the supervised methods and the baseline SSL methods employ a 1D-CNN as the encoder. During pre-training, the model is trained for 100 epochs using the unlabeled training data from all subjects. Fine-tuning is conducted separately for each subject's training data over 50 epochs. Validation datasets are used to optimize hyperparameters, and final performance is evaluated on test datasets. The pre-training learning rate is set to 3e-4 with a batch size of 256, while fine-tuning uses a learning rate of 1e-3. The fine-tuning batch size varies: 16 for datasets $\leq$5%, 32 for datasets $\leq$10%, and 256 otherwise. The encoder employs a 1D-CNN with a kernel size of 8. The temporal prediction step K is 12, with a masking rate of 50%. For the loss function, we set $\lambda_1 = 1$, $\lambda_2 = 0.7$, and $\lambda_3 = 1$.

Experiment Results

Comparison with Baseline Approaches We compare our proposed EEG-TPSR model with the baseline methods under both fine-tuning and linear fine-tuning conditions. In the fine-tuning setup, both the encoder and the linear layer are optimized, while in linear fine-tuning, the encoder is frozen, and only the linear layer is updated. The results are presented in Table 1. Our model consistently achieves state-of-the-art performance across most scenarios on all three datasets, demonstrating the effectiveness of integrating spatio-temporal information during the pre-training phase. Notably, in the linear fine-tuning configuration, our method achieves over a 20% relative accuracy improvement compared to the supervised method, highlighting the strong representational learning capability of EEG-TPSR.

Fine-tuning with Different Amounts of Data To evaluate the robustness of our EEG-TCSR model under low amount of data scenarios, we fine-tune it using varying proportions of the training data: 1%, 10%, 50%, and 100% of randomly selected samples. The performance results are shown in Figure 2, with baseline models consistent with previous configurations.

We observe that supervised training performs poorly with limited labeled data, while our EEG-TCSR achieves significantly better performance than supervised training with only 1% of labeled data. Especially, EEG-TCSR gains a more than

15% relative accuracy improvement with only 1% data volume in all of the three datasets, and it achieves an accuracy of 65.21%, 44.75%, and 37.02% on the SEED, SEED-IV, and SEED-V datasets with only 10% data, respectively.

Cross-Dataset Transfer Experiments We further evaluate the cross-dataset transfer performance of EEG-TPSR by pre-training and fine-tuning on different combinations of the SEED, SEED-IV, and SEED-V datasets. The results, presented in Table 2, demonstrate that in some cases, transferring between datasets yields better performance than training and fine-tuning on the same dataset. This highlights the strong generalizability of our EEG-TPSR model across varying dataset distributions.

Table 2: The fine-tuning accuracy and standard deviation (acc/std %) in cross dataset scenarios.

Pre-training Dataset	Fine-tuning Dataset		
	SEED	SEED-IV	SEED-V
SEED	69.67/7.99	46.90/15.10	43.98/10.03
SEED-IV	69.80/8.64	49.18/13.78	45.70/11.85
SEED-V	70.09/8.43	46.78/15.10	45.42/11.42

Ablation Study

We evaluate the effectiveness of each component in the EEG-TPSR model through an ablation study by isolating the temporal or spatial modules. Table 3 presents the performance of EEG-TPSR, the Spatial Only model, and the Temporal Only model under varying amounts of labeled fine-tuning data on the SEED, SEED-IV, and SEED-V datasets. Our EEG-TPSR consistently outperforms the other models, demonstrating the importance of integrating both spatial and temporal information during pre-training. Additionally, due to the high temporal resolution and low spatial resolution of EEG signals, the Temporal Only model performs better than the Spatial Only model.

Table 3: Ablation study for the classification performance (acc/std %) with different ratio of labeled fine-tuning data on SEED, SEED-IV, and SEED-V datasets.

Model	Ratio	Dataset		
		SEED	SEED-IV	SEED-V
Spatial Only	1%	50.73/11.55	26.31/6.12	21.35/2.75
	10%	61.12/10.22	35.02/11.73	31.66/8.66
	50%	65.44/11.19	42.08/16.09	39.49/13.59
	100%	67.37/9.56	44.15/15.66	41.91/12.49
Temporal Only	1%	56.03/8.08	32.83/8.71	28.36/7.58
	10%	64.94/8.22	43.18/13.31	36.32/12.35
	50%	69.99/7.90	44.70/14.85	44.03/12.74
	100%	69.39/9.61	47.51/14.71	43.75/12.74
EEG-TCSR	1%	56.58/7.33	34.34 /8.39	29.85/6.67
	10%	65.21/8.10	44.75/12.48	37.02/11.92
	50%	70.13/8.46	45.62/14.94	45.13/11.93
	100%	69.67/7.99	49.18/13.78	45.42/11.42

Visualization of Key Brain Regions

By freezing the pre-trained encoders and performing linear fine-tuning, we obtain a set of channel-specific weights, which are visualized as a topographic map in Figure 3. In this map, regions shaded closer to red indicate higher weight values, while those closer to blue represent lower values. The map highlights the importance of different channels in emotional cognition. Notably, regions with high weights align closely with findings from previous studies (Zheng & Lu, 2015; Zheng et al., 2019), particularly in the prefrontal and temporal lobes. This suggests that these brain regions play a critical role in emotion-related tasks, further demonstrating that our model effectively captures the generative representational information of EEG signals.

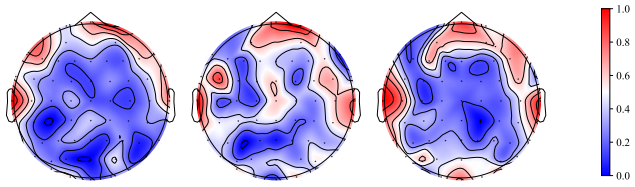

Figure 3: The topography maps represent the average weights of each channel obtained from linear fine-tuning on the SEED (left) , SEED-V (middle) and SEED-V (right) datasets, respectively.

Conclusion

In this paper, we propose EEG-TPSR, a self-supervised EEG representation learning framework based on Temporal Prediction and Spatial Reconstruction to address the challenge of decoding emotions from limited labeled EEG data. The framework comprises an encoder, a temporal predictor, and a spatial reconstruction module, leveraging the spatio-temporal characteristics of EEG signals to enhance the performance of EEG-based emotion recognition. Extensive experiments on the SEED, SEED-IV, and SEED-V datasets demonstrate the superior performance of EEG-TPSR compared to supervised learning and other baseline self-supervised methods. Notably, the model effectively learns robust EEG representations from abundant unlabeled data, achieving high accuracy in decoding emotions with minimal labeled data. Furthermore, our method performs end-to-end emotion decoding directly from raw EEG signals, in contrast to previous approaches that rely on hand-crafted features, and it learns generalized, transferable representations with strong potential to scale up to more diverse and larger EEG datasets.

Acknowledgments

This work was supported in part by grants from National Natural Science Foundation of China (Grant No. 62376158), STI 2030-Major Projects+2022ZD0208500, Shanghai Jiao Tong University 2030 Initiative, Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZD ZX), Medical-Engineering Interdisciplinary Research Foundation of Shanghai Jiao Tong University "Jiao Tong Star" Program (YG2023ZD25, YG2024ZD25 and YG2024QNA03), Shanghai Pujiang Program (Grant No. 22PJ1408600), Shanghai Pilot Program for Basic Research - Shanghai Jiao Tong University (No. 21TQ1400203), and Shanghai Jiao Tong University SCS-Shanghai EmoRays Technology Co., Ltd Joint Laboratory of Affective Brain-Computer Interfaces.

References

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems.*

Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., … others (2023). A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* (pp. 1597–1607).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics.*

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science*, *298*(5596), 1191–1194.

Duan, R.-N., Zhu, J.-Y., & Lu, B.-L. (2013). Differential entropy feature for EEG-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering* (pp. 81–84).

Eldele, E., Ragab, M., Chen, Z., Wu, M., Kwoh, C.-K., & Li, X. (2023). Self-supervised learning for label- efficient sleep stage classification: A comprehensive evaluation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 1333–1342.

Eldele, E., Ragab, M., Chen, Z., Wu, M., Kwoh, C. K., Li, X., & Guan, C. (2021). Time-series representation learning via temporal and contextual contrasting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21* (pp. 2352–2359).

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16000–16009).

He, Z., Cai, M., Li, L., Tian, S., & Dai, R.-J. (2024). EEG-EMG FAConformer: Frequency aware conv-transformer for the fusion of EEG and EMG. In *2024 IEEE International Conference on Bioinformatics and Biomedicine* (pp. 3258–3261).

Hu, K., Dai, R.-J., Chen, W.-T., Yin, H.-L., Lu, B.-L., & Zheng, W.-L. (2024). Contrastive self-supervised EEG representation learning for emotion classification. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 1–4).

Jia, Z., Lin, Y., Cai, X., Chen, H., Gou, H., & Wang, J. (2020). Sst-emotionnet: Spatial-spectral-temporal based attention 3d dense network for EEG emotion recognition. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 2909–2917).

Jiang, W.-B., Zhao, L.-M., Guo, P., & Lu, B.-L. (2021). Discriminating surprise and anger from EEG and eye movements with a graph network. In *2021 IEEE International Conference on Bioinformatics and Biomedicine* (pp. 1353–1357).

Li, R., Wang, Y., & Lu, B.-L. (2021). A multi-domain adaptive graph convolutional network for EEG-based emotion recognition. In *Proceedings of the 29th ACM International Conference on Multimedia* (pp. 5565–5573).

Li, R., Wang, Y., Zheng, W.-L., & Lu, B.-L. (2022). A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 6–14).

Li, X., Zhang, Y., Tiwari, P., Song, D., Hu, B., Yang, M., … Marttinen, P. (2022). EEG based emotion recognition: A tutorial and review. *ACM Computing Surveys*, *55*(4), 1–57.

Li, Y., Chen, J., Li, F., Fu, B., Wu, H., Ji, Y., … Zheng, W. (2023). GMSS: Graph-based multi-task self-supervised learning for EEG emotion recognition. *IEEE Transactions on Affective Computing*, *14*(3), 2512–2525.

Liu, W., Qiu, J.-L., Zheng, W.-L., & Lu, B.-L. (2021). Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2021). Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, *35*(1), 857–876.

Liu, Y., & Sourina, O. (2013). Real-time fractal-based valence level recognition from EEG. In *Transactions on Computational Science XVIII: Special Issue on Cyberworlds* (pp. 101–120).

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

Shi, B., Hsu, W.-N., Lakhotia, K., & Mohamed, A. (2022). Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all

you need. In *Advances in Neural Information Processing Systems* (Vol. 30).

Wang, X.-W., Nie, D., & Lu, B.-L. (2014). Emotional state classification from EEG data using machine learning approach. *Neurocomputing*, *129*, 94–106.

Wang, Y., Zhang, B., & Di, L. (2024). Research progress of EEG-Based emotion recognition: A survey. *ACM Computing Surveys*, *56*(11), 1–49.

Wu, D., Lu, B.-L., Hu, B., & Zeng, Z. (2023). Affective brain–computer interfaces (aBCIs): A tutorial. *Proceedings of the IEEE*, *111*(10), 1314–1332.

Zhang, Z., Liu, Y., & Zhong, S.-h. (2023). GANSER: A self-supervised data augmentation framework for EEG-Based emotion recognition. *IEEE Transactions on Affective Computing*, *14*(3), 2048–2063.

Zheng, W.-L., Liu, W., Lu, Y., Lu, B.-L., & Cichocki, A. (2019). Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics*, *49*(3), 1110–1122.

Zheng, W.-L., & Lu, B.-L. (2015). Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development*, *7*(3), 162–175.

Zhong, P., Wang, D., & Miao, C. (2020). EEG-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing*, *13*(3), 1290–1301.