EEG-Based Emotion Recognition via Convolutional Transformer with Class Confusion-Aware Attention

Jiahui Pan (panjiahui@m.scnu.edu.cn)
School of Software, South China Normal University, Guangzhou, 510631, China

Chenyu Bai (2023024312@m.scnu.edu.cn)
School of Software, South China Normal University, Guangzhou, 510631, China

Abstract

Currently, emotion recognition based on electroencephalograms (EEGs) has a wide range of applications. Although many approaches have been proposed for automatic emotion recognition with favorable performance, there are still several challenges: (1) how to sufficiently model the long- and short-term temporal feature discrepancies and redundant spatial information of EEGs and (2) how to alleviate the negative impact of the ambiguity of emotion classes. To tackle these issues, we propose the CSET-CCA, a novel framework for EEG-based emotion recognition. The feature extractor of this model combines the 1D convolutional neural network (CNN), channel Squeeze-and-Excitation (SE) module and transformer. It can extract the temporal features of EEG signals from local and global perspectives and select the critical channels in emotion recognition. Moreover, to adaptively perceive the confusion degrees of classes and increase the model's attention on confusing emotion classes, we design class confusion-aware (CCA) attention. We evaluate the CSET-CCA with the SEED and SEED-V datasets. The experimental results show that the proposed approach outperforms state-of-the-art methods.

Keywords: EEG; emotion recognition; convolutional transformer; class confusion-aware attention

Introduction

Emotion is the physiological arousal state of an individual and the cognitive state that adapts to this arousal state (Schachter & Singer, 1962). In recent years, studies on automatic emotion recognition have attracted considerable attention. Currently, there are two primary signal types used in emotion recognition tasks: overt behavioral signals and human physiological signals (Dzedzickis, Kaklauskas, & Bucinskas, 2020). Emotion, as a complex cognitive process, has been proven to be the result of the coordinated action of the cerebral cortex and subcortical nerves (Malfliet et al., 2017). Therefore, physiological signals represented by electroencephalograms (EEGs) have an inherent advantage in emotion recognition tasks.

To automate EEG-based emotion recognition, several approaches have been applied to achieve state-of-the-art performance (R. Li, Wang, & Lu, 2021; Y. Li et al., 2020). However, there are still several challenges:

(1) Differences between local and global temporal features and redundant spatial information are underutilized. An EEG signal is a type of human physiological electrical signal with high temporal resolution (Burle et al., 2015). Previous studies have also focused on the extraction of temporal features from EEGs, but these features tend to be considered from a single scale, local or global. For example, convolutional neural networks (CNNs) (Ozdemir et al., 2021) or long short-term memory (LSTM) networks (Feng et al., 2022) have been used. In emotional EEG signals, there is a strong relationship between adjacent temporal points, and a high degree of continuity (Mitchell, 2021). Additionally, there is a correlation between distant temporal points and similar neural patterns and representations (Riberto et al., 2022). Figure 1 (a) demonstrates the "short-term temporal continuity" and "long-term temporal similarity" of emotions. In addition, from a spatial perspective, when the subject is in an emotional state, distinct brain regions are activated at varying levels. Figure 1 (b) shows an example of this phenomenon. Multi-channel EEG signals often contain redundant spatial information. The different channels have varying degrees of significance according to their actual placement positions. Therefore, we capture the long- and short-term temporal relationships via a convolutional transformer and identify the critical channels using the Squeeze-and-Excitation (SE) module (Hu, Shen, & Sun, 2018).

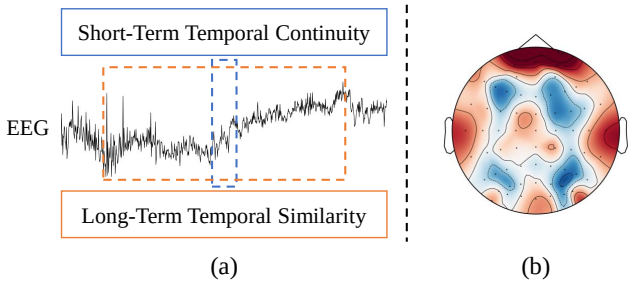

Figure 1: Spatio-temporal characteristics of emotional EEG signals. (a) Short-term temporal continuity and long-term temporal similarity of emotions. (b) An example of a topography of brain activation under positive emotion.

(2) The ambiguity of emotion classes strongly influences the model training and recognition results. Since emotion is a highly subjective personal experience, it is commonly complicated and ambiguous (Berrios, 2019). This phenomenon is also reflected in the emotion recognition task. Specifically, there are still significant variances in feature distributions within the same emotion class, but these disparities are quite minimal when compared to other classes (W. Li et al., 2021). There are several emotion classes with high neural

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.*

similarity, which often seriously influence the overall recognition accuracy. One study by R. Li et al. (2021) showed that the classification accuracy of disgust was only 71.51% in a five-classification task, with 9.95%, 8.66% and 6.13% of the samples being misidentified as happy, sad and fear, respectively. We hope that the model will focus more on emotion classes with higher confusion degrees. Therefore, we propose class confusion-aware (CCA) attention. It can intelligently perceive the degree of class confusion in a model's outputs and apply attentional weighting.

Overall, our contributions can be summarized as follows:

(1) We propose a novel emotional EEG model named CSET-CCA, which can be used to extract the long- and short-term temporal features of EEG signals comprehensively and select critical spatial information.

(2) We design the CCA attention mechanism to address the ambiguity of emotion classes. It can adaptively perceive the class confusion degree and focus more on confusing classes through attentional weighting.

(3) The experimental results show that our model achieves state-of-the-art performance, with the superior results especially on the confusing emotion classes.

Related Work

EEG-Based Emotion Recognition

Emotion recognition based on traditional machine learning requires the human extraction of EEG features and the design of classifiers. The commonly used features include power spectral density (PSD), differential entropy (DE) and rational asymmetry (RASM) (X. Li et al., 2022). In terms of choosing classifiers, support vector machine (SVM) (Zhao et al., 2019) and XGBoost (Xefteris et al., 2022), among others, are widely used. However, machine learning methods for processing raw data are limited (LeCun, Bengio, & Hinton, 2015). Researchers have further attempted to decode EEG signals using an end-to-end artificial neural network (ANN). CNN and recurrent neural network (RNN), among others, are being utilized increasingly frequently. For example, Miao et al. (2023) proposed a multiband parallel spatio-temporal 3D deep residual CNN learning framework for emotion recognition. Y. Li et al. (2020) introduced the BiHDM, which consists of four RNNs, to capture the information of each hemispheric EEG electrode from horizontal and vertical streams and achieved 93.12% accuracy on the SEED. Shen et al. (2020) combined CNNs and RNNs to extract frequency, temporal and spatial features of EEGs, and achieved 94.74% accuracy on the SEED.

However, the size of the convolutional kernel—a large kernel limits the extraction of deep information, whereas a small kernel limits the perceptual field of view—tends to be the limiting factor for CNNs (J. He et al., 2019). The RNN represented by LSTM is limited by its own network structure, which cannot realize parallel computing (Zhang, 2020). The emergence of an attention mechanism effectively alleviates the problems mentioned above. Tao et al. (2020) used a CNN incorporating channel-wise attention to extract more discriminative spatial information and explored temporal relationships via RNNs. A transformer based on a self-attention mechanism has the inherent ability to perceive global dependencies (Vaswani et al., 2017). Song et al. (2022) proposed the EEG Conformer, a convolutional transformer model used to extract EEG temporal features. However, this model ignores spatial variations in brain activation, which is also crucial for decoding EEG signals. We summarize the advantages and disadvantages of these previous works. On this basis, we design the CSET with the 1D temporal CNN, channel SE module and transformer.

CCA Attention

The minimum class confusion (MCC) loss was proposed by Y. Jin et al. (2020). This loss function is mainly applied in the process of domain adaptation to solve the problem of poor generalization performance for models trained on the source domain and tested in the target domain. The MCC loss function measures the degree of confusion between classes in the target domain and constructs the class confusion matrix from it. Next, the model optimizes the class confusion matrix in the target domain to achieve multiple domain adaptations. Inspired by this, we intend to perceive the degree of confusion for each emotion class and make specific optimizations for confusing emotion classes. However, unlike the MCC loss, the emotion recognition task is a supervised learning classification task. Therefore, after constructing the class confusion matrix, we achieve CCA attention by weighting the cross-entropy (CE) loss.

Proposed Methods

Model Architecture

An overview of the CSET-CCA is shown in Figure 2 (a). In our proposed model, the 1D CNN in the feature extractor is used to extract short-term temporal features. The Channel SE module is utilized to determine channel relevance and to extract and aggregate spatial features. The transformer is used to extract long-term temporal dependencies in EEGs. The outputs from the FC layer are subsequently inputted into the CCA Attention module.

Short-Term Temporal 1D CNN: Inspired by the work of EEGNet (Lawhern et al., 2018), we separate temporal and spatial convolutions. The raw EEG signals are fed into the model, and the 1D convolutional kernel extracts short-term features in the temporal dimension. This process involves three 1D CNN layers in total to enhance the model's capacity to capture short-term features. The number of kernels in the last convolutional layer is k, and the resulting feature is $X \in \mathbb{R}^{t \times k \times c}$, where c is the number of channels and t is the temporal feature.

Channel SE and Spatial Feature Aggregation: To extract more discriminative spatial features from EEGs, the model initially selects critical channels and assigns them higher weights through the SE module. The SE module con-

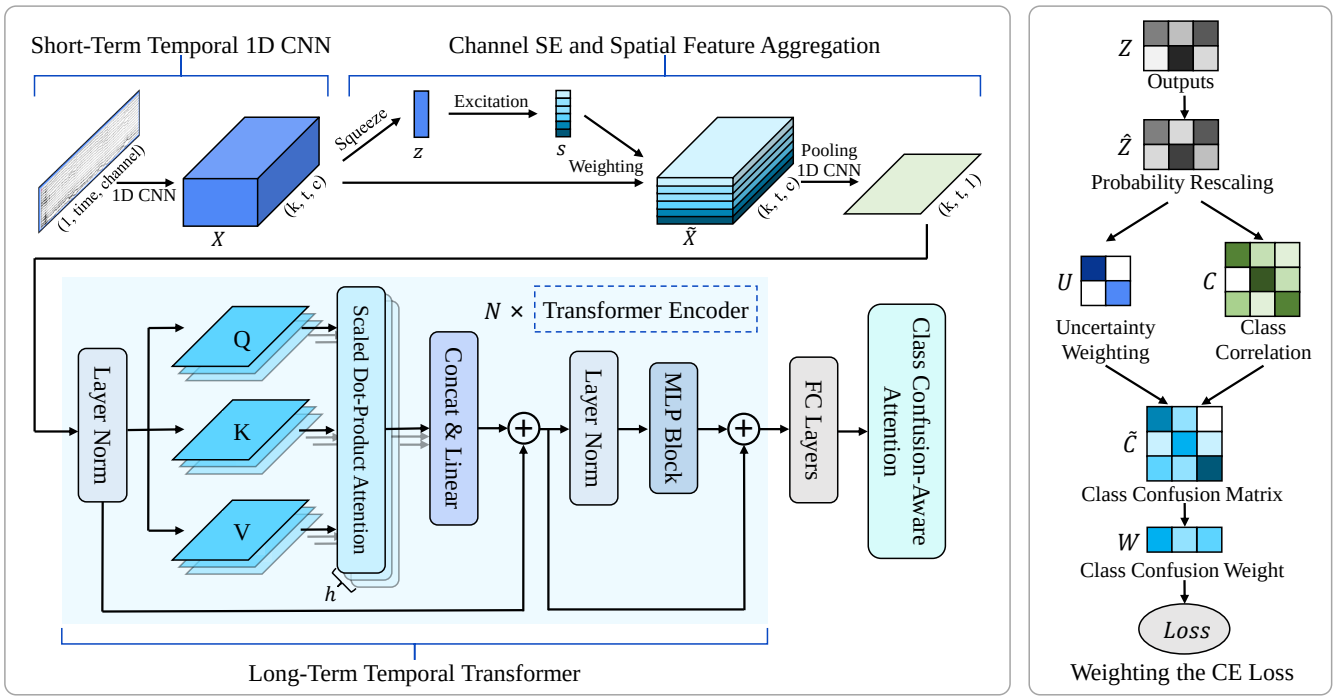

Figure 2: Overview diagram of our proposed model with a schematic diagram of the CCA attention module. (a) The model structure diagram of the CSET-CCA. There are three primary components in the feature extractor: a short-term temporal 1D CNN, a channel SE and spatial feature aggregation, and a long-term temporal transformer. In the end, the classification results are fed into the CCA attention. (b) Schematic diagram of the CCA attention module.

sists of two primary phases, squeezing and excitation. In the squeezing process, to extract the relationships between channels, features in each channel are initially squeezed into a global feature. This process is implemented by global average pooling to obtain the output $z \in \mathbb{R}^{1\times1\times c}$. The importance of each channel is predicted by two fully connected (FC) layers during the excitation process. The obtained $s \in \mathbb{R}^{1\times1\times c}$ is the weight matrix of the channels. The computation of s is shown below:

$$s = \sigma(W_2\delta(W_1 z)) \tag{1}$$

where δ and σ represent the ReLU and Sigmoid activation function, respectively, the $W_1 \in \mathbb{R}^{\frac{c}{r}\times c}$, $W_2 \in \mathbb{R}^{c\times\frac{c}{r}}$ and r is the reduction ratio. Weighting is achieved via multiplying the weight s by the original feature map X along the channel dimension. This in turn results in a feature map $\tilde{X} \in \mathbb{R}^{t\times k\times c}$ with the same dimensional size that has been weighted according to each channel's importance. To further determine the significance of the critical channels, max pooling is used in the channel dimension. Only the elements with the largest values in the pooling window are retained. Eventually, to facilitate the input of features into the transformer, spatial feature aggregation is achieved via 1D spatial convolution.

Long-Term Temporal Transformer: Due to the coherence of neural activities, global temporal dependence is also crucial for decoding EEG signals. Therefore, we input all temporal features as tokens into the transformer. The transformer consists of N transformer encoders. Initially, the model carries out a layer norm operation, which serves to ensure the stability of the sample feature distribution and avoid vanishing gradients. The multi-head mechanism can effectively improve the diversity of extracted representations, and each head is an independent representation subspace. In each head, we perform three linear transforms of feature maps, to obtain three copies of the query (Q), key (K) and value (V). In Scaled Dot-Product Attention, we use the obtained Q and K to carry out the dot product operation to calculate the similarity between every two tokens. The obtained similarity is divided by the scale factor $\sqrt{d_k}$ to avoid gradient vanishing, where d_k is the dimension of the key. After normalization by Softmax, the weight matrix is acquired. Finally, the weight matrix is multiplied by V to complete the process of weighting. The computation of the process is shown below:

$$Attention(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

After that, we need to fuse the outputs obtained from each head. The model combines the outputs of each head together, and it subsequently performs a linear transform and residual addition (K. He et al., 2016). The calculations are shown

below:

$$MultiHead(Q,K,V) = Concat(head_1,...,head_h)W^O \quad (3)$$

where, $W^O \in \mathbb{R}^{hd_v \times d_{model}}$, h represents the number of heads, d_v represents the dimensions of V, and d_{model} represents the dimension of the model outputs. To further increase the model's fitting ability, the layer norm, MLP block, and residual addition processes are added later. The operation process of one transformer encoder is as described above. This procedure needs to be repeated a total of N times.

CCA Attention

The confusion degrees of emotion classes largely influence the model training process. To adaptively perceive confusing emotion classes, we apply CCA attention to the output results of the classifier. The structure of the CCA module is shown in Figure 2 (b). This module combines the degree of uncertainty of the samples with the correlation between classes. On this basis the class confusion matrix is constructed, and the class confusion weights are determined. The final weighting to the cross-entropy loss function realizes the CCA attention.

Probability Rescaling: The output of deep neural networks (DNNs) is not a probability distribution, and DNNs tend to make overconfident predictions (Guo et al., 2017). To alleviate the negative impact of this problem when modeling the class confusion, we apply temperature rescaling. Suppose the output of the classifier is $Z_{ij} \in \mathbb{R}^{b \times n}$, where b is the batch size and n represents the number of classes. $\hat{Z}_{ij}$ to denote the probability of the j-th class being assigned to the i-th sample. Its calculation is shown below:

$$\hat{Z}_{ij} = \frac{exp(Y_{ij}/T)}{\sum_{j'=1}^{b} exp(Y_{ij'}/T)} \quad (4)$$

where T is the hyperparameter of temperature rescaling, and when $T = 1$, the above equation becomes a Softmax function.

Class Correlation: A preliminary estimate of the correlation between the j-th class and the j'-th class is shown below:

$$C_{jj'} = \hat{Z}_{\cdot j}^{\top} \hat{Z}_{\cdot j'} \quad (5)$$

Uncertainty Weighting: The confusion degrees of emotion classes are quantified differently for each sample. When the prediction results of a sample are distributed relatively evenly, the classifier is ignorant of this sample. When there are specific peaks in the prediction results of a sample, the classifier has difficulty in choosing among these confusing classes. Obviously, these samples that cause the classifier to produce cross-class ambiguity are more reflective of class confusion. We define the uncertainty of a sample with the concept of entropy in information theory. The uncertainty of the i-th sample is:

$$H(\hat{Z}_{i\cdot}) = -\sum_{j=1}^{n} \hat{Z}_{ij} log \hat{Z}_{ij} \quad (6)$$

The result of calculating the uncertainty of a sample through the entropy function is not a probability distribution. Therefore, we apply the Softmax function to the result. In turn, we construct a matrix $U \in \mathbb{R}^{b \times b}$ representing the contribution of each sample to class confusion. U is a diagonal matrix, where U_{ii} is the contribution of the i-th sample to class confusion. Its calculation is shown below:

$$U_{ii} = \frac{b(1 + exp(-H(\hat{Z}_{i\cdot})))}{\sum_{i'=1}^{b}(1 + exp(-H(\hat{Z}_{i'\cdot})))} \quad (7)$$

Class Confusion Matrix: After calculating the sample uncertainty, we can update the preliminary estimate of the class correlation matrix to class confusion matrix $C \in \mathbb{R}^{n \times n}$, where the confusion between the j-th class and the j'-th class is calculated as follows:

$$C_{jj'} = \hat{Z}_{\cdot j}^{\top} U \hat{Z}_{\cdot j'} \quad (8)$$

To ensure the stability of the class confusion degree and avoid the problem of serious class imbalance within a batch. We perform a class normalization operation. The calculation is shown below:

$$\tilde{C}_{jj'} = \frac{C_{jj'}}{\sum_{j''=1}^{n} C_{jj''}} \quad (9)$$

Weighting the CE Loss: $\tilde{C}_{jj'}$ defines the cross-class confusion between class j and class j'. Afterwards, the average confusion degree of all classes within a batch is calculated based on $\tilde{C}$. The confusion degree within a class ($j = j'$) is ignored in the calculation. The weight matrix after Softmax normalization is $W \in \mathbb{R}^{1 \times n}$. It is assumed that Y and $\hat{Y}$ are the true label and the predicted value, respectively. The loss within a batch is computed as shown below:

$$Loss = -\frac{1}{b} \sum_{i=1}^{b} \sum_{j=1}^{n} W_j Y_{ij} log(\tilde{Y}_{ij}) \quad (10)$$

Eventually, the class confusion weight matrix is weighted to the CE loss function, as shown in Equation (10). In this way, intelligent awareness and attention to confusing classes are realized.

Experiments

Datasets and Settings

We evaluate the performance of the model on the SEED (Zheng & Lu, 2015) and SEED-V (Zhao et al., 2019) datasets. Both datasets use video materials to induce the corresponding emotions of the subjects, and the number of electrode channels is 62 for both datasets. The SEED dataset records the emotional EEG data of 15 subjects. Each subject needed to participate in 3 sessions, and each session included 15 trials. The emotion classes included positive, negative and neutral. There are 16 subjects in the SEED-V dataset. Similarly, each subject included 3 sessions, each session consisted of 15 trials, and the emotion classes were happy, disgust, sad, neutral

and fear. The EEG signals were sliced into 4-second segments without overlapping windows as a way to increase the number of samples. In addition, our experiment followed a subject-dependent setting. We use the same divisions of the training and testing sets as Zheng and Lu (2015) and Zhao et al. (2019) for the SEED and SEED-V datasets, respectively.

The parameters of the model are set as follows. In the 1D CNN, the number of kernels in the last convolutional layer k is 60 (which is also the size of a token in transformer). In the channel SE, reduction ratio r is set to 0.5. In the transformer, the number of encoders N and the number of heads h are set to 6 and 10, respectively. In the CCA, we set the temperature rescaling T to 2.5. The model was implemented on the PyTorch framework and trained using a Tesla T4 GPU. We trained the model using Adam optimizer, with the learning rate and batch size of 0.0001 and 256, respectively.

Baselines

The baselines used for comparison are shown below.

(1) **SVM** (Zheng & Lu, 2015; Zhao et al., 2019): Machine learning method using DE features with the SVM classifier. (2) **BDAE** (Zhao et al., 2019): Extraction of high-level representations for emotion recognition via bimodal deep autoencoder. (3) **R2G-STNN** (Y. Li et al., 2019): Extracting spatial relationships within and between brain regions and dynamic temporal information, respectively. (4) **BiHDM** (Y. Li et al., 2020): This method explores the discrepancy between left and right hemisphere features, and capture temporal information from two directions via the RNNs. (5) **RGNN** (Zhong, Wang, & Miao, 2020): The regularized GNN explores the topology of EEG signal channels using two regularizers. (6) **4D-CRNN** (Shen et al., 2020): A model combining a CNN and an RNN is used to extract spatial, spectral and temporal domain features of EEG signals. (7) **MD-AGCN** (R. Li et al., 2021): This model is an adaptive GCN model that fuses frequency domain features and temporal domain features. (8) **PGCN** (M. Jin et al., 2023): A graph convolution model that incorporates local, mesoscopic and global features at different scales.

Classification Results

We compare the proposed CSET-CCA model with baselines on the SEED and SEED-V datasets.

Table 1 shows the accuracy (ACC) and standard deviation (STD) of these models on the SEED and SEED-V datasets. Compared with that of the baseline models, the ACC of our model is further improved. CSET-CCA achieves state-of-the-art performance. On the SEED dataset, an ACC of 95.18% and an STD of 5.35% are achieved. The CSET-CCA simultaneously considers long- and short-term temporal features, spatial features and the degrees of class confusion, which enables our model to adequately capture valuable information in EEG signals. CSET-CCA still achieves the best result on the SEED-V dataset, with an ACC of 82.06% and an STD of 8.42%. The similarity between emotion classes in the five-classification task is further improved. However, CCA attention can adaptively perceive the confusion degrees of classes and pay more attention to confusing emotion classes, which further improves the performance of the model.

Table 1: Comparison of classification performance on the SEED and SEED-V datasets.

Models	SEED ACC ± STD (%)	SEED-V ACC ± STD (%)
SVM	83.99 ± 9.72	69.50 ± 10.28
BDAE	-	79.70 ± 4.76
R2G-STNN	93.38 ± 5.96	-
BiHDM	93.12 ± 6.06	-
RGNN	94.24 ± 5.95	-
4D-CRNN	94.74 ± 2.32	-
MD-AGCN	94.81 ± 4.52	80.77 ± 6.61
PGCN	-	81.69 ± 10.57
CSET-CCA	**95.18 ± 5.35**	**82.06 ± 8.42**

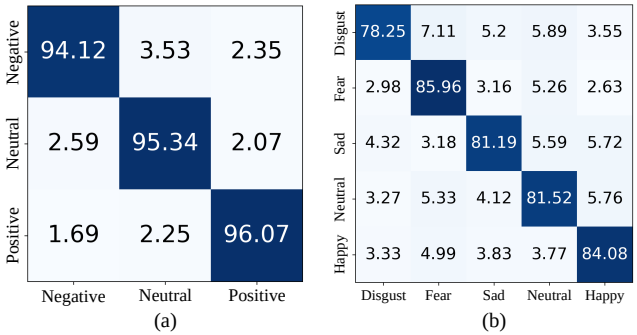

Figure 3: Confusion matrices of the CSET-CCA. Each column represents the predicted classes that our model outputs and each row represents the true classes. (a) Confusion matrix on the SEED. (b) Confusion matrix on the SEED-V.

Figure 3 shows the confusion matrix of the model on these two datasets. On the SEED dataset, our model outperforms the recognition of positive and neutral emotions (96.07% and 95.34%, respectively) than that of negative (94.12%). The proportions of samples that are misclassified as negative or neutral are relatively high, at 3.53% and 2.59%, respectively. For the SEED-V dataset, our model has a relatively low recognition accuracy of 78.25% for the disgust emotion, but this accuracy is still higher than that of the baseline model MD-AGCN (71.51%). This result is also superior considering that disgust is a confusing emotion class and that only 7.11% and 3.55% of samples are misidentified as fear and happy, respectively. Additionally, the accuracies of the other classes still remain high (85.96% for fear, 81.19% for sad, 81.52% for neutral and 84.08% for happy). This is attributed to the strong representation ability of the CSET and the adaptive perception of confusing classes by CCA attention.

Table 2: Ablation study on the SEED dataset.

Models	ACC±STD(%)	Variations(%)
1D CNN-removed	94.73 ± 4.96	-0.45
channel SE-removed	93.81 ± 5.76	-1.37
transformer-removed	86.59 ± 6.22	-8.59
CCA-removed	93.64 ± 4.89	-1.54
CSET-CCA	**95.18 ± 5.35**	**0.00**

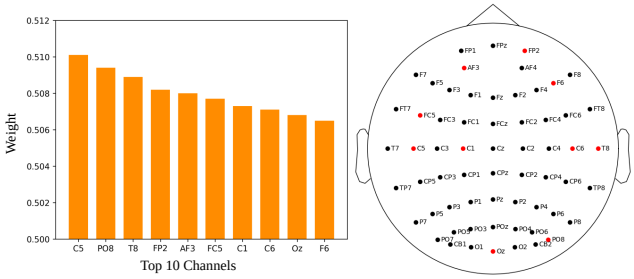

Figure 4: Top 10 channels (62 in total) selected by the SE module and the topological map of the distribution of these channels (in red).

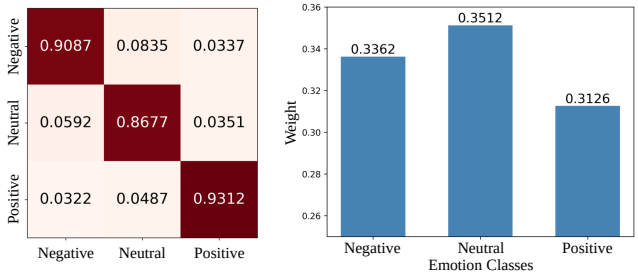

Figure 5: Average class confusion matrix and class confusion weights constructed on the SEED dataset.

Ablation Study

We verify the contribution of each module through an ablation study. The experiments are based on the SEED dataset. The results are shown in Table 2.

(1) When the 1D CNN is removed, the model's ability to extract short-term temporal features decreases. However, since the captured long-term temporal dependencies also contain short-term features, the accuracy decreases by only 0.45% and can still reach 94.73%.

(2) When the channel SE module is removed, the model treats all channels indiscriminately and the spatial information is underutilized, with an accuracy of 93.81%, and a decrease of 1.37%.

(3) When the transformer is removed, the model is constrained by the limited perceptual field of the CNNs, which is unable to effectively model long-term temporal relationships, with an accuracy of only 86.59%, and a decrease of 8.59%.

(4) When the CCA attention is removed, the model's recognition accuracy for the confusing emotion classes decreases, with an overall accuracy of 93.64%, a decrease of 1.54%.

In summary, the long-term temporal dependence extracted by the transformer contributes the most to the model. This is mainly due to the high temporal resolution of EEG signals and the powerful global perception of the transformer. The CCA attention and channel SE also favorably contribute to the final results.

Visualization

To identify the critical channels for emotion recognition, we calculate the average weights learned by the channel SE module. Then, all of the channel weights are sorted, and the ten channels—C5, PO8, T8, FP2, AF3, FC5, C1, C6, Oz and F6—with the highest weights are chosen. Fig 4 shows the top 10 channel weights and the topological map of their electrode distributions. Most of these channels are concentrated on the frontal, temporal and occipital lobes of the brain. According to the study by Phan et al. (2002), emotions mainly activate prefrontal and temporal lobe sites, and evoked materials based on visual stimuli activate the occipital cortex. Our experimental results are generally consistent with neuroscience studies. Therefore, the channel SE module effectively selects the critical channels in emotion recognition.

In addition, to verify the effectiveness of the CCA attention, we also visualize the average class confusion matrix and class confusion weights constructed on the SEED dataset, as shown in Figure 5. The nondiagonal elements in the class confusion matrix represent cross-class confusion, and it can be seen that there is a relatively high degree of class confusion between negative and neutral emotions. For the corresponding class confusion weights, neutral is assigned the highest confusion weight of 0.3512. With a value of 0.3362, negative emotion is the second highest. The lowest value for positive emotion is 0.3126. This result suggests that the model highlights the importance of neutral emotion during the training process. In conclusion, CCA attention is effective in perceiving the confusion degrees of classes and making the model more attentive to confusing emotion classes.

Conclusion

In this paper, we propose the CSET-CCA, a novel emotion recognition model. It can effectively extract long- and short-term temporal features and discriminative spatial information from EEG signals. It also takes the ambiguity of emotion classes into consideration. The CCA attention mechanism achieves adaptive perception and weighting of confusing emotion classes. The experimental results show that our model can achieve the best classification performance. An ablation experiment and visualization also validate the effectiveness of each module. Due to the limitations of the current work, our model does not perform well enough in cross-subject scenarios. In our future work, we will further attempt to optimize our model and improve its generalization performance to make it more valuable in practical applications.

Acknowledgments

This work was supported by the STI 2030-Major Projects under grant 2022ZD0208900, the Guangdong Basic and Applied Basic Research Foundation under grant 2024A1515010524, and the Special Innovation Projects of Colleges and Universities in Guangdong Province under grant 2022KTSCX035.

References

Berrios, R. (2019). What is complex/emotional about emotional complexity? *Frontiers in Psychology*, *10*, 1606.

Burle, B., Spieser, L., Roger, C., Casini, L., Hasbroucq, T., & Vidal, F. (2015). Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view. *International Journal of Psychophysiology*, *97*(3), 210–220.

Cui, H., Liu, A., Zhang, X., Chen, X., Wang, K., & Chen, X. (2020). EEG-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network. *Knowledge-Based Systems*, *205*, 106243.

Dzedzickis, A., Kaklauskas, A., & Bucinskas, V. (2020). Human emotion recognition: Review of sensors and methods. *Sensors*, *20*(3), 592.

Feng, L., Cheng, C., Zhao, M., Deng, H., & Zhang, Y. (2022). EEG-based emotion recognition using spatial-temporal graph convolutional LSTM with attention mechanism. *IEEE Journal of Biomedical and Health Informatics*, *26*(11), 5406–5417.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330).

He, J., Zhao, L., Yang, H., Zhang, M., & Li, W. (2019). HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(1), 165–178.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7132–7141).

Jin, M., Zhu, E., Du, C., He, H., & Li, J. (2023). PGCN: Pyramidal graph convolutional network for EEG emotion recognition. *arXiv preprint arXiv:2302.02520*.

Jin, Y., Wang, X., Long, M., & Wang, J. (2020). Minimum class confusion for versatile domain adaptation. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)* (pp. 464–480).

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, *15*(5), 056013.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

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, W., Huan, W., Hou, B., Tian, Y., Zhang, Z., & Song, A. (2021). Can emotion be transferred?—A review on transfer learning for EEG-based emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(3), 833–846.

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., 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., 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.

Malfliet, A., Coppieters, I., Van Wilgen, P., Kregel, J., De Pauw, R., Dolphens, M., & Ickmans, K. (2017). Brain changes associated with cognitive and emotional factors in chronic pain: A systematic review. *European Journal of Pain*, *21*(5), 769–786.

Miao, M., Zheng, L., Xu, B., Yang, Z., & Hu, W. (2023). A multiple frequency bands parallel spatial–temporal 3D deep residual learning framework for EEG-based emotion recognition. *Biomedical Signal Processing and Control*, *79*, 104141.

Mitchell, J. (2021). Affective shifts: Mood, emotion and well-being. *Synthese*, *199*(5-6), 11793–11820.

Ozdemir, M. A., Degirmenci, M., Izci, E., & Akan, A. (2021). EEG-based emotion recognition with deep convolutional neural networks. *Biomedical Engineering/Biomedizinische Technik*, *66*(1), 43–57.

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage*, *16*(2), 331–348.

Riberto, M., Paz, R., Pobric, G., & Talmi, D. (2022). The neural representations of emotional experiences are more similar than those of neutral experiences. *Journal of Neuroscience*, *42*(13), 2772–2785.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*(5), 379.

Shen, F., Dai, G., Lin, G., Zhang, J., Kong, W., & Zeng, H. (2020). EEG-based emotion recognition using 4D convolutional recurrent neural network. *Cognitive Neurodynamics*,

14, 815–828.

Song, Y., Zheng, Q., Liu, B., & Gao, X. (2022). EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 710–719.

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*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Xefteris, V., Tsanousa, A., Georgakopoulou, N., Diplaris, S., Vrochidis, S., & Kompatsiaris, I. (2022). Graph theoretical analysis of EEG functional connectivity patterns and fusion with physiological signals for emotion recognition. *Sensors*, *22*(21), 8198.

Zhang, N. (2020). Learning adversarial transformer for symbolic music generation. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhao, L., Li, R., Zheng, W., & Lu, B. (2019). Classification of five emotions from EEG and eye movement signals: Complementary representation properties. In *Proceedings of the 9th International IEEE/EMBS Conference on Neural Engineering (NER)* (pp. 611–614).

Zheng, W., & Lu, B. (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.