

Article

MCF-SCA: A Multi-Scale Spatio-Temporal Convolution and Multi-Order Gated Spatial-Channel Aggregation Networks for Cross-Subject EEG-Based Emotion Recognition

Yinghui Meng , Jiaoshuai Song, Duan Li , Jiaofen Nan, Wen Feng, Yongquan Xia, Fubao Zhu 
and Changxiang Yuan

School of Computing and Artificial Intelligence, Zhengzhou University of Light Industry, Zhengzhou 450001, China; yinghuimeng@zzuli.edu.cn (Y.M.); 332305080534@zzuli.edu.cn (J.S.); nanjiaofen@zzuli.edu.cn (J.N.); fengwen@zzuli.edu.cn (W.F.); xiayongquan@zzuli.edu.cn (Y.X.); fbzhu@zzuli.edu.cn (F.Z.); 332305090555@zzuli.edu.cn (C.Y.)

* Correspondence: liduan@hpu.edu.cn

Abstract

Cross-subject emotion recognition using EEG remains challenging due to substantial inter-individual variability. To address this, we propose a Multi-scale Spatio-Temporal Convolution and Multi-order Gated Spatial-Channel Aggregation Network (MCF-SCA). The model leverages multi-scale spatio-temporal convolution to capture rich temporal and spatial features and applies Fast Fourier Transform to transform EEG signals into the frequency domain, enhancing emotion-related representations. A multi-order spatial-channel aggregation module is then introduced, which adaptively integrates features across spatial and channel dimensions through a gating mechanism, enabling dynamic feature weighting and more expressive emotional representations. Experiments on the DEAP dataset show accuracy gains of up to 11–30% for arousal and 12–31% for valence compared with TSception, CNN, LSTM, EEGNet, and MLP. On the DREAMER dataset, improvements reach 5–33% and 3.7–34%, respectively. These results confirm that MCF-SCA achieves superior accuracy and cross-subject adaptability, providing strong support for emotion-based brain-computer interface applications.

Keywords: EEG; emotion recognition; multi-scale spatio-temporal convolution; multi-order spatial-channel aggregation



Academic Editor: Ognjen Arandjelović

Received: 19 January 2026

Revised: 10 February 2026

Accepted: 3 March 2026

Published: 5 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Emotion serves as an umbrella term encompassing various mental states triggered by changes in an individual's environment, which can lead to behavioral and cognitive alterations [1]. A deeper understanding of human emotional mechanisms will propel the development of AI models exhibiting human-like attributes [2]. Existing research indicates that investigations into emotional mechanisms primarily depend on two categories of data: non-physiological signals and physiological signals. Typical non-physiological signals include facial expressions [3], speech [4], text and video. In the field of affective computing, emotion recognition methods based on facial expressions, speech, and multimodal information have achieved remarkable progress [5,6]. However, in real-world scenarios, such approaches remain susceptible to subjective emotional masking, environmental noise, and pronounced inter-individual variability. In contrast, physiological signals encompass electroencephalogram (EEG), electrocardiogram (ECG), and electromyogram (EMG), among

others [7,8]. Compared with non-physiological modalities such as facial expressions, textual content, and video data, physiological signals are inherently more difficult to disguise and therefore provide greater objectivity. Notably, EEG data, which directly reflects the brain's internal neural activity, serves as one of the most objective indicators for assessing human emotional states [9]. Consequently, EEG-based emotion recognition has attracted growing research interest in recent years.

Currently, emotion recognition based on EEG can be approached in two ways: within-subject and cross-subject [10]. Within-subject emotion recognition typically trains a model using a subset of data from the same subject and predicts the remaining data from that subject, focusing on modeling intra-individual patterns. In contrast, cross-subject emotion recognition uses data from a group of subjects for training and evaluates the model on unseen subjects during testing, emphasizing the robustness and transferability of the model to inter-individual variability. In this study, we focus on the cross-subject setting, as it more closely reflects practical application scenarios where labeled data from new subjects are unavailable. Research has shown that cross-subject models generally perform worse than within-subject models. For example, Kim et al. [11] used linear discriminant analysis with bimodal data fusion, and in their cross-subject experiments, the average accuracy was only 55%, much lower than the 92% accuracy in within-subject tests. Similarly, Liu et al. [12] proposed a dynamic graph attention network for emotion recognition, and while subject-dependent accuracies on the SEED [13], SEED-IV [14], and DEAP [15] datasets were 99.86%, 99.53%, and 94.69%, respectively, these dropped to 80.02%, 70.22%, and 64.76% in cross-subject experiments. SEED and SEED-IV are EEG-based emotion datasets collected under a video-based emotion elicitation paradigm, where subjects watch emotionally evocative film clips to induce affective states; SEED focuses on three discrete emotion categories, whereas SEED-IV extends this setting to four discrete emotions. DEAP is a multimodal affective dataset in which participants are exposed to music video stimuli and provide self-assessment ratings along the arousal and valence dimensions, making it suitable for dimensional emotion recognition. Similarly, the DREAMER [16] (Dataset for Emotion Analysis using EEG and Peripheral Physiological Signals) dataset employs emotion-eliciting video stimuli and records EEG together with peripheral physiological signals, with emotion labels also defined in terms of arousal and valence. Li et al. [8] introduced a trainable adjacency relationship-driven graph convolutional network, and in subject-dependent experiments on the DEAP dataset, the accuracies for arousal and valence were 85.6% and 89.2%, respectively, but dropped to 68.9% and 71.3% in cross-subject tests. Wang et al. [17] proposed an end-to-end spatiotemporal model based on raw EEG signals, which achieved 97.15% and 96.33% accuracy in subject-dependent experiments on the DREAMER dataset, but the accuracy dropped to 81.22% and 81.83% in cross-subject experiments. Wu et al. [18] proposed a hybrid architecture combining CNN and GRU. In experiments on the DEAP and DREAMER datasets, the subject-dependent accuracies reached 99.51%, 99.60%, and 99.59% (98.63%, 98.7% and 98.71%), while the cross-subject accuracies were 58.67%, 65.74% and 67.05% (75.65%, 75.89%, and 71.71%), respectively. Despite the rapid progress of deep learning-based EEG emotion recognition, cross-subject emotion decoding remains fundamentally unsolved. Unlike subject-dependent settings, cross-subject EEG analysis suffers from severe distribution mismatch caused by inter-individual neurophysiological variability, including differences in cortical functional organization, baseline neural activity, and frequency-dependent emotional responses. These discrepancies lead to significant representation shifts across subjects, making features learned from one group difficult to generalize to unseen individuals. Existing deep models often implicitly assume that emotional EEG patterns are shared across subjects at the feature level; however, empirical evidence suggests that emotional neural signatures exhibit strong subject-specific modulations in both

temporal dynamics and spatial distributions. As a result, even high-capacity deep networks tend to overfit subject-specific characteristics, leading to a substantial performance degradation in cross-subject scenarios. Therefore, improving cross-subject generalization requires not only deeper architectures, but also principled feature representations that are robust to inter-subject variability.

In recent years, several novel approaches have been proposed to improve cross-subject emotion decoding models. For instance, Samavat et al. [19] introduced a multi-input deep model combining CNNs and LSTMs, applying adaptive regularization to enhance the model's generalization across subjects. Behrooz Zali-Vargahan et al. [20] proposed a cross-subject emotion recognition method based on EEG signals, validating the effectiveness of CNN feature extraction with various classifiers. Li et al. [21] introduced the Dynamic Stream Selection Network (DSSN), a method specifically designed to address the inherent challenges in cross-subject EEG emotion recognition. Arjun et al. [22] developed an unsupervised LSTM with a channeled attentional autoencoder to obtain subject-invariant latent vector subspaces. They combined this with a CNN using an attentional framework for cross-subject emotion recognition, achieving accuracies of 65.93% and 69.51%. Pandey et al. [23] developed a cross-subject EEG-based emotion recognition model achieving accuracies of 75.93% and 81.48% for arousal and valence, respectively, on the DREAMER dataset. Ding et al. [24], inspired by GoogleNet's initial block [25], proposed TSception, which extracts time-frequency features effectively using dynamic time and asymmetric space layers. TSception achieved classification results of 61.57% and 60.61% on the DEAP and MAHNOB-HCI [26] (Multimodal Database for Human–Computer Interaction) datasets, respectively. Later, Han, L et al. [27] combined multi-scale convolution with the TimesNet network [28]. They achieved impressive results in both subject-dependent and cross-subject experiments on the DEAP dataset. These studies, which build on CNN improvements, introduced new methods for cross-subject emotion decoding. However, the cross-subject decoding accuracy and generalization ability of these models still need further enhancement.

Compared to traditional CNNs, multi-scale convolution has shown strong performance in both the temporal and spatial domains of EEG signals. However, existing multi-scale convolution-based methods often lack sufficient flexibility and adaptability, limiting their ability to extract deep features across multiple dimensions. The Multi-Order Gated Aggregation Network (MOGANET) model [29], through multi-order feature aggregation and context information extraction, overcame this challenge and enhanced feature representation. Given the multi-scale nature of EEG signals, MOGANET provides an adaptive context aggregation mechanism to precisely extract deep emotion features. Building on this, the current study proposes a multi-order interaction mechanism to improve both the accuracy and generalizability of emotion decoding from EEG signals. This approach leverages feature mining across different dimensions and levels and dynamically adjusts the feature aggregation pattern through a gating mechanism to mitigate the offset of features at different levels due to individual differences.

The main contributions of this work are summarized as follows:

(1) A new cross-subject EEG emotion recognition method, called Multi-scale Spatio-Temporal Convolution and Multi-Order Gated Spatial-Channel Aggregation Networks (MCF-SCA), is proposed. Unlike traditional emotion recognition techniques, our method integrates local and global feature representations, effectively minimizing subject-specific differences and boosting the accuracy of cross-subject EEG emotion decoding.

(2) We propose a multi-order gated spatial-channel aggregation module that extracts rich temporal and spatial features from EEG signals of different individuals using multi-scale convolutions. Meanwhile, FFT is utilized to convert the time-domain signals into frequency-domain signals for deeper feature extraction. After passing through the spatial-

channel aggregation (SCA) module, low-order, mid-order, and high-order features are extracted, respectively, and the gated mechanism is employed to dynamically adjust the weights of each layer's features, thereby mining the most expressive emotional features.

(3) Extensive experiments on the DEAP and DREAMER datasets confirm the effectiveness and cross-subject generalization of MCF-SCA. Its average classification accuracy and F1-score for arousal and valence classification are significantly higher than those of most existing methods. This significantly demonstrates the robustness and excellent cross-subject adaptability of our model in cross-subject scenarios.

2. Materials and Methods

This section provides a detailed description of the overall architecture of the proposed MCF-SCA, as shown in Figure 1. The architecture consists of the Pre-processing module, the MCF-SCA module, and the final classifier module. The following sections describe the functions of each module in detail.

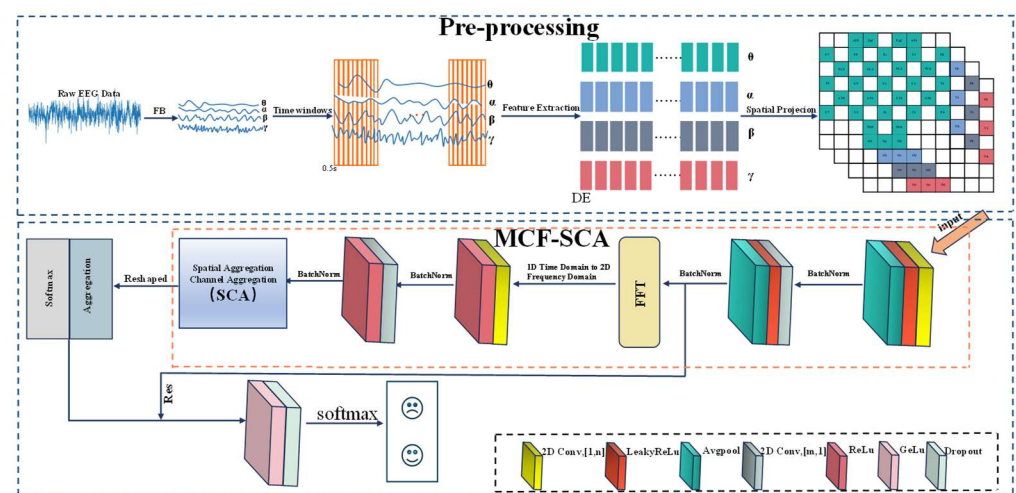


Figure 1. The architecture of the MCF-SCA.

2.1. Pre-Processing

As shown in Figure 1, for the DEAP dataset, we first apply a band-pass filter to extract EEG signals in the α , β , γ , and θ bands. To address the large number of trainable parameters and the small number of trials in both DEAP and DREAMER, the sliding window length is set to 0.5 s with no overlap to balance temporal resolution and signal stationarity. Previous EEG studies have shown that windows shorter than 0.25 s may fail to capture stable spectral characteristics, while excessively long windows tend to smooth transient emotional responses and introduce redundant temporal information. A 0.5 s window provides sufficient frequency resolution for emotion-related EEG rhythms while maintaining quasi-stationary signal properties, which is particularly important for reliable feature extraction in cross-subject emotion recognition tasks. For each frequency band, we compute Differential Entropy [30] by approximating the differential entropy using a Gaussian distribution, as shown in the formula below:

$$\begin{aligned}
 DE &= - \int_a^b f(x) \log(f(x)) dx \\
 &= - \int_a^b \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right) dx \\
 &= \frac{1}{2} \log(2\pi e \sigma^2)
 \end{aligned} \tag{1}$$

where x is the EEG signal, μ represents the mean value of the signal within the time interval, σ is the standard deviation of the signal amplitude within this frequency band, and a and b define the value range of the signal.

After computing DE for each time window, baseline DE values calculated from the pre-stimulus segment are subtracted to emphasize emotion-related relative changes. These features are then mapped onto a 2D matrix representing the spatial distribution of the electrodes, resulting in higher-dimensional features. The two-dimensional electrode mapping does not aim to reconstruct the exact cortical geometry or model volume conduction effects. Instead, it provides a topology-preserving approximation of electrode neighborhood relationships to facilitate spatial convolution. By projecting electrodes according to their relative scalp positions, the proposed mapping enables convolutional kernels to exploit local spatial correlations while maintaining consistency with the physical layout of EEG sensors. This representation is particularly advantageous for convolution-based architectures, as it allows the network to capture spatially adjacent activation patterns that are difficult to model using one-dimensional channel-wise representations. The mapping for the DEAP dataset (32 electrodes) and the DREAMER dataset (14 electrodes) is shown in Figure 2.

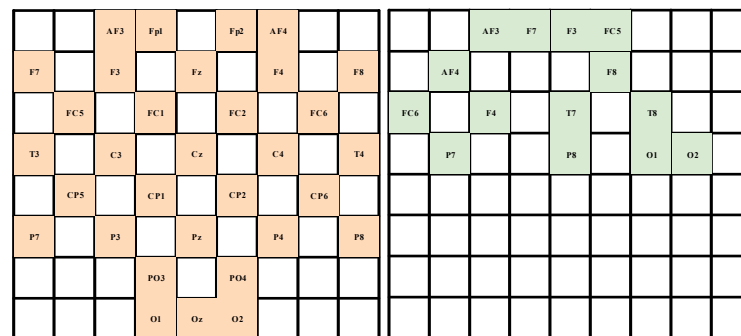


Figure 2. DEAP 32 electrode mapping matrix and DREAMER 14 electrode mapping matrix.

2.2. MCF

The input data is the preprocessed DE feature matrix with dimensions $N \times 1 \times C \times L$, where N is the number of samples, C is the 4 frequency bands, and L is the 2D matrix that we mapped out to be 8×9 (i.e., 72).

This layer aims to capture dynamic temporal and spatial patterns from the mapped DE features, allowing the fine-grained temporal variations in EEG signals to be represented. The core idea is multi-scale parallel convolution, which introduces convolution kernels with different receptive fields in the time dimension, specifically $(1, n)$ and $(m, 1)$, where n and m vary depending on the position. The input EEG features are represented as $X = [X_1, X_2, \dots, X_t], X_t \in R^{c \times l}$, where t is the number of DE feature samples, c is the number of channels, and l is the length of each feature. The features are initially convolved with a 2D kernel of size $(1, n)$, where $n \in (4, 8, 16)$, and the stride is set to $(1, 1)$, extracting features in the temporal dimension. After the multi-scale temporal convolution, the LeakyReLU activation function is applied, followed by average pooling to downsample the feature map, reducing spatial resolution and mitigating noise. This step controls model complexity while extracting deeper features. Let $Z_{temporal}^i$ represent the output of the i -th multi-scale temporal convolution, where $Z_{temporal}^i \in R^{N \times M \times c \times l}$. Here, N is the number of samples, M is the number of output channels of the convolution, c is the number of channels, and l is the convolution length. It is defined as follows:

$$Z_{temporal}^i = AvgPool(\Phi_{L-ReLU}(Conv2D(X))) \quad (2)$$

where X is the input sample, $Conv2D(\cdot)$ represents a 2D convolution along the temporal dimension with kernel size $(1, n)$, $\Phi_{L-ReLU}(\cdot)$ is the LeakyReLU activation function, and $AvgPool(\cdot)$ denotes an average pooling. After multi-scale temporal processing, the feature maps are concatenated along both the time and channel dimensions to enrich the representation. To ensure stable input distribution during training, batch normalization [31] is applied after the dynamic temporal layer. The output of the multi-scale temporal layer is $Z_B^{(1)}, Z_B^{(1)} \in R^{N \times M \times c \times l}$, defined as:

$$Z_B^{(1)} = BN\left(\left[Z_{temporal}^1, \dots, Z_{temporal}^i\right]\right), i \in [1, 2, 3] \quad (3)$$

where $BN(\cdot)$ denotes the batch normalization operation, which normalizes the input features using batch-wise statistics and includes learnable scaling and shifting parameters to stabilize and accelerate network training, and $[Z_{temporal}^i]$ denotes that it is spliced over the output feature dimension of the i -th convolution. The processed features are then passed into the spatial convolution module to further explore the data's intrinsic structure. For the multi-scale spatial convolution module, we use 2D convolution kernels of size $m \times 1$, where m is either 4 or 2, and the stride is set to 1×1 . The $(4, 1)$ kernel captures global, large-scale features, while the $(2, 1)$ kernel captures local, fine-grained features. Let $Z_{spatial}^i$ represent the output of the i -th multi-scale spatial convolution layer, where $Z_{spatial}^j \in R^{N \times M \times c \times l}$. The specific definition:

$$Z_{spatial}^j = AvgPool\left(\Phi_{L-ReLU}\left(Conv2D\left(Z_B^{(1)}\right)\right)\right) \quad (4)$$

where $Z_B^{(1)}$ is the normalized output from the previous layer, $Conv2D(\cdot)$ represents the 2D convolution with kernel size of $(m, 1)$. Each output is concatenated along the channel dimension. Then, similar to the multi-scale time configuration, normalization is applied. The output is:

$$Z_B^{(2)} = BN\left(\left[Z_{spatial}^1, \dots, Z_{spatial}^j\right]\right), j \in [1, 2, 3] \quad (5)$$

where $BN(\cdot)$ is the batch normalization operation, and $[Z_{spatial}^j]$ represents that it is spliced on the output feature dimension of the j -th convolution. To capture the frequency-domain features of the EEG signal, we apply the FFT to convert the one-dimensional time series into the frequency domain, allowing us to obtain the signal's feature representation in both the frequency and spatial domains. First, we transform the output $Z_B^{(2)}$ from the previous layer into a new shape $Z_B^{(2)} \in R^{N \times c \times l}$, where N is the number of samples, c is the number of frequency bands after the spatial convolution, and l is the number of features. After applying the FFT, we extract information like period changes and signal amplitude variations. If the time series length is not a multiple of the period, we apply padding to make it a multiple. The padded data is then reorganized into a two-dimensional tensor for the subsequent convolution operation. The specific definition is:

$$Z_F^i = Reshape_{q_i, r_i}\left(Padding\left(FFT\left(Z_B^{(2)}\right)\right)\right), i \in \{1, \dots, k\} \quad (6)$$

where $FFT(\cdot)$ denotes the Fast Fourier Transform, the primary role of the FFT module is to automatically identify latent dominant periodic patterns in the signal and to provide a principled basis for structural reorganization prior to subsequent two-dimensional convolution operations. Specifically, a one-dimensional real-valued fast Fourier transform (rFFT) is applied solely along the temporal dimension to reduce computational complexity. Based on the resulting amplitude spectrum, the top $k = 2$ dominant frequencies are used to infer candidate periods used for reshaping. $Padding(\cdot)$ refers to zero-padding along

the time dimension, and $Reshape_{q_i, r_i}(\cdot)$ is used to address the issue of directly applying a 2D transformation. Here, q_i, r_i represent the rows and columns into which the data is expanded, respectively.

Through these operations, we convert the 1D time dimension into a 2D frequency dimension, while simultaneously obtaining a set of 2D frequency-domain tensors $\{Z_F^1, \dots, Z_F^k\}$. These processed 2D tensors are then passed into the multi-scale convolutional layer, which can be defined as:

$$\bar{Z}_F^i = MCNN(Z_F^i), i \in \{1, \dots, k\} \quad (7)$$

where Z_F^i represents the i -th 2D feature map, and $MCNN(\cdot)$ denotes the multi-scale convolution. This method is similar to the previously discussed multi-scale temporal and spatial convolution operations, with the key difference being the variation in the size of the convolution kernel.

2.3. SCA

The Spatial-Channel Aggregation (SCA) layer is designed for adaptive multi-domain feature extraction and consists of three main components: the embedding stem (Stem), the spatial aggregation module (SA), and the channel aggregation module (CA). By jointly exploiting temporal, frequency, and spatial characteristics through multi-order feature modeling, the SCA layer enhances the network's capability to capture both local details and global contextual information. The overall architecture of the SCA layer is illustrated in Figure 3.

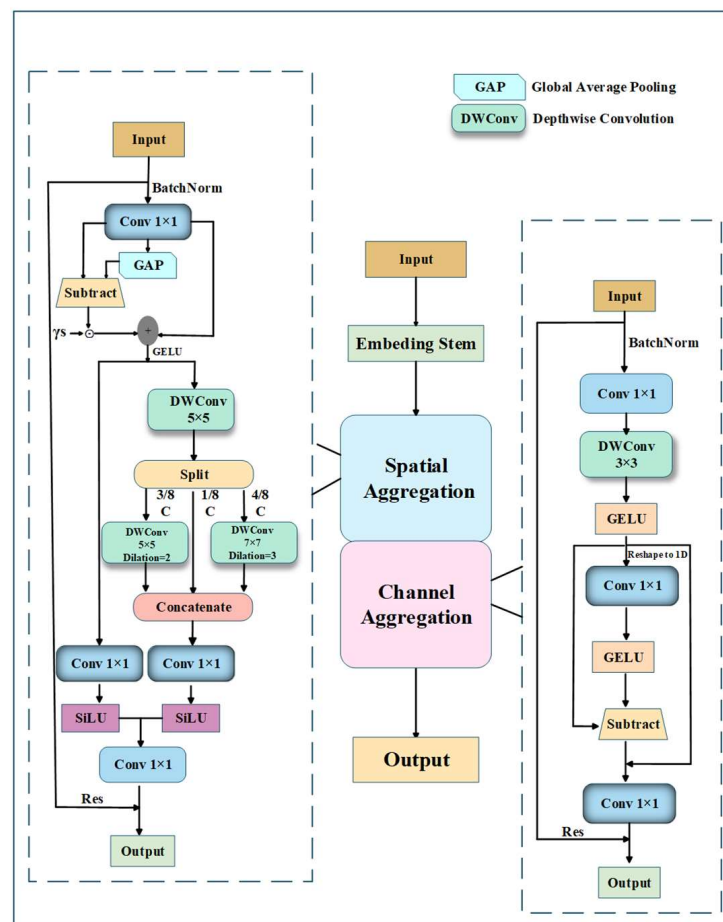


Figure 3. Spatial-Channel Aggregation (SCA).

The embedding stem is first employed to downsample the input feature map, aiming to reduce feature redundancy and alleviate computational burden while preserving essential structural information. Let the input feature be the output of the previous layer, denoted as $\bar{Z}_F^i$. The stem operation can be formulated as:

$$Z_{OUT}^{Stem} = Stem(\bar{Z}_F^i) \quad (8)$$

where $Stem(\cdot)$ denotes the downsampling operation; in our implementation, it is an Avg-Pool with kernel (1, 4) and stride (1, 4) applied along the temporal dimension. For simplicity, we denote $u = Z_{OUT}^{Stem}$ in the following sections.

2.3.1. Spatial Aggregation

The spatial aggregation (SA) module aims to enhance spatial representation capability by capturing both fine-grained local textures and middle-order global structures through multi-order feature extraction. Specifically, local spatial patterns are first modeled using pointwise convolution, while global contextual information is incorporated via global average pooling (GAP). The spatial feature embedding is initially obtained by:

$$Y = Conv1 \times 1(u) \quad (9)$$

Followed by a global–local interaction mechanism defined as:

$$Z^{(1)} = GELU(Y + \gamma_s \odot (Y - GAP(Y))) \quad (10)$$

where $\gamma_s \in R^{c \times 1}$ is a scale factor initialized to zero. $GAP(\cdot)$ extracts global information, and $Y - GAP(Y)$ emphasizes spatial feature diversity by highlighting deviations from the global context.

To further capture middle-order spatial interactions, depthwise convolutions (DW-Conv) with multiple dilation rates $d \in \{1, 2, 3\}$ are employed. A gated aggregation mechanism is then applied to adaptively fuse the extracted features:

$$Z^{(2)} = SiLU(Conv1 \times 1(Z^{(1)})) \odot SiLU(Conv1 \times 1(DWConv(Z^{(1)}))) \quad (11)$$

Finally, the spatial aggregation module is integrated into a residual framework to preserve the original feature representation:

$$Z_{OUT}^{SA} = u + FD(Norm(Z^{(2)})) \quad (12)$$

where $FD(\cdot)$ is the feature decomposition module. Through this design, the SA module effectively enhances spatial modeling capability while maintaining stable feature propagation.

2.3.2. Channel Aggregation

After spatial aggregation, the channel aggregation (CA) module is introduced to further exploit inter-channel dependencies and improve information interaction in the high-dimensional feature space. Given the spatially aggregated features Z_{OUT}^{SA} , channel-wise feature transformation is first performed as:

$$Z^{(3)} = GELU(DW_{3 \times 3}(Conv_{1 \times 1}(Norm(Z_{OUT}^{SA})))) \quad (13)$$

where $Norm(\cdot)$ is a normalization operation, $DW_{3 \times 3}(\cdot)$ is a depthwise convolution, $Conv_{1 \times 1}(\cdot)$ is a 1×1 convolution operation, and $GELU(\cdot)$ is an activation function. To

enable adaptive channel reweighting, a complementary residual modulation mechanism is applied:

$$Z^{(4)} = Z^{(3)} + \gamma_e \odot \left(Z^{(3)} - \text{GELU}\left(Z^{(3)} W_r\right) \right) \quad (14)$$

where γ_e is the channel scaling factor initialized to zero and W_r is the trainable weight. This operation reassigns the channel features through complementary interactions. The final output after the channel aggregation module is:

$$Z_{OUT}^{CA} = \text{Conv1} \times 1 \left(Z^{(4)} \right) + Z_{OUT}^{SA} \quad (15)$$

Through this process, the channel features are fused while maintaining the spatial dimensions, thereby enhancing the interaction of information between channels. Finally, the output of the channel aggregation, Z_{OUT}^{CA} , is rearranged and reshaped into a one-dimensional form. The results from all cycles are then stacked along the last dimension, and the final output is obtained through adaptive aggregation:

$$\bar{z}_{out} = \sum_{i=1}^k \text{Softmax}\left(A_{f_i}\right) \times Z_{OUT}^{CA} + Z_B^{(2)} \quad (16)$$

where the magnitude A is aggregated after $\text{Softmax}(\cdot)$ as weight coefficients.

2.4. Classifier

$$\text{Output} = \text{FC}(\text{Dropout}(\text{GeLU}(\bar{z}_{out}))) \quad (17)$$

where $\text{GeLU}(\cdot)$ is an activation function. $\text{Dropout}(\cdot)$ is employed to prevent overfitting and enhance the model's ability to generalize to new data. $\text{FC}(\cdot)$ refers to the fully connected layer, which returns the final predicted result.

3. Experiment

3.1. Experimental Details

The proposed model is implemented using the PyTorch framework (version 1.12.1), and all experiments are conducted under a unified training configuration to ensure reproducibility. The Adam optimizer is used for parameter optimization with a fixed learning rate of 1×10^{-4} , and no learning rate scheduling strategy is applied. The batch size is set to 512. To alleviate overfitting, dropout regularization is employed at different stages of the network, with a dropout rate of 0.2 applied in the main network and a higher dropout rate of 0.8 used in the classification head. The cross-entropy loss function is adopted for emotion classification. To ensure fair comparisons and reproducibility, the random seed is fixed to 1234. The parameters of the core components are summarized in Supplementary Tables S1 and S2. Different numbers of cross-validation folds are adopted for different datasets based on a comprehensive consideration of dataset scale and the number of subjects. Specifically, the DEAP dataset includes 32 subjects with a relatively sufficient number of samples; therefore, a 10-fold cross-validation strategy is employed to ensure training stability while enabling a more thorough evaluation of the model's average performance under different subject partitions. In contrast, the DREAMER dataset contains only 23 subjects, and using a larger number of folds would result in an excessively small test set per fold, leading to increased performance variance. Consequently, a 5-fold cross-validation scheme is adopted for the DREAMER dataset to achieve a balance between evaluation stability and statistical reliability.

3.2. Datasets

To thoroughly assess the performance of our proposed model, we performed a series of experiments on two publicly available benchmark datasets: DEAP [15] and DREAMER [16]. The specific parameters are listed in Table 1.

Table 1. Introduction of Datasets.

Factor	DEAP	DREAMER
subjects	32	23
EEG channels	32	14
Trials	40	18
Trial duration	60s	67–393 s
Stimuli	Audio-Video	Audio-Video
Sampling rate	128 Hz	128 Hz
Label	V/A	V/A

The DEAP dataset includes physiological signal recordings from 32 healthy subjects (16 males and 16 females, mean age 27 years) while watching 40 emotionally evocative video clips. EEG signals were recorded using a 32-channel electrode system at a 512 Hz sampling rate, which was later downsampled to 128 Hz for this study. After watching each 1 min video, subjects self-assessed the four dimensions of valence, arousal, dominance, and preference on a scale of 1–9. The DREAMER dataset contains physiological signal recordings from 23 subjects (14 males, 9 females, mean age 26.6 years) while watching 18 emotionally evocative movie clips. EEG signals were collected using a 14-channel wireless device at a 128 Hz sampling rate. Subjects self-assessed the three dimensions of valence, arousal, and sense of dominance on a scale of 1–5 after watching videos ranging from 67 to 393 s in duration. The EEG reference scheme follows the original experimental settings of each public dataset, and no additional re-referencing or manual artifact rejection was applied beyond the preprocessing provided by the dataset publishers. Detailed information about data acquisition, reference schemes, and basic preprocessing procedures is provided in the Supplementary Material. Table 1 provides a detailed list of the key attributes of the two datasets and their specific usage in this study. Among them, V/A in the Label represents valence and arousal.

3.3. Experimental Results

To ensure a strict cross-subject evaluation, all samples belonging to the same subject are exclusively assigned to either the training or testing set within each fold. Sliding window segmentation is performed prior to data partitioning, and all windows extracted from the same trial are kept within the same subject-specific split. This protocol prevents subject-specific information leakage across folds and ensures that the reported performance reflects genuine cross-subject generalization rather than temporal redundancy.

Tables 2–5 summarizes the cross-validation results on the DEAP and DREAMER datasets, where 10-fold and 5-fold cross-validation strategies are employed, respectively. Across all evaluation metrics, performance variations among different folds remain limited, indicating that the proposed MCF-SCA model exhibits stable and reliable behavior under varying subject partitions and data splits. As shown in Tables 2 and 3, on the DEAP dataset, the proposed method demonstrates consistently strong performance in both arousal and valence classification tasks. The model maintains a balanced trade-off between sensitivity and specificity across folds, without noticeable performance degradation under different cross-validation splits. This stability suggests that the learned representations are robust to inter-subject variability, which is a key challenge in cross-subject EEG emotion recognition. Similarly, the results on the DREAMER dataset, reported in Tables 4 and 5, further confirm

the effectiveness of the proposed approach. The model achieves consistently high classification performance for both arousal and valence dimensions, with only minor fluctuations across folds. Overall, the cross-validation results on both datasets demonstrate that the proposed method exhibits strong adaptability to cross-subject emotion recognition tasks and maintains stable performance across different data partition settings.

Table 2. 10-fold cross-validation results of Arousal on the DEAP dataset.

Experiment	ACC	F1	Recall	Precision	Specificity	Balanced Acc	FNR
Run-1	0.8846	0.9020	0.9186	0.8859	0.8379	0.8783	0.0814
Run-2	0.8842	0.9005	0.9146	0.8870	0.8437	0.8792	0.0854
Run-3	0.8901	0.9057	0.9202	0.8914	0.8497	0.8849	0.0798
Run-4	0.8868	0.9031	0.9141	0.8921	0.8496	0.8819	0.0859
Run-5	0.8922	0.9064	0.9163	0.8968	0.8603	0.8883	0.0837
Run-6	0.8906	0.9049	0.9072	0.9027	0.8682	0.8877	0.0928
Run-7	0.8892	0.9051	0.9095	0.9009	0.8610	0.8853	0.0905
Run-8	0.8921	0.9073	0.9156	0.8993	0.8599	0.8878	0.0844
Run-9	0.8832	0.8984	0.8924	0.9046	0.8705	0.8814	0.1076
Run-10	0.8951	0.9089	0.9095	0.9084	0.8753	0.8924	0.0905
Average	0.8888	0.9042	0.9138	0.8969	0.8576	0.8857	0.0862

Table 3. 10-fold cross-validation results of Valence on the DEAP dataset.

Experiment	ACC	F1	Recall	Precision	Specificity	Balanced Acc	FNR
Run-1	0.8759	0.8901	0.9131	0.8684	0.8304	0.8718	0.0869
Run-2	0.8758	0.8892	0.8961	0.8825	0.8503	0.8732	0.1039
Run-3	0.8751	0.8877	0.8906	0.8848	0.8561	0.8734	0.1094
Run-4	0.8726	0.8850	0.8799	0.8902	0.8635	0.8717	0.1201
Run-5	0.8770	0.8895	0.8986	0.8804	0.8506	0.8746	0.1014
Run-6	0.8706	0.8804	0.8746	0.8863	0.8661	0.8704	0.1254
Run-7	0.8762	0.8896	0.8966	0.8830	0.8510	0.8738	0.1034
Run-8	0.8750	0.8880	0.8933	0.8831	0.8524	0.8729	0.1067
Run-9	0.8705	0.8845	0.8783	0.8908	0.8452	0.8680	0.1092
Run-10	0.8699	0.8827	0.8883	0.8770	0.8474	0.8679	0.1117
Average	0.8739	0.8867	0.8918	0.8814	0.8513	0.8716	0.1082

Table 4. 5-fold cross-validation results of Arousal on the DREAMER dataset.

Experiment	ACC	F1	Recall	Precision	Specificity	Balanced Acc	FNR
Run-1	0.9638	0.9701	0.9704	0.9378	0.9535	0.9620	0.0296
Run-2	0.9582	0.9659	0.9745	0.9330	0.9329	0.9537	0.0255
Run-3	0.9585	0.9663	0.9688	0.9366	0.9423	0.9556	0.0312
Run-4	0.9653	0.9721	0.9782	0.9366	0.9445	0.9614	0.0218
Run-5	0.9621	0.9691	0.9730	0.9388	0.9450	0.9590	0.0270
Average	0.9615	0.9687	0.9729	0.9366	0.9436	0.9583	0.0271

Table 5. 5-fold cross-validation results of Valence on the DREAMER dataset.

Experiment	ACC	F1	Recall	Precision	Specificity	Balanced Acc	FNR
Run-1	0.9708	0.9799	0.9815	0.9751	0.9427	0.9621	0.0185
Run-2	0.9697	0.9792	0.9830	0.9734	0.9344	0.9587	0.0170
Run-3	0.9678	0.9778	0.9851	0.9676	0.9232	0.9541	0.0149
Run-4	0.9728	0.9815	0.9860	0.9752	0.9368	0.9614	0.0140
Run-5	0.9662	0.9765	0.9755	0.9746	0.9422	0.9589	0.0245
Average	0.9694	0.9789	0.9822	0.9732	0.9359	0.9590	0.0178

Figure 4a–d presents the confusion matrices of the proposed MCF-SCA model for the first cross-validation fold on both datasets. Specifically, Figure 4a,b illustrate the arousal and valence classification results on the DEAP dataset, while Figure 4c,d show the corresponding results on the DREAMER dataset. On DEAP, the model correctly classified 8161 high- and 5426 low-arousal samples, as well as 7718 high- and 5736 low-valence samples. On DREAMER, correct classifications reached 6531 high- and 4162 low-arousal samples, and 7878 high- and 2894 low-valence samples. Across both datasets and affective dimensions, correct predictions (true positives and true negatives) clearly dominate over misclassifications, indicating stable and reliable overall performance. Further analysis reveals that false positives (FP) and false negatives (FN) are relatively limited and well balanced, suggesting that the model does not exhibit a systematic bias toward any specific class. FN errors mainly arise from samples with weak or ambiguous emotional expressions, whereas FP errors are often associated with transient high-energy EEG patterns in low-emotion states. Notably, the model demonstrates stronger performance in the valence dimension, particularly for high-valence samples, with fewer FN errors. Overall, the pronounced diagonal dominance and controlled FP/FN distributions in the confusion matrices provide strong evidence of the robustness of the proposed MCF-SCA model under cross-subject conditions.

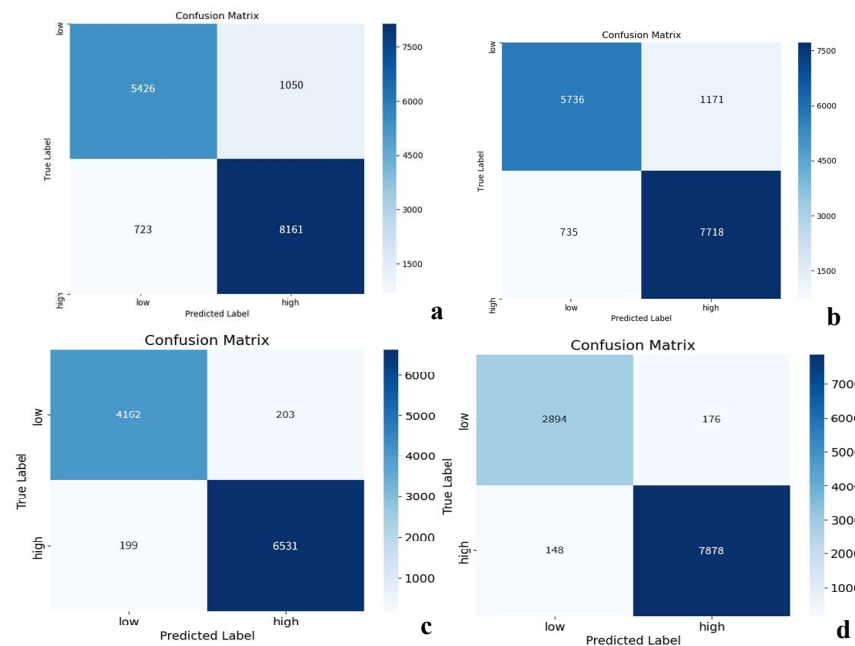


Figure 4. Confusion matrix for arousal and valence. (a) Results for the arousal dimension on the DEAP dataset; (b) Results for the valence dimension on the DEAP dataset; (c) Results for the arousal dimension on the DREAMER dataset; (d) Results for the valence dimension on the DREAMER dataset.

In this test, we implemented the TSception, CNN, LSTM, EEGNet, and MLP models using the same feature representations across both datasets and compared their performance with that of our proposed method. Figure 5 presents the comparative results. Figure 5a–d displays the 10-fold and 5-fold cross-validation results for arousal and valence on the DEAP and DREAMER datasets, respectively.

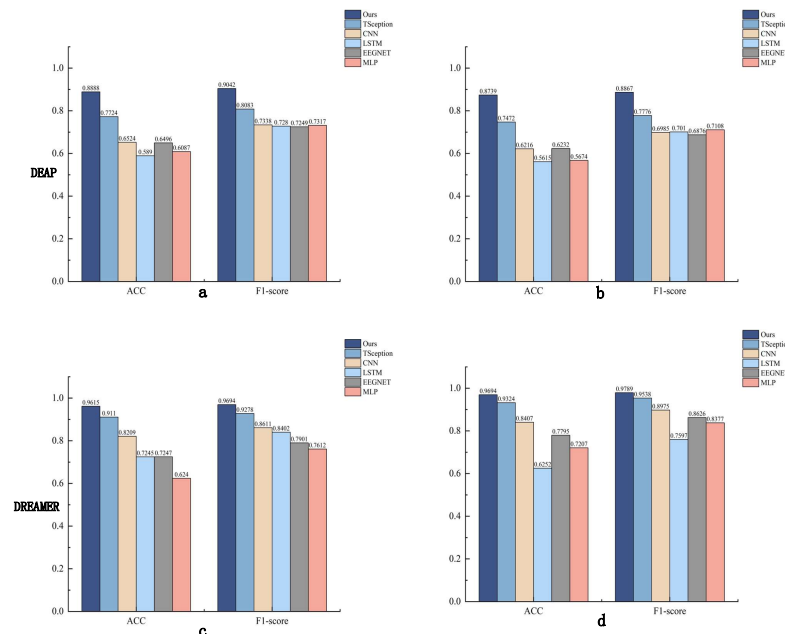


Figure 5. Classification results for six models on arousal and valence. (a) Comparison results for the arousal dimension on the DEAP dataset; (b) Comparison results for the valence dimension on the DEAP dataset; (c) Comparison results for the arousal dimension on the DREAMER dataset; (d) Comparison results for the valence dimension on the DREAMER dataset.

On the DEAP dataset, our method achieved the highest classification accuracy for the arousal dimension, outperforming all baseline methods. TSception obtained the second-best results, while EEGNet and MLP showed relatively lower performance. Similarly, on the DREAMER dataset, our method again achieved the best performance, followed by TSception. MLP recorded the lowest accuracy for arousal classification, whereas LSTM performed the worst in the valence dimension.

In summary, extensive experimental evaluations substantiate the effectiveness and high accuracy of our approach in cross-subject emotion recognition. The proposed model consistently outperforms existing methods across multiple datasets, demonstrating strong potential for EEG-based cross-subject emotion recognition applications.

3.4. Feature Visualization

To validate the classification performance of the proposed model, t-SNE [32] is employed to visualize the high-dimensional feature representations. As shown in Figure 6, the x -axis (“comp-1”) and y -axis (“comp-2”) represent the reduced principal components. We compare the feature distributions for the DEAP dataset (with 10-fold cross-validation) in Figure 6a and the DREAMER dataset (with 5-fold cross-validation) in Figure 6b, focusing on the arousal dimension. In Figure 6, “BEST” and “WORST” correspond to the folds with the highest and lowest accuracy, respectively, in the cross-validation process.

The experimental results indicate that the proposed method achieves superior performance compared with all baseline models in both the “BEST” and “WORST” settings on both datasets. The red (high arousal) and blue (low arousal) points are well separated, with minimal intra-class variance and large inter-class variance, demonstrating our model’s strong ability to distinguish between categories. Other models, like TSception and MLP, show significant performance discrepancies. MLP, in particular, has the worst performance due to severe mixing of red and blue points. CNN and LSTM show large fluctuations across folds. While some models improve in the “BEST” scenario, they still suffer from noticeable class mixing in the “WORST” case. In contrast, our method consistently maintains clear

category separation, validating its robust feature extraction ability for cross-subject emotion recognition tasks.

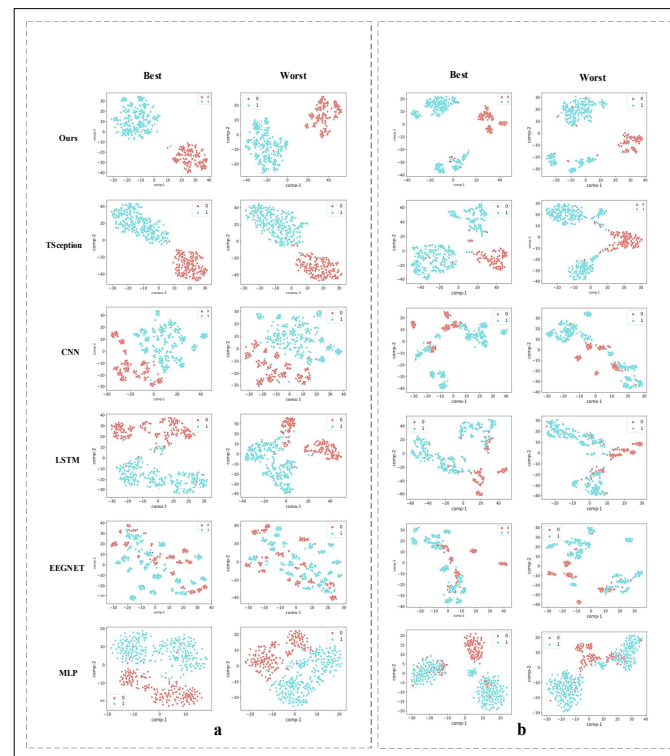


Figure 6. t-SNE visualization. (a) Feature distribution of the best and worst-performing models in the 10-fold cross-validation on the DEAP dataset; (b) Feature distribution of the best and worst-performing models in the 5-fold cross-validation on the DREAMER dataset.

3.5. Comparison with Studies in Recent Years

To further highlight the superiority of our model, we compared it with recent advanced emotion recognition methods, as shown in Figure 7.

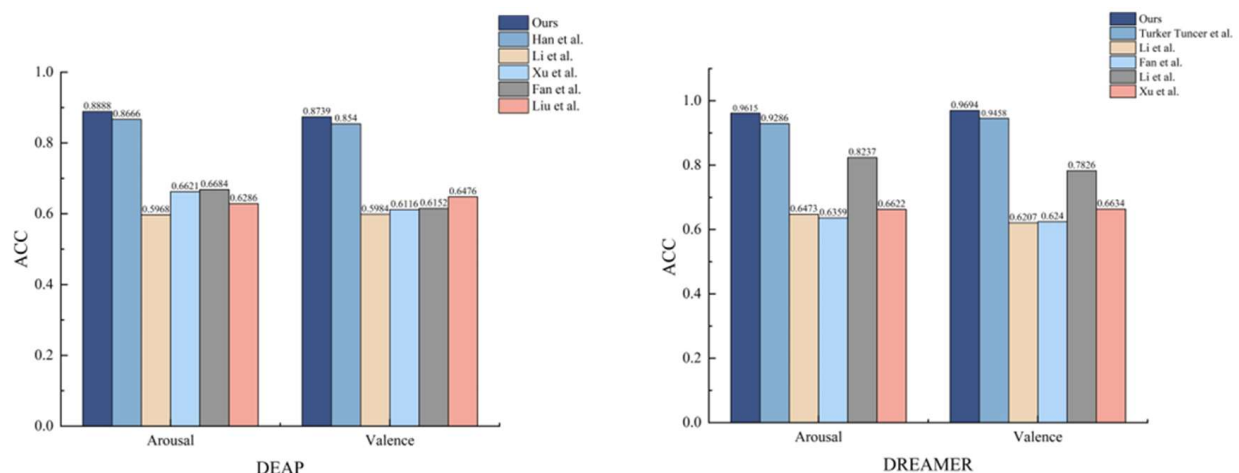


Figure 7. Comparison with other methods [12,21,27,33–36].

In the DEAP dataset, Han et al. [27] proposed an innovative deep learning-based multi-scale emotion recognition method. Our model achieves a 2.22% higher accuracy than theirs. Compared with the dynamic stream selection network proposed by Li et al. [21], our approach demonstrates an improvement of over 20 percentage points in accuracy. Additionally, our model outperforms the modified depthwise separable convolution and

graph neural network proposed by Xu et al. [33] by approximately 20% in accuracy. When compared with the residual capsule network developed by Fan et al. [34], our architecture shows a 20-percentage-point advantage in accuracy. Finally, in subject-independent experiments, our model achieves a 24-percentage-point higher accuracy than the dynamic graph attention network with multi-branch feature extraction and stage fusion proposed by Liu et al. [12].

On the DREAMER dataset, we compared our model with recently published state-of-the-art techniques. Our model outperforms the multi-level fusion feature generation network and deep classifier approach proposed by Turker Tuncer et al. [35] by 3 percentage points. Compared to the dynamic flow selection network proposed by Li et al. [21], our accuracy is over 30 percentage points higher. Our model also outperforms the residual capsule network proposed by Fan et al. [34] by approximately 30 percentage points. When compared to the improved deep separable convolution and graph neural network method proposed by Xu et al. [33], our accuracy shows an improvement of around 30%. Finally, compared to the end-to-end network STSNet based on ManifoldNet and BiLSTM proposed by Li et al. [36], our accuracy is about 15% higher. Overall, a range of data results demonstrated that our MCF-SCA approach performed very effectively in cross-subject experiments, outperforming multiple state-of-the-art approaches.

3.6. Ablation Experiments

To assess the individual contribution of each module to the overall model performance, we conducted ablation experiments on two datasets to systematically validate the impact of different modules on model performance. In separate experiments, the FFT, SA, and CA modules were individually removed. The feature extraction method and the maximum number of iterations in all comparison experiments remained consistent with those used in previous studies. The detailed experimental results are shown in Tables 6 and 7.

Table 6. Ablation experiments on the DEAP dataset.

FFT	CA	SA	Arousal		Valence	
			ACC	F1	ACC	F1
✓	✓	✓	0.8888	0.9042	0.8739	0.8867
	✓	✓	0.8751	0.8826	0.8644	0.8680
✓	✓		0.8722	0.8827	0.8607	0.8639
✓		✓	0.8721	0.8821	0.8574	0.8612

Note. Boldface indicates the best results. “✓” indicates that the corresponding module is applied.

Table 7. Ablation experiments on the DREAMER dataset.

FFT	CA	SA	Arousal		Valence	
			ACC	F1	ACC	F1
✓	✓	✓	0.9615	0.9687	0.9694	0.9790
	✓	✓	0.9408	0.9481	0.9487	0.9585
✓	✓		0.9387	0.9465	0.9467	0.9571
✓		✓	0.9405	0.9478	0.9475	0.9576

Note. Boldface indicates the best results. “✓” indicates that the corresponding module is applied.

Tables 6 and 7 present the ablation results on the DEAP and DREAMER datasets. For both datasets, the complete model consistently achieves the best performance in terms of accuracy and F1-score for arousal and valence classification. On the DEAP dataset, removing the FFT, CA, or SA module leads to clear performance degradation, indicating that each component contributes positively to emotion recognition. Similar trends are

observed on the DREAMER dataset, where excluding any module results in reduced classification performance. Notably, removing the SA module causes the most significant performance drop on DREAMER, highlighting its importance in capturing discriminative emotional information. Moreover, configurations integrating both CA and SA modules outperform those with only one of them, suggesting complementary effects between channel aggregation and self-attention. Overall, the ablation results demonstrate that the superior performance of the proposed MCF-SCA model arises from the joint contribution of all modules.

4. Discussion

This study, based on the MCF-SCA method, conducted extensive experiments on the DEAP and DREAMER datasets to explore the key challenges in cross-subject emotion recognition. The complexity of this task primarily arises from the subjectivity of emotions, the high sensitivity of EEG signals to noise, the interdependence between emotional dimensions (valence and arousal), and limitations related to dataset size, quality, and labeling. The observed performance differences between arousal and valence can be largely attributed to their distinct neurophysiological representations. Arousal is generally associated with global cortical activation levels and increased high-frequency neural activity, which tend to exhibit stronger cross-subject consistency. In contrast, valence relies more heavily on localized and asymmetric activation patterns across specific brain regions, making it more sensitive to individual differences and therefore more prone to confusion in certain subjects. From a frequency-domain perspective, the model shows higher sensitivity to arousal-related patterns in higher-frequency bands, such as the β and γ bands, whereas valence discrimination depends more on the joint contribution of low- and mid-frequency bands, particularly the θ and α bands. This frequency-dependent behavior is consistent with findings reported in prior EEG-based emotion studies and provides a plausible explanation for the differing classification characteristics observed between the two emotional dimensions. As illustrated in Figure 8, substantial inter-subject variability can be observed in both the energy distribution and spatial activation patterns within the same frequency band (e.g., the α band). Such subject-specific amplitude shifts and spatial redistributions lead to inconsistent feature representations across subjects, making fixed-weight or single-scale feature fusion strategies difficult to generalize reliably in cross-subject scenarios. To mitigate inter-individual variability in cross-subject scenarios, we introduce the Spatial-Channel Aggregation (SCA) module in our model. The spatial aggregation (SA) component employs multi-order depthwise convolutions combined with a gating mechanism to jointly capture local and global spatial context, thereby adaptively emphasizing more stable and transferable spatial patterns across different subjects. Furthermore, the channel aggregation (CA) component dynamically recalibrates channel-wise contributions in the high-dimensional feature space, suppressing noise-prone or subject-sensitive channels and redundant information. Through this sample- and subject-adaptive weighting mechanism, the SCA module enables more robust and generalized feature fusion for cross-subject emotion recognition.

Although EEG signals exhibit strong temporal characteristics, their ability to express spatial information is relatively weak. To enhance the representation of spatial features, EEG channels are mapped into a two-dimensional matrix according to their electrode locations. This spatial arrangement preserves the relative neighborhood relationships between channels, allowing convolutional operations to better exploit local spatial dependencies. It should be noted that this representation is intended for feature modeling rather than for reconstructing true cortical source activity, and it does not eliminate the volume conduction effect inherent in EEG signals. Subsequently, the FFT is applied to achieve time-frequency

domain conversion. This strategy effectively integrates multi-domain information and enhances the ability to capture variations in frequency bands. On this basis, the SCA module deeply fuses multi-domain features, effectively mitigating inter-subject variability and further exploring frequency-domain variations across spatial channels. This results in superior feature aggregation and significantly improves cross-subject classification performance. The final results demonstrate that MCF-SCA achieves high accuracy and F1-scores on both the DEAP and DREAMER datasets, validating its strong adaptability and performance in cross-subject emotion recognition.

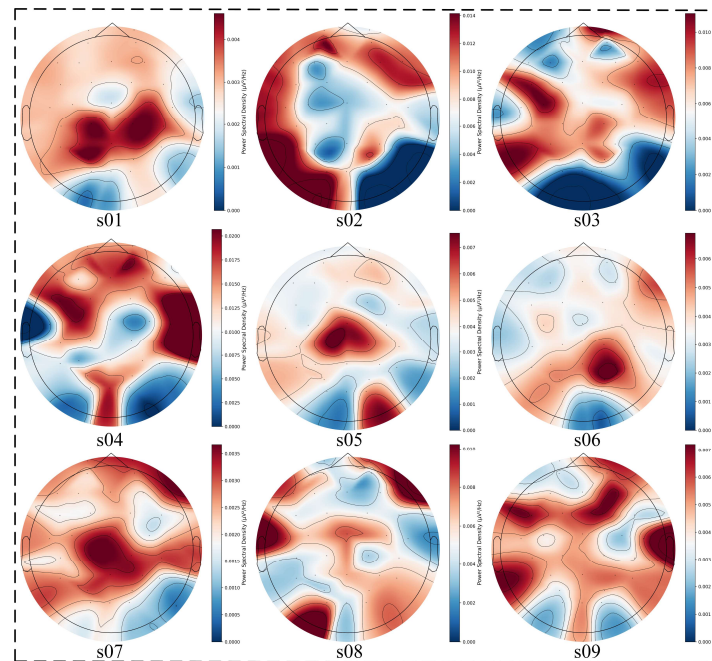


Figure 8. Distribution of features in the alpha band for different subjects.

The feature visualization results in Figure 6 highlight the advantages of our method. In both the “BEST” and “WORST” cases, the model shows concentrated data point distributions and clear clustering, indicating strong intra-class consistency and inter-class separability. In contrast, other models struggle because of their limited capacity to capture complex EEG signals, which hampers feature extraction and reduces classification accuracy.

The ablation experiment results demonstrate that each module of the MCF-SCA model plays a crucial role in classification outcomes. The SCA module, as a lightweight aggregation mechanism, captures collaborative relationships across channels, models complex dependencies between channels, and highlights task-related key information by dynamically adjusting channel weights, while suppressing irrelevant noise. By combining spatial and channel aggregation, the SCA module achieves the fusion of local and global features, dynamically selecting key spatial and channel features, greatly enhancing the spatial feature representation ability of EEG signals. Furthermore, compared to other complex attention mechanisms, the SCA module adopts a modular design, significantly improving model performance while reducing the parameter count and computational complexity (specific parameter count is shown in Table 8). This design enables the MCF-SCA model to more effectively adapt to individual differences, reducing intra-class variation while maintaining sensitivity to individual features, thereby significantly improving the accuracy of cross-subject emotion recognition tasks.

Table 8. Parameter presentation in SCA module.

Module	Parameter (M)
SCA	0.0323
CA	0.0119
SA	0.0201

The experimental results presented in Figures 5 and 7 clearly demonstrate that the MCF-SCA method outperforms other approaches in classification accuracy across most experiments. As shown in Figure 5, baseline models perform poorly due to their limited ability to capture complex EEG signals and extract meaningful time and spatial features. In contrast, our method uses multi-scale convolutions for hierarchical feature extraction and the SCA module for effective feature fusion, improving classification performance. In Figure 7, we observe the limitations of other methods in feature extraction and integration. For example, Han et al.'s approach with multi-scale convolution and TimesNet lacks effective aggregation, resulting in lower accuracy. Li et al.'s dynamic flow selection network struggles in cross-subject experiments due to ineffective spatiotemporal decoupling and static weight allocation. On the other hand, our approach efficiently extracts and uses spatiotemporal features, achieving better accuracy. Compared to Xu et al.'s improved depth-wise separable convolution and graph neural network method, our multi-scale temporal and spatial fusion strategy offers more effective feature representation. Xu's method relies solely on LSTM for temporal feature extraction, missing the spatial characteristics of EEG signals. FAN et al.'s residual convolutional capsule network has limited feature extraction capacity, particularly in the spatial domain. Liu et al.'s dynamic graph attention network performs well in subject-dependent experiments but shows poor generalization in cross-subject settings. Our method's excellent cross-validation results on different datasets (as shown in Tables 2–5) also fully prove its robustness and excellent cross-subject adaptability.

Although our MCF-SCA model has achieved promising results, there are still areas for improvement. Future research could focus on further optimizing the SCA module to enhance its adaptability across a wider range of classification tasks. Additionally, exploring multimodal fusion approaches (such as combining facial expressions, speech, etc.) could provide a more comprehensive representation of EEG signals. Furthermore, incorporating graph neural networks or other graph-based modeling techniques could help better capture the spatial information in EEG signals, thereby improving the accuracy and fine-grained differentiation ability of emotion classification.

5. Conclusions

In conclusion, the proposed MCF-SCA model offers an innovative solution for emotion recognition tasks. It not only utilizes spatial projection to map DE features onto a 2D matrix but also demonstrates exceptional performance in extracting spatial and temporal features through multi-scale convolution. By applying FFT to transform the time domain into the frequency domain and performing deeper feature extraction via multi-scale convolution, the model further enhances its representation of EEG signals. Finally, the SCA module enables precise modeling of complex emotional signals. Feature visualization and ten-fold cross-validation results fully demonstrate the robustness, stability, and, most importantly, the generalization capability of the proposed model. In the future, further optimization of this approach—such as incorporating more physiologically informed spatial transformations to mitigate volume conduction effects and exploring stricter cross-subject evaluation protocols—will promote the application and development of EEG-based emotion recognition technologies in areas such as intelligent human–machine interaction and mental health monitoring, offering new possibilities and directions for emotion computing research.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/info17030257/s1>, Table S1: Model configuration of the proposed MCF-SCA network; Table S2: Training setup and hyper-parameters; Section S1: Data Acquisition and Basic EEG Preprocessing.

Author Contributions: Conceptualization, Y.M. and J.S.; methodology, Y.M. and J.S.; validation, J.N. and D.L.; formal analysis, Y.M., J.N. and W.F.; software, W.F.; visualisation, C.Y.; Data curation, Y.X.; investigation, F.Z.; writing—original draft preparation, Y.M. and J.S.; review and editing, D.L., J.N., Y.X., W.F. and F.Z.; supervision, D.L. and C.Y.; project administration, Y.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Henan Province [No. 252300421518], the National Natural Science Foundation of China [No. 62476255 and No. 62303427], the Young Teacher Foundation of Henan Province [No. 2021GGJS093], the Key Science Research Project of Colleges and Universities in Henan Province of China [No. 25A520003, No. 23B413006 and No. 26A510009], the Key Science and Technology Program of Henan Province [No. 242102211058, No. 242102211018 and No. 252102211056], and the Young Backbone Teacher Program of Zhengzhou University of Light Industry.

Data Availability Statement: The DEAP and DREAMER datasets are publicly available benchmark datasets. Due to potential accessibility issues with the original hosting websites, the datasets are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Choi, G.-Y.; Shin, J.-G.; Lee, J.-Y.; Lee, J.-S.; Heo, I.-S.; Yoon, H.-Y.; Lim, W.; Jeong, J.-W.; Kim, S.-H.; Hwang, H.-J. EEG Dataset for the Recognition of Different Emotions Induced in Voice-User Interaction. *Sci. Data* **2024**, *11*, 1084. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lee, M.H. *EAV: EEG-Audio-Video Dataset for Emotion Recognition in Conversational Contexts*; Zenodo: Geneva, Switzerland, 2023.
- Baygin, M.; Tuncer, I.; Dogan, S.; Barua, P.D.; Tuncer, T.; Cheong, K.H.; Acharya, U.R. Automated Facial Expression Recognition Using Exemplar Hybrid Deep Feature Generation Technique. *Soft Comput.* **2023**, *27*, 8721–8737. [\[CrossRef\]](#)
- Li, H.; Zhang, X.; Duan, S.; Liang, H. Speech Emotion Recognition Based on Bi-Directional Acoustic–Articulatory Conversion. *Knowl.-Based Syst.* **2024**, *299*, 112123. [\[CrossRef\]](#)
- Pathirana, A.; Rajakaruna, D.K.; Kasthurirathna, D.; Atukorale, A.; Aththidiye, R.; Yatipansalawa, M. A Reinforcement Learning-Based Approach for Promoting Mental Health Using Multimodal Emotion Recognition. *J. Future Artif. Intell. Technol.* **2024**, *1*, 124–142. [\[CrossRef\]](#)
- Nugroho, K.; Al Amin, I.H.; Noviasari, N.A.; Setiadi, D.R.I.M. Prosodic Spatio-Temporal Feature Fusion with Attention Mechanisms for Speech Emotion Recognition. *Computers* **2025**, *14*, 361. [\[CrossRef\]](#)
- Liu, H.; Zhang, Y.; Li, Y.; Kong, X. Review on Emotion Recognition Based on Electroencephalography. *Front. Comput. Neurosci.* **2021**, *15*, 758212. [\[CrossRef\]](#) [\[PubMed\]](#)
- Li, W.; Wang, M.; Zhu, J.; Song, A. EEG-Based Emotion Recognition Using Trainable Adjacency Relation Driven Graph Convolutional Network. *IEEE Trans. Cogn. Dev. Syst.* **2023**, *15*, 1656–1672. [\[CrossRef\]](#)
- He, Z.; Zhong, Y.; Pan, J. An Adversarial Discriminative Temporal Convolutional Network for EEG-Based Cross-Domain Emotion Recognition. *Comput. Biol. Med.* **2022**, *141*, 105048. [\[CrossRef\]](#)
- Nath, D.; Anubhav; Singh, M.; Sethia, D.; Kalra, D.; Indu, S. A Comparative Study of Subject-Dependent and Subject-Independent Strategies for EEG-Based Emotion Recognition Using LSTM Network. In Proceedings of the 2020 4th International Conference on Compute and Data Analysis, New York, NY, USA, 9–12 March 2020; pp. 142–147.
- Kim, J. Bimodal Emotion Recognition Using Speech and Physiological Changes. *Robust Speech Recognit. Underst.* **2007**, *265*, 280.
- Liu, S.; Wang, X.; Jiang, M.; An, Y.; Gu, Z.; Li, B.; Zhang, Y. MAS-DGAT-Net: A Dynamic Graph Attention Network with Multibranch Feature Extraction and Staged Fusion for EEG Emotion Recognition. *Knowl.-Based Syst.* **2024**, *305*, 112599. [\[CrossRef\]](#)
- Zheng, W.-L.; Lu, B.-L. Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. *IEEE Trans. Auton. Ment. Dev.* **2015**, *7*, 162–175. [\[CrossRef\]](#)
- Zheng, W.-L.; Liu, W.; Lu, Y.; Lu, B.-L.; Cichocki, A. Emotionmeter: A Multimodal Framework for Recognizing Human Emotions. *IEEE Trans. Cybern.* **2018**, *49*, 1110–1122. [\[CrossRef\]](#) [\[PubMed\]](#)
- Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. Deap: A Database for Emotion Analysis; Using Physiological Signals. *IEEE Trans. Affect. Comput.* **2011**, *3*, 18–31. [\[CrossRef\]](#)

16. Katsigiannis, S.; Ramzan, N. DREAMER: A Database for Emotion Recognition through EEG and ECG Signals from Wireless Low-Cost off-the-Shelf Devices. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 98–107. [\[CrossRef\]](#)
17. Wang, Y.; Zhou, Y.; Lu, W.; Wu, Q.; Li, Q.; Zhang, R. AC-CfC: An Attention-Based Convolutional Closed-Form Continuous-Time Neural Network for Raw Multi-Channel EEG-Based Emotion Recognition. *Biomed. Signal Process. Control* **2024**, *94*, 106249. [\[CrossRef\]](#)
18. Wu, X.; Zhang, Y.; Li, J.; Yang, H.; Wu, X. FC-TFS-CGRU: A Temporal-Frequency-Spatial Electroencephalography Emotion Recognition Model Based on Functional Connectivity and a Convolutional Gated Recurrent Unit Hybrid Architecture. *Sensors* **2024**, *24*, 1979. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Samavat, A.; Khalili, E.; Ayati, B.; Ayati, M. Deep Learning Model with Adaptive Regularization for EEG-Based Emotion Recognition Using Temporal and Frequency Features. *IEEE Access* **2022**, *10*, 24520–24527. [\[CrossRef\]](#)
20. Zali-Vargahan, B.; Charmin, A.; Kalbkhani, H.; Barghandan, S. Deep Time-Frequency Features and Semi-Supervised Dimension Reduction for Subject-Independent Emotion Recognition from Multi-Channel EEG Signals. *Biomed. Signal Process. Control* **2023**, *85*, 104806. [\[CrossRef\]](#)
21. Li, W.; Dong, J.; Liu, S.; Fan, L.; Wang, S. Dynamic Stream Selection Network for Subject-Independent EEG-Based Emotion Recognition. *IEEE Sens. J.* **2024**, *24*, 19336–19343. [\[CrossRef\]](#)
22. Arjun; Rajpoot, A.S.; Panicker, M.R. Subject Independent Emotion Recognition Using EEG Signals Employing Attention Driven Neural Networks. *Biomed. Signal Process. Control* **2022**, *75*, 103547. [\[CrossRef\]](#)
23. Pandey, P.; Seeja, K.R. A One-Dimensional CNN Model for Subject Independent Emotion Recognition Using EEG Signals. In *Proceedings of the International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021*; Springer: Berlin/Heidelberg, Germany, 2021; Volume 2, pp. 509–515.
24. Ding, Y.; Robinson, N.; Zhang, S.; Zeng, Q.; Guan, C. TSception: Capturing Temporal Dynamics and Spatial Asymmetry from EEG for Emotion Recognition. *IEEE Trans. Affect. Comput.* **2022**, *14*, 2238–2250. [\[CrossRef\]](#)
25. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
26. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Trans. Affect. Comput.* **2011**, *3*, 42–55. [\[CrossRef\]](#)
27. Han, L.; Zhang, X.; Yin, J. EEG Emotion Recognition Based on the TimesNet Fusion Model. *Appl. Soft Comput.* **2024**, *159*, 111635. [\[CrossRef\]](#)
28. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. Timesnet: Temporal 2d-Variation Modeling for General Time Series Analysis. *arXiv* **2022**, arXiv:2210.02186.
29. Li, S.; Wang, Z.; Liu, Z.; Tan, C.; Lin, H.; Wu, D.; Chen, Z.; Zheng, J.; Li, S.Z. Moganet: Multi-Order Gated Aggregation Network. *arXiv* **2022**, arXiv:2211.03295.
30. Shi, L.-C.; Jiao, Y.-Y.; Lu, B.-L. Differential Entropy Feature for EEG-Based Vigilance Estimation. In *Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*; IEEE: Osaka, Japan, 2013; pp. 6627–6630.
31. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the International Conference on Machine Learning*; PMLR: New York, NY, USA, 2015; pp. 448–456.
32. van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
33. Xu, Z.; Chen, N.; Li, G.; Li, J.; Zhu, H.; Zhu, Z. The Mitigation of Heterogeneity in Temporal Scale among Different Cortical Regions for EEG Emotion Recognition. *Knowl.-Based Syst.* **2025**, *309*, 112826. [\[CrossRef\]](#)
34. Fan, C.; Wang, J.; Huang, W.; Yang, X.; Pei, G.; Li, T.; Lv, Z. Light-Weight Residual Convolution-Based Capsule Network for EEG Emotion Recognition. *Adv. Eng. Inform.* **2024**, *61*, 102522. [\[CrossRef\]](#)
35. Tuncer, T.; Dogan, S.; Subasi, A. LEDPatNet19: Automated Emotion Recognition Model Based on Nonlinear LED Pattern Feature Extraction Function Using EEG Signals. *Cogn. Neurodyn.* **2022**, *16*, 779–790. [\[CrossRef\]](#)
36. Li, R.; Ren, C.; Zhang, S.; Yang, Y.; Zhao, Q.; Hou, K.; Yuan, W.; Zhang, X.; Hu, B. STSNet: A Novel Spatio-Temporal-Spectral Network for Subject-Independent EEG-Based Emotion Recognition. *Health Inf. Sci. Syst.* **2023**, *11*, 25. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.