Multi-view Time-frequency Contrastive Learning for Emotion Recognition

Lei Wang (wanglei2611@mail.dlut.edu.cn)
DaLian University of Technology
Dalian City, Liaoning Province,China

Jianping Zhu (zhujp@mail.dlut.edu.cn)
DaLian University of Technology
Dalian City, Liaoning Province,China

Bo Jin *(jinbo@dlut.edu.cn)
DaLian University of Technology
Dalian City, Liaoning Province,China

Xiaopeng Wei *(xpwei@dlut.edu.cn)
DaLian University of Technology
Dalian City, Liaoning Province,China

Abstract

Electroencephalogram (EEG) signals are physiological indicators of brain activity, offering the advantage of high temporal resolution for capturing subtle emotional changes and providing rich information for emotion recognition. However, extracting effective features from EEG data with a low signal-to-noise ratio poses a significant challenge that hinders progress in this research field. To address this issue, we propose a multi-view time-frequency contrastive learning framework called MV-TFCL to enhance the information representation capability of EEG signals from multiple perspectives. Firstly, we introduce a recursive neural network based on multi-scale time-frequency consistency, which integrates global semantic information across different scales through gated units. To our knowledge, this is the first proposal of the theory of multi-scale time-frequency consistency applied in emotion recognition research. Subsequently, we design a tree-structured time-frequency encoder to capture local semantic information within the time-frequency domain. Finally, we incorporate semantic consistency constraints from both global and local perspectives to learn more generalizable and robust features. Extensive experimental results on two publicly available datasets demonstrate the effectiveness and superiority of our proposed method.

Keywords: Emotion recognition; Contrastive-learning; Multi-scale Time-Frequency Representation

Introduction

Electroencephalogram (EEG) signals serve as a direct reflection of brain activity, offering inherent advantages such as non-invasiveness, real-time monitoring, and portability. Consequently, they have become an indispensable tool for investigating human emotions and consciousness (Guo, Zhu, Zhang, Jin, & Wei, 2023).EEG signals can be directly acquired from the cerebral cortex, rendering them arduous to disguise or conceal, thereby enhancing their authenticity and reliability in comparison to behavioral signals. By harnessing EEG signals, we can explore numerous cutting-edge applications in the realm of emotion recognition, such as EEG-based systems for computing emotions (Liu, Sourina, & Nguyen, 2011), EEG-driven emotional robots (Liu, Habibnezhad, & Jebelli, 2021), and EEG-assisted emotional interventions (Tian et al., 2023). Hence, the accurate recognition of emotions based on EEG signals is a fundamental prerequisite for realizing these applications. Although the contrast learning paradigm has demonstrated some success in enhancing model generalization ability, there remains substantial room for improvement in emotion recognition tasks.

Firstly, the entanglement of time-frequency semantics in EEG signals presents a formidable challenge for achieving universal representation learning. Currently, emotion recognition approaches based on contrastive learning strategies, such as SeqCLR (Mohsenvand, Izadi, & Maes, 2020) and SGMC (Kan et al., 2023), enhance the generalization ability of models through data augmentation and positive sample constraints. However, these methods still possess certain limitations. One concern pertains to the low signal-to-noise ratio (SNR) of EEG signals due to potential interference from brain activity unrelated to the task at hand, which can be attributed to equipment or environmental factors. This results in traditional data augmentation methods further reducing the amount of effective information in the data. Another crucial aspect is the composition of EEG signals, which consist of sinusoidal waves with varying frequencies (Sanei & Chambers, 2013), Each frequency component corresponds to a distinct brain rhythm associated with different cognitive functions and states, such as sleep, attention, and emotions. Conventional approaches often overlook the extraction of key frequency domain signals, thereby compromising the ultimate recognition performance. Therefore, we believe that fully utilizing the time-frequency characteristics of EEG signals can aid in the capture of effective information.

Secondly, the integration of information from different temporal scales constitutes a pivotal factor in constructing an efficacious model for emotion recognition. Numerous studies, such as GMSS (Li et al., 2022), MSTGCN (Jia, Lin, Wang, Ning, et al., 2021), and HetEmotionNet (Jia, Lin, Wang, Feng, et al., 2021), employ a multi-view and multi-task framework to extract comprehensive features from EEG signals by amalgamating diverse pieces of information. However, the annotation

*Corresponding authors.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

process for EEG data is extremely time-consuming and requires extensive medical training or intricate experimental design. As a result, supervised multi-perspective construction incurs additional costs in terms of human and other data resources and often lacks broader generality. In contrast, self-supervised multi-view semantic capture, leveraging the inherent characteristics of EEG signals themselves, presents a more cost-effective approach to explore the interdependence and specificity of each scale.

To address the aforementioned challenges, this paper proposes a multi-view time-frequency contrastive learning framework, termed MV-TFCL, which enhances the information representation capability of EEG signals from multiple perspectives. Specifically, we initially employ a recursive downsampling strategy to divide the data into multiple scales, considering global information at different scales. Next, we design a TF-Block module for capturing global time-frequency representations at each scale and constructing time-frequency consistency constraints within the module to enhance the stability of these representations. Simultaneously, we connect TF-Blocks of different scales through gated units to achieve cross-scale dynamic perception. Following this, we adopt a tree-structured time-frequency encoder for extracting local semantic information of time-frequency. Each node (LocalBlock) consists of a one-dimensional convolutional layer and a frequency enhancement layer, which are responsible for extracting local time-frequency representations. Finally, we further construct consistency constraints from a global-local perspective to improve the robustness of learned representations.

The main contributions of our MV-TFCL can be summarized as follows:

- We propose a multi-view time-frequency contrastive learning framework. To our knowledge, we are the first to introduce and apply the theory of multi-scale time-frequency consistency in the field of emotion recognition.
- We have successfully developed a global-local contrastive learning perspective, unifying the learned global and local representations through consistency constraints, thereby enhancing the robustness of the model.
- Extensive empirical evidence demonstrates that our model achieves optimal performance in emotion recognition tasks.

Related Work

Electroencephalogram (EEG) signals, which reflect brain activity, can be sampled from multiple areas of the cerebral cortex through multi-channel electrodes, providing information in three dimensions: time, space, and frequency. Traditional machine learning approaches often employ handcrafted features, such as statistical measures, discrete wavelet transforms, or power spectral density (Lin et al., 2010), to characterize the temporal and spectral properties of EEG signals. Subsequently, linear or nonlinear classifiers, such as Support Vector Machines (SVM) (Hearst, Dumais, Osuna, Platt, & Scholkopf, 1998), Linear Discriminant Analysis (LDA) (Xanthopoulos et al., 2013), or Logistic Regression (LR) (LaValley, 2008), are utilized for feature classification. while these methods are straightforward and easy to implement, they overlook the spatial dimension of EEG signals, i.e., the interactions between different brain regions, as well as the nonlinearity and high-dimensionality of EEG signals, making it difficult to capture the complex and variable features of EEG signals.

To address these limitations, recent deep learning-based methods have employed multi-layer neural networks to extract high-level and abstract features from EEG signals, thereby enhancing the accuracy of emotion recognition. For instance, Li et al. (Li, Zheng, Wang, Zong, & Cui, 2019) proposed a hierarchical spatio-temporal neural network (R2G-STNN) based on Bidirectional LSTM (BiLSTM) networks, which effectively capture long-term dependencies of EEG signals and consider spatial relationships among different brain regions, resulting in remarkable performance improvements in EEG-based emotion recognition. Song et al. (Song, Zheng, Song, & Cui, 2018) proposed a multi-channel EEG emotion recognition method based on a novel Dynamic Graph Convolutional Neural Network (DGCNN), which can dynamically learn the correlations between different EEG channels, thereby enhancing the expressiveness of the features. Zhong et al. (Zhong, Wang, & Miao, 2020) proposed a Regularized Graph Neural Network (RGNN) with two regularizers, which can handle both the variations in cross-subject EEGs and the problem of noisy labels, thereby improving the robustness of EEG emotion recognition.

Whether employing machine learning or deep learning methods, a large amount of annotated data is essential for training the model. However, acquiring labelled EEG data in real-world scenarios is often costly and challenging, thus self-supervised learning methods have greater potential and practical value in this field. For instance, Kan et al. (Kan et al., 2023) utilized a group projector and devised a new genetic-based data augmentation technique to address the limited labeled data issue in emotion recognition. Zeng et al. (Zeng et al., 2022) extracted deeper and more valuable features using a deep attention layer implemented with multi-head attention mechanisms, and used a Siamese network to cluster the outputs of the GNN based on Euclidean distance, ensuring that the learned information has certain class separability.

Methodology

In this section, we initially provide a concise introduction to the definitions employed in this work. Subsequently, we present both an overview and a comprehensive explanation of the architecture.

Problem Definition

Given a time series $X \in \mathbb{R}^{T_x \times V}$, where T_x represents the length of the input sequence and V represents the number of variables, our goal is to learn a nonlinear function f_θ that maps the input sequence X to a representation $Z \in \mathbb{R}^{Fe}$, where Fe is the dimension of the representation vector. To capture global

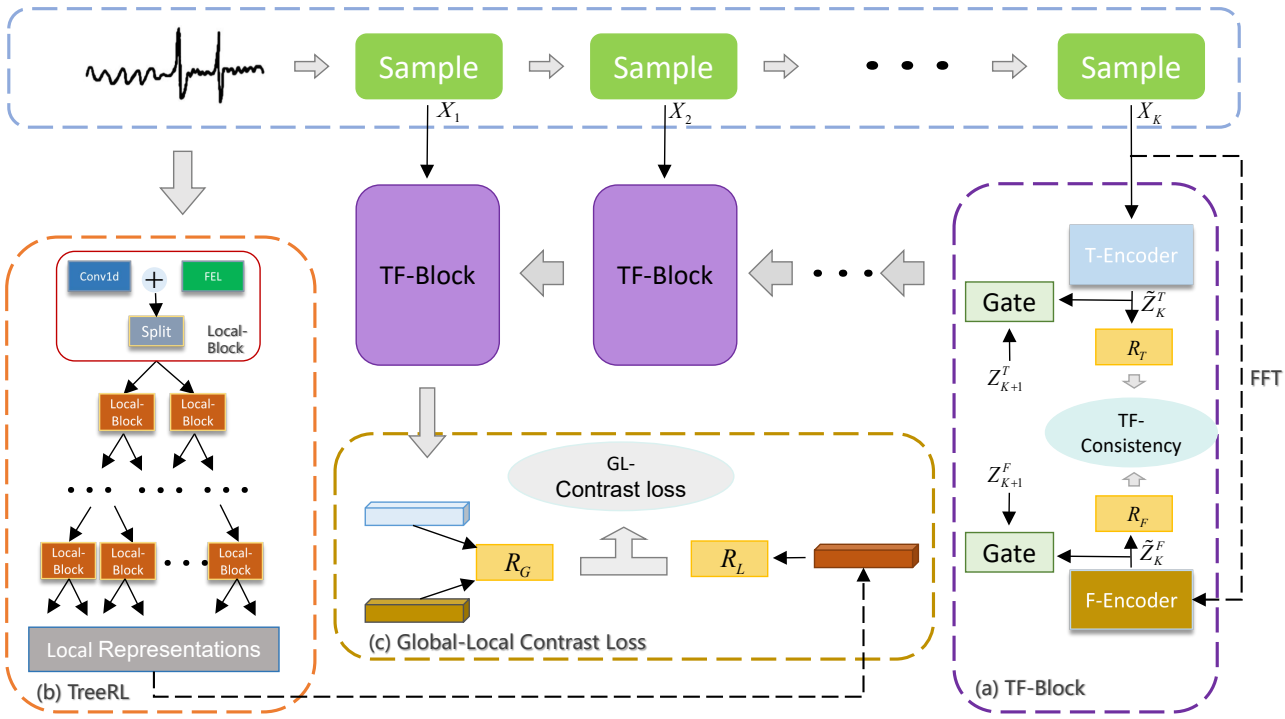

Figure 1: Overview of the proposed MV-TFCL architecture. Specifically, the original time series undergoes recursive downsampling to obtain $\{X_1, X_2, \cdots, X_K\}$, which serves as the input for the (a) TF-Block at the corresponding scale. Then, multiple TF-Blocks are connected by gating units, recursively forming a multi-scale global recurrent neural network (MGRNN). Each TF-Block uses a T-Encoder and an F-Encoder to learn the time-frequency representation at the corresponding scale and completes the calculation of the time-frequency consistency contrast loss. Subsequently, (b) TreeRL learns local representations through a tree decomposition structure. Finally, (c) completes the calculation of the global-local contrast loss.

information at different scales, we first use a recursive downsampling strategy to downsample the data X_{k-1} at the $k-1$ scale along the time dimension every other time step, resulting in $X_k \in \mathbb{R}^{\frac{T_x}{2^k} \times C}$. We then use MGRNN and TreeLR to capture the global representation $Z_G \in \mathbb{R}^{Fe}$ and the local representation $H_L \in \mathbb{R}^{Fe}$, respectively. Finally, we have $Z = Z_G + H_L$.

Multi-Scale Global Recurrent Neural Network

In time series data, capturing global information often implies the need to directly model a time series with a long time span, which poses a significant challenge to the modeling capability of the model. Firstly, substantial noise interference hampers the extraction of representations containing sufficient information content. Secondly, when modeling multiple time steps simultaneously, the model frequently gets trapped in local features, thereby overlooking global information. To address these challenges, we propose a multi-scale global recurrent neural network (MGRNN). Specifically, we first recursively downsample the original time series X along the time dimension, resulting in K time series $\{X_1, X_2, \cdots, X_K\}$ at different scales. Next, we design a time-frequency recursive neural network, where each TF-Block is used to learn the time-frequency representation at the corresponding scale. Different TF-Blocks are connected through gated units, thereby integrating time-domain and frequency-domain representations at different scales. Finally, the final global representation is obtained by concatenating the output representations from both time-domain and frequency-domain at the last scale.

Encoder The TF-Block is composed of a T-Encoder, an F-Encoder, and time-frequency consistency constraints. Specifically, the T-Encoder and F-Encoder are used to learn the global time-frequency representations $\tilde{Z}_k^T \in \mathbb{R}^{Fe}$ and $\tilde{Z}_k^F \in \mathbb{R}^{Fe}$ at the k-th scale, respectively, where Fe is the dimension of the feature vector. The T-Encoder is constructed by stacking one-dimensional convolutions and linear layers:

$$\tilde{Z}_k^T = Fc_k(vec(Conv1d_k(X_k))), \tag{1}$$

where Fc_k and $Conv1d_k$ represent the linear layer and one-dimensional convolutional layer, respectively.

The F-Encoder is used to capture the frequency domain information of EEG signals. For the sample $X_k \in \mathbb{R}^{\frac{T_x}{2^k} \times C}$ at the k-th scale, we first map it to the frequency domain through the Fast Fourier Transform (FFT), i.e., $\mathcal{F}(X_k) \in \mathbb{R}^{I \times C}$, where $I = \lfloor \frac{T_x}{2^{k+1}} \rfloor + 1$. Subsequently, we complete the extraction of

frequency domain information through a linear mapping:

$$\tilde{Z}_k^F = AvgPool(\mathcal{F}^{-1}(\mathcal{F}(X_k) \odot Q_k^F)), \tag{2}$$

where, Q_k^F is a parameterized kernel that is initialized randomly, $\mathcal{F}$ represents the FFT, and $\mathcal{F}^{-1}$ is its inverse. Finally, considering the aggregation of multi-scale time-frequency representations, we use a gating unit to aggregate the time-frequency representations Z_{k+1}^T and Z_{k+1}^F corresponding to the previous scale:

$$\begin{aligned} \eta_T &= Sigmoid(W_\eta^T \cdot [Z_{k+1}^T, \tilde{Z}_k^T] + b_\eta^T), \\ Z_k^T &= \eta_T \odot Z_{k+1}^T + (1-\eta_T) \odot \tilde{Z}_k^T, \\ \eta_F &= Sigmoid(W_\eta^F \cdot [Z_{k+1}^F, \tilde{Z}_k^F] + b_\eta^F), \\ Z_k^F &= \eta_F \odot Z_{k+1}^F + (1-\eta_F) \odot \tilde{Z}_k^F, \end{aligned} \tag{3}$$

where W_η^T, W_η^F, b_η^T, and b_η^F are trainable parameters. We obtain the gating aggregation signals η_F and η_T in the frequency domain and time domain, respectively, through a sigmoid function. We initially set Z_{K+1}^T and Z_{K+1}^F to be zero vectors.

Multi-Scale Time-Frequency Consistency The core idea of time-frequency consistency lies in fully exploiting the inherent attributes of temporal data and leveraging its potential for consistent characteristics to accomplish unsupervised learning tasks. Multi-scale global modeling helps to further enhance the information modeling and capture at various levels. In the task of emotion recognition, the time-domain information of EEG signals captures subtle variations in human emotions, such as instantaneous transitions between happiness and sadness, while the frequency-domain information reveals neural activity patterns associated with emotional processing in the human brain, such as high-frequency oscillations during anger or low-frequency fluctuations during calmness. The utilization of both time and frequency domains ensures effective invariance irrespective of the distribution characteristics inherent to time series data based on signal processing theory. (Flandrin, 1998; Papandreou-Suppappola, 2018). However, existing research relies solely on time-domain modeling and is difficult to capture the richer frequency-domain features of EEG signals. Therefore, we continue the time-frequency modeling idea of TF-C (X. Zhang, Zhao, Tsiligkaridis, & Zitnik, 2022) and further propose multi-scale time-frequency consistency.

Specifically, under the framework of contrastive learning, we regard the time-frequency representation under each scale as a positive sample pair to ensure that the time-frequency representation under the same scale is close to each other. In order to ensure that the distance between representations is measurable, we map the time-frequency representation Z_k^T and Z_k^F corresponding to the k-th scale to the time-frequency joint space through the projector R_k^T and R_k^F, and get e_k^T and e_k^F:

$$\begin{aligned} e_k^T &= R_k^T(Z_k^T), \\ e_k^F &= R_k^F(Z_k^F), \end{aligned} \tag{4}$$

where R_k^T and R_k^F are linear mappings. We adopt the contrastive learning variant of MoCo (He, Fan, Wu, Xie, & Girshick, 2020) to construct the time-frequency contrast loss.

Algorithm 1 The pre-training process of MV-TFCL.

Input: The entire training set; max epoch $Epoch$; batch size B.

Use the uniform distribution to initialize model parameters $\theta \sim U(-1,1)$.

Output: Well-trained MV-TFCL.

for $epoch = 1, 2, \cdots, Epoch$ **do**
  Recursive downsampling is performed along the time dimension on the original sequence X, resulting in K time series $\{X_1, X_2, \cdots, X_K\}$ at different scales;
  for $k = K, K-1, \cdots, 1$ **do**
    $\tilde{Z}_k^T = T\text{-}Encoder(X_k)$;
    $\tilde{Z}_k^F = F\text{-}Encoder(X_k)$;
    $Z_k^T = Gating(Z_{k+1}^T, \tilde{Z}_k^T)$;
    $Z_k^F = Gating(Z_{k+1}^F, \tilde{Z}_k^F)$;
  end for
  Calculate the time-frequency consistency contrastive loss at each scale using Eq.(5);
  $\{h_C^1, h_C^2, \cdots, h_C^{2^C}\} = TreeRL(X)$;
  Calculate the global-local contrastive loss using Eq.(7);
  Update model parameters by optimizing the loss function in Eq.(8);
end for

Negative pairs are obtained through a dynamic dictionary with a queue. We define the sets of negative sample vectors in the time domain and frequency domain under the k-th scale as $\mathcal{D}_T^k$ and $\mathcal{D}_F^k$, respectively. The time-frequency contrast loss can be expressed as:

$$\mathcal{L}_{TF} = \frac{1}{K}\sum_{k=1}^{K} -\log \frac{\exp(e_k^T \cdot e_k^F / \tau_{TF})}{\sum_{e'_k \in \mathcal{D}_F^k} \exp(e_k^T \cdot e'_k / \tau_{TF}) + \sum_{e'_k \in \mathcal{D}_T^k} \exp(e_k^F \cdot e'_k / \tau_{TF})}, \tag{5}$$

where τ_{TF} is an adjustable hyperparameter.

Tree Local Representation Learning

Inspired by the multi-head concept in Transformer, we further adopt a tree-structured time-frequency encoder with a total of C layers to extract more fine-grained local semantic information of time-frequency. In TreeRL, we define each node as a Local-Block, which consists of a Feed-Forward Neural Network (FFN) and a Frequency Enhancement Layer (FEL). The FFN is used to capture the time-domain information of each timestamp, and the FEL is used to capture the frequency-domain information. For the input $h_c^i \in \mathbb{R}^{\frac{T_x}{2^{c-1}} \times Fe}$ of the i-th node in the c-th layer, we sum up the local time-frequency features extracted by each Local-Block, split it in half along the time dimension, and transmit it to the corresponding two leaf nodes. The i-th Local-Block in the c-th layer can be represented as:

$$h_{c+1}^{2i}, h_{c+1}^{2i+1} = Split(FFN_c^i(h_c^i) + \mathcal{F}^{-1}(\mathcal{F}(h_c^i) \odot Q_c^i)), \tag{6}$$

where Q_c^i is a parameterized kernel initialized randomly, and $Split()$ represents the subsequence splitting function. Ultimately, at the C-th layer, we obtain 2^C local representations $\{h_C^1, h_C^2, \cdots, h_C^{2^C}\}$.

Global-Local Contrast Loss

Ultimately, we further construct the contrast loss from a global-local perspective. In the task of emotion recognition, EEG signals possess emotional continuity. Therefore, global and local representations have semantic associations. Similarly, we project the global and local representations into a joint space through the projectors R^G and R^L. We express the global-local contrast loss as:

$$\mathcal{L}_{GL} = \frac{1}{2^C} \sum_{i=1}^{2^C} -\log \frac{\exp(R^G([Z_1^T, Z_1^F]) \cdot R^L(h_C^i)/\tau_{GL})}{\exp(R^G([Z_1^T, Z_1^F]) \cdot R^L(h_C^i)/\tau_{GL}) + \sum_{h' \in \mathcal{D}_L} \exp(R^G([Z_1^T, Z_1^F]) \cdot R^L(h')/\tau_{GL})}, \tag{7}$$

where $\mathcal{D}_L$ represents the set of negative local representation samples randomly drawn, $[\cdot,\cdot]$ denotes the concatenation operation of two variables, and τ_{GL} is an adjustable hyperparameter. Ultimately, we define the loss function as:

$$\mathcal{L} = \mathcal{L}_{TF} + \mathcal{L}_{GL}. \tag{8}$$

Table 1: Performances on DEAP dataset.

Methods	Accuracy(%)		
	Valence	Arousal	Four
CNN-LSTM	90.82	86.13	-
CDCN	92.24	92.92	-
MMResLSTM	92.87	92.30	-
ARCNN	93.72	93.38	-
MCLFS-GAN	-	-	81.32
GANSER	93.52	94.21	89.74
SGMC	95.31	95.79	93.42
MV-TFCL	**96.19**	**96.59**	**94.82**

Experiments

To comprehensively evaluate the performance of MV-TFCL, we conducted a series of experiments and extensively compared it with state-of-the-art deep learning algorithms. In addition, we also validated the effectiveness of each component of the model through ablation experiments and representation visualization.

Settings

Datasets In our study, we aimed to validate the effectiveness of our method in the task of emotion recognition. To this end, we selected two publicly available emotion recognition datasets, namely DEAP (Koelstra et al., 2011) and SEED (Zheng & Lu, 2015; Duan, Zhu, & Lu, 2013).

The DEAP dataset is a rich repository that includes 32 channels of EEG signals and 8 channels of peripheral physiological signals from 32 subjects while they were watching emotional videos. This dataset covers 40 video trials, each of which includes 3 seconds of resting signals and 60 seconds of signals related to emotional video clips. These EEG signals were downsampled to 128 Hz and processed through a bandpass filter of 4-45 Hz. Finally, each subject rated the videos. When the rating exceeded 5.0, the corresponding EEG signals were labeled as high arousal or high valence. Otherwise, these signals were labeled as low arousal or low valence. Based on these labels, we performed binary classification tasks on the dimensions of emotion and arousal, and combined these two dimensions to perform four-class classification tasks.

The SEED dataset, on the other hand, recorded the EEG signals of 15 subjects while they were watching 15 movie videos, with emotion labels being positive, neutral, and negative. Each video trial lasted approximately 4 minutes. Each subject watched the same videos three times with an interval of more than a week. These EEG signals were recorded by 62 electrodes, sampled at 1000 Hz, then resampled to 200 Hz, and filtered within the range of 0-75 Hz.

Baselines To validate the effectiveness of our proposed MV-TFCL method, we conducted comparative experiments with the current state-of-the-art deep learning models. On the DEAP dataset, we selected four supervised models based on deep neural networks, namely CNN-LSTM (Yang, Wu, Qiu, Wang, & Chen, 2018), MMResLSTM (Ma, Tang, Zheng, & Lu, 2019), CDCN (Gao et al., 2020), and ACRNN (Tao et al., 2020). In addition, we also selected three models based on self-supervised learning, namely MCLFS-GAN (Dong & Ren, 2020), GANSER (Z. Zhang, Zhong, & Liu, 2022), and SGMC (Kan et al., 2023). On the SEED dataset, we likewise selected four supervised models, including GRSLR (Li, Zheng, Cui, Zong, & Ge, 2019), DGCNN (Song et al., 2018), BiHDM (Li et al., 2020), and ResNet18 1D kernel (Cheah et al., 2021). Furthermore, we also compared SGMC (Kan et al., 2023) as a self-supervised model.

Implementation Details We optimized our method using the Adam optimizer, with a batch size set to 64, a learning rate of 0.001, and a feature vector dimension of 1024. We set the sampling times K to 6, corresponding to the 6 TF-Blocks contained in MGRNN. Both τ_{TF} and τ_{GL} were set to 0.07. The height C of TreeRL was set to 4. All models were trained/tested on an NVIDIA Tesla V100 32G GPU.

Table 3: Ablation study of each module of MV-TFCL on the DEAP dataset.

Methods	Accuracy(%)		
	Valence	Arousal	Four
MV-TFCL$_{Multi\text{-}scale-}$	95.43	95.29	93.66
MV-TFCL$_{MGRNN-}$	92.01	92.81	91.26
MV-TFCL$_{TreeRL-}$	93.39	93.83	92.34
MV-TFCL	**96.19**	**96.59**	**94.82**

Performance Comparison

To verify the generalization ability of our model on different datasets, we conducted experiments on two publicly available EEG emotion recognition datasets, DEAP and SEED, with the results shown in Table 1 and Table 2. Our model, MV-TFCL, adopts a multi-view representation joint modeling method,

Table 2: Performances on SEED dataset.

Methods	Accuracy(%)			
Percentage of labels	1%	10%	50%	100%
GRSLR	-	-	-	87.39
DGCNN	-	-	-	90.40
BiHDM	-	-	-	93.12
ResNet18 1D kernel	-	-	-	93.43
SGMC	92.58	94.28	94.63	94.96
MV-TFCL	**92.94**	**95.37**	**95.88**	**96.03**

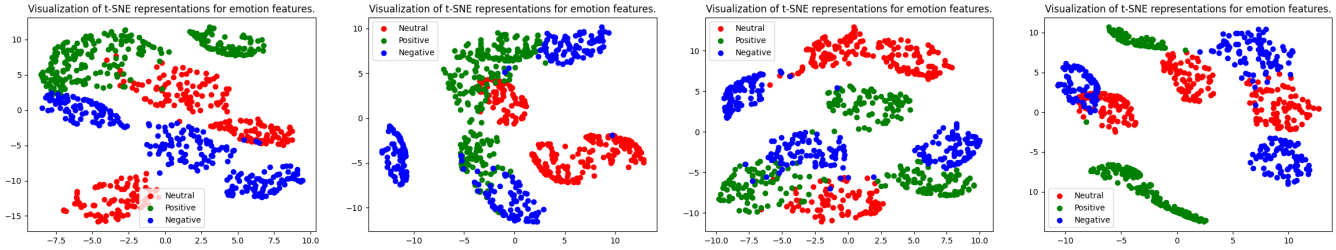

Figure 2: Visualization analysis of different emotional representations on the SEED dataset.

which can effectively extract and fuse emotional information in EEG signals. During the pre-training stage, we used a fine-tuning strategy similar to SGMC to improve the performance of emotion classification. The experimental results show that our model significantly outperforms the current state-of-the-art self-supervised pre-training model, SGMC, on all datasets and tasks. Specifically, on the DEAP dataset, our model improved the accuracy of emotion, arousal, and four-class classification tasks by 0.9%, 0.8%, and 1.5% respectively compared to SGMC. On the SEED dataset, when using all the annotated data, our model improved the accuracy by 1.1% compared to SGMC. In addition, when using part of the annotated data, our model also demonstrated strong robustness, improving the accuracy by 0.4%, 1.2%, and 1.3% respectively when using 1%, 10%, and 50% of the annotated data compared to SGMC.

Ablation Study

To investigate the contribution and role of each module in the framework to the performance of the model, we conducted three types of ablation experiments on the DEAP dataset and presented the results in Table 3. First, we simplified the overall Multi-scale Global Recurrent Neural Network module to a single TF-Block to learn the global time-frequency representation, resulting in the MV-TFCL$_{Multi-scale-}$ model. This was done to verify the importance of multi-scale modeling for the extraction of information from EEG signals. Second, we removed the MGRNN and multi-scale time-frequency consistency constraints, and used a simple linear mapping to obtain the global representation, resulting in the MV-TFCL$_{MGRNN-}$ model. This was done to assess the necessity of multi-scale time-frequency representation modeling and time-frequency consistency constraints. Finally, we replaced TreeRL with 2^C linear layers, resulting in the MV-TFCL$_{TreeRL-}$ model. This was done to verify the local representation capture capability of TreeRL. From Table 3, it can be clearly seen that each module in the framework is meaningful and effective, playing a key role in enhancing the performance of the model.

Representation Visualization

To evaluate the ability of MV-TFCL to learn representations of different emotions, we used the t-SNE algorithm to visualize the representations of three emotions (Positive, Neutral, Negative) in the SEED dataset. Figure 2 shows the distribution of EEG features of four subjects on a two-dimensional plane. From the figure, we can clearly see that MV-TFCL can effectively distinguish the representations of the three emotions, forming a clear boundary in space. At the same time, we can also find that the representations of the same emotion category have strong clustering, indicating that MV-TFCL can capture the inherent consistency of emotions.

Conclusion

This paper presents a novel framework for emotion recognition tasks based on EEG signals, termed Multi-View Time-Frequency Contrastive Learning (MV-TFCL). The proposed MV-TFCL framework everages multi-scale time-frequency features along with a global-local contrastive loss to effectively extract emotional information from EEG signals. Experimental results on two public datasets demonstrate that MV-TFCL outperforms existing algorithms in terms of accuracy and robustness in emotion recognition. In the future, we intend to apply MV-TFCL to more intricate non-stationary EEG data to further enhance its performance and generalization capabilities.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No. 62172074), Program of Introducing Talents of Discipline to Universities (Plan 111) (No. B20070)

References

Cheah, K. H., Nisar, H., Yap, V. V., Lee, C.-Y., Sinha, G., et al. (2021). Optimizing residual networks and vgg for classification of eeg signals: Identifying ideal channels for emotion recognition. *Journal of Healthcare Engineering*, *2021*.

Dong, Y., & Ren, F. (2020). Multi-reservoirs eeg signal feature sensing and recognition method based on generative adversarial networks. *Computer Communications*, *164*, 177–184.

Duan, R.-N., Zhu, J.-Y., & Lu, B.-L. (2013). Differential entropy feature for EEG-based emotion classification. In *6th international ieee/embs conference on neural engineering (ner)* (pp. 81–84).

Flandrin, P. (1998). *Time-frequency/time-scale analysis*. Academic press.

Gao, Z., Wang, X., Yang, Y., Li, Y., Ma, K., & Chen, G. (2020). A channel-fused dense convolutional network for eeg-based emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(4), 945–954.

Guo, X., Zhu, J., Zhang, L., Jin, B., & Wei, X. (2023). Adaptive bayesian meta-learning for eeg signal classification. In *2023 ieee international conference on bioinformatics and biomedicine (bibm)* (pp. 1935–1940).

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9729–9738).

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

Jia, Z., Lin, Y., Wang, J., Feng, Z., Xie, X., & Chen, C. (2021). Hetemotionnet: two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition. In *Proceedings of the 29th acm international conference on multimedia* (pp. 1047–1056).

Jia, Z., Lin, Y., Wang, J., Ning, X., He, Y., Zhou, R., ... Li-wei, H. L. (2021). Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *29*, 1977–1986.

Kan, H., Yu, J., Huang, J., Liu, Z., Wang, H., & Zhou, H. (2023). Self-supervised group meiosis contrastive learning for eeg-based emotion recognition. *Applied Intelligence*, *53*(22), 27207–27225.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., ... Patras, I. (2011). Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, *3*(1), 18–31.

LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395–2399.

Li, Y., Chen, J., Li, F., Fu, B., Wu, H., Ji, Y., ... Zheng, W. (2022). Gmss: Graph-based multi-task self-supervised learning for eeg emotion recognition. *IEEE Transactions on Affective Computing*.

Li, Y., Wang, L., Zheng, W., Zong, Y., Qi, L., Cui, Z., ... Song, T. (2020). A novel bi-hemispheric discrepancy model for eeg emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(2), 354–367.

Li, Y., Zheng, W., Cui, Z., Zong, Y., & Ge, S. (2019). Eeg emotion recognition based on graph regularized sparse linear regression. *Neural Processing Letters*, *49*, 555–571.

Li, Y., Zheng, W., Wang, L., Zong, Y., & Cui, Z. (2019). From regional to global brain: A novel hierarchical spatial-temporal neural network model for eeg emotion recognition. *IEEE Transactions on Affective Computing*, *13*(2), 568–578.

Lin, Y.-P., Wang, C.-H., Jung, T.-P., Wu, T.-L., Jeng, S.-K., Duann, J.-R., & Chen, J.-H. (2010). Eeg-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering*, *57*(7), 1798–1806.

Liu, Y., Habibnezhad, M., & Jebelli, H. (2021). Brainwave-driven human-robot collaboration in construction. *Automation in Construction*, *124*, 103556.

Liu, Y., Sourina, O., & Nguyen, M. K. (2011). Real-time eeg-based emotion recognition and its applications. *Transactions on Computational Science XII: Special Issue on Cyberworlds*, 256–277.

Ma, J., Tang, H., Zheng, W.-L., & Lu, B.-L. (2019). Emotion recognition using multimodal residual lstm network. In *Proceedings of the 27th acm international conference on multimedia* (pp. 176–183).

Mohsenvand, M. N., Izadi, M. R., & Maes, P. (2020). Contrastive representation learning for electroencephalogram classification. In *Machine learning for health* (pp. 238–253).

Papandreou-Suppappola, A. (2018). *Applications in time-frequency signal processing*. CRC press.

Sanei, S., & Chambers, J. A. (2013). *Eeg signal processing*. John Wiley & Sons.

Song, T., Zheng, W., Song, P., & Cui, Z. (2018). Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, *11*(3), 532–541.

Tao, W., Li, C., Song, R., Cheng, J., Liu, Y., Wan, F., & Chen, X. (2020). Eeg-based emotion recognition via channel-wise attention and self attention. *IEEE Transactions on Affective Computing*.

Tian, F., Zhu, L., Shi, Q., Wang, R., Zhang, L., Dong, Q., ... Hu, B. (2023). The three-lead eeg sensor: Introducing an eeg-assisted depression diagnosis system based on ant lion optimization. *IEEE Transactions on Biomedical Circuits and Systems*.

Xanthopoulos, P., Pardalos, P. M., Trafalis, T. B., Xanthopou-

los, P., Pardalos, P. M., & Trafalis, T. B. (2013). Linear discriminant analysis. *Robust data mining*, 27–33.

Yang, Y., Wu, Q., Qiu, M., Wang, Y., & Chen, X. (2018). Emotion recognition from multi-channel eeg through parallel convolutional recurrent neural network. In *2018 international joint conference on neural networks (ijcnn)* (pp. 1–7).

Zeng, H., Wu, Q., Jin, Y., Zheng, H., Li, M., Zhao, Y., ... Kong, W. (2022). Siam-gcan: a siamese graph convolutional attention network for eeg emotion recognition. *IEEE Transactions on Instrumentation and Measurement*, *71*, 1–9.

Zhang, X., Zhao, Z., Tsiligkaridis, T., & Zitnik, M. (2022). Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems*, *35*, 3988–4003.

Zhang, Z., Zhong, S.-h., & Liu, Y. (2022). Ganser: A self-supervised data augmentation framework for eeg-based emotion recognition. *IEEE Transactions on Affective Computing*.

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. doi: 10.1109/TAMD.2015.2431497

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.