Multi-Scale Masked Autoencoders for Cross-Session Emotion Recognition

Miaoqi Pang, Hongtao Wang, *Senior Member, IEEE*, Jiayang Huang, Chi-Man Vong, *Senior Member, IEEE*, Zhiqiang Zeng, and Chuangquan Chen, *Member, IEEE*

***Abstract*— Affective brain-computer interfaces (aBCIs) have garnered widespread applications, with remarkable advancements in utilizing electroencephalogram (EEG) technology for emotion recognition. However, the time-consuming process of annotating EEG data, inherent individual differences, non-stationary characteristics of EEG data, and noise artifacts in EEG data collection pose formidable challenges in developing subject-specific cross-session emotion recognition models. To simultaneously address these challenges, we propose a unified pre-training framework based on multi-scale masked autoencoders (MSMAE), which utilizes large-scale unlabeled EEG signals from multiple subjects and sessions to extract noise-robust, subject-invariant, and temporal-invariant features. We subsequently fine-tune the obtained generalized features with only a small amount of labeled data from a specific subject for personalization and enable cross-session emotion recognition. Our framework emphasizes: 1) multi-scale representation to capture diverse aspects of EEG signals, obtaining comprehensive information; 2) an improved masking mechanism for robust channel-level representation learning, addressing missing channel issues while preserving inter-channel relationships; and 3) invariance learning for regional correlations in spatial-level representation, minimizing inter-subject and inter-session variances. Under these elaborate designs, the proposed MSMAE exhibits a remarkable ability to decode emotional states from a different session of EEG data during the testing phase. Extensive experiments conducted on the two publicly available datasets, i.e., SEED and SEED-IV, demonstrate that the proposed MSMAE consistently achieves stable results and outperforms competitive baseline methods in cross-session emotion recognition.**

Manuscript received 12 October 2023; revised 28 February 2024; accepted 5 April 2024. Date of publication 15 April 2024; date of current version 23 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62201402, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011978 and Grant 2020A1515111154, in part by the Projects for International Scientific and Technological Cooperation of Guangdong Province under Grant 2023A0505050144, in part by the Educational Commission of Guangdong Province under Grant 2021KTSCX136, and in part by Hong Kong and Macau Joint Research and Development Fund of Wuyi University under Grant 2021WGALH19. *(Corresponding author: Chuangquan Chen.)*

Miaoqi Pang, Hongtao Wang, Jiayang Huang, Zhiqiang Zeng, and Chuangquan Chen are with the School of Electronics and Information Engineering, Wuyi University, Jiangmen 529020, China (e-mail: chenchuangquan87@163.com).

Chi-Man Vong is with the Department of Computer and Information Science, University of Macau, Macau, China.

Digital Object Identifier 10.1109/TNSRE.2024.3389037

***Index Terms*— EEG-based emotion recognition, self-supervised learning, cross-session, transformer.**

I. Introduction

AFFECTIVE Brain-computer Interfaces (aBCIs) employ brain imaging techniques to capture and interpret human emotional states, aiming to achieve emotional communication and expression between humans and computers. This endeavor enhances both the immersive user experience and the efficiency of human-computer interaction. Additionally, aBCIs exhibit promising applications in fields such as healthcare and education for long-term monitoring and prediction of emotional states, enabling personalized psychological interventions and treatment plans [1], [2]. With aBCIs, a variety of modalities have been utilized, including functional magnetic resonance imaging (fMRI), Near-Infrared Spectroscopy (NIRS), and electroencephalography (EEG). In particular, EEG-based aBCIs have garnered increasing attention due to the rapid advancements in noninvasive, user-friendly, and low-cost EEG recording devices, particularly with the aid of portable dry electrode devices [3].

EEG-based aBCIs have demonstrated their capability to decode users' intentions from brain recordings and have showcased potential applications in neural rehabilitation systems [4]. However, individual differences and the non-stationary characteristic of EEG [5] render the development of stable EEG-based emotion recognition models a challenging task. Consequently, it is necessary to collect labeled samples for each subject at each time to train new models, leading to time-consuming and expensive labeling work. To mitigate the reliance on the labeled data, in recent years, an increasing number of researchers have turned their focus on applying transfer learning methods to reduce individual differences [5], [6], [7], [8], [9] and improve feature invariance representation [10], [11], [12].

Currently, the predominant transfer learning methods employed in EEG-based aBCIs include domain adaptation (DA) and domain generalization (DG). These methods are designed to reduce the distribution discrepancy between the source and target domains, thus resulting in an improved recognition performance in the target domain. Nevertheless, DA methods require utilizing the target domain during the training stage and typically assume that the data distribution

remains invariant or changes minimally between the source domain and target domain. In scenarios where the data distribution continuously evolves during real-time data acquisition, DA methods cannot effectively adapt these variations. On the other hand, DG generates domain-invariant representations from the source domains without exposure to data from the target domain, thus being more suitable for practical applications. However, DG methods require large numbers of source domains to train the model and enhance its generalization capabilities.

DA methods require access to target domains with data distributions, while DG methods need large numbers of source domains. These approaches are impractical for the following cross-session emotion recognition scenario: when only one session (i.e., one source domain) of labeled data is available for a specific subject during the training stage. In this context, the primary concern is effectively utilizing the limited labeled data to train a subject-specific model for cross-session emotion recognition.

Within the context of the brain-big-data center, real-time EEG data from a vast group of individuals are continuously transmitted, resulting in an abundance of unlabeled signals from various subjects and sessions, potentially containing some degree of corruption. Therefore, this situation presents an intriguing challenge: Can these unlabeled data be combined with the limited labeled data to train a subject-specific model for cross-session emotion recognition? This paper addresses this challenge by proposing Multi-Scale Masked Autoencoders (MSMAE). The MSMAE model is based on a multi-scale Vision-Transformer hybrid architecture, incorporating spectrum embedding, multi-head spatial attention, and multi-scale feature fusion to capture channel and spatial information of the EEG signals effectively. Specifically, MSMAE is pre-trained using unlabeled EEG data from multiple subjects and sessions, encoding and reconstructing channel-level and spatial-level representations of EEG signals to extract noise-robust, subject-invariant, and temporal-invariant features. Subsequently, only a small amount of labeled data from specific subjects is necessary to fine-tune the model for personalization. Under this comprehensive training, the subject-specific model demonstrates a remarkable ability to decode emotional states from a different session of EEG data during the testing phase.

The main contributions of this study can be summarized in three aspects:

1) We introduce a unified multi-scale pre-training framework aimed at addressing challenges related to missing EEG channels and limited labeled data in emotion recognition. This framework significantly enhances the practicality and effectiveness of EEG-based emotion recognition in real-world applications.

2) We present an innovative multi-scale fusion approach that combines channel-level and spatial-level learning. Our model aligns spatial-level correlations between pre-training and fine-tuning data to mitigate inter-subject and inter-session variations. Furthermore, it fine-tunes channel-level representation to ensure the exclusivity of subject-specific features. These techniques enhance adaptability and robustness for subject-specific cross-session emotion recognition tasks.

3) Our proposed model exhibits superior performance on two publicly available datasets for cross-session emotion recognition, even when only one session of labeled data is accessible for training.

The organization of this paper is structured as follows: Section II offers a brief review of related works. Section III elaborates on the proposed method. Section IV conducts a comprehensive evaluation of the proposed method. Finally, Section V concludes the paper.

II. Related Work

A. EEG Emotion Recognition

EEG-based emotion recognition depends on extracting sufficiently discriminative EEG features. The widely used EEG features can be categorized into four groups: temporal-domain features, frequency-domain features, time-frequency-domain features, and brain connectivity features. The commonly employed statistical information in the temporal domain includes entropy, the fractal dimension, and higher-order crossings [13], [14]. Within the frequency domain, power spectral density (PSD) [15] and differential entropy (DE) [16] stand out as two of the most frequently employed features. Several approaches [17], [18], [19], [20] have demonstrated excellent performance for time-frequency-domain features. Nalwaya et al. [19] employed the Fourier-Bessel domain adaptive wavelet transform (FBDAWT) to analyze multi-sensor EEG signals, accurately identifying emotional states. Bhattacharyya et al. [20] integrated the empirical wavelet transform (EWT) with Fourier–Bessel series expansion (FBSE), resulting in enhanced time-frequency representation of multi-component signals. For brain connectivity features, two crucial features, namely the Phase Lag Index (PLI) and the Phase Lock Value (PLV), were utilized to assess the phase synchronization among electrode signals across various brain regions. Liu et al. [21] employed the PLI feature to discern the emotional states of individual subjects, highlighting its remarkable discriminative capability. Chen et al. [22] integrated frequency-domain features with brain connectivity features for cross-subject emotion recognition, demonstrating superior performance. Furthermore, with the widespread adoption of deep learning methods, Alhagry et al. [23] utilized a two-layer long short-term memory network to extract temporal features. Zhang et al. [24] employed a recurrent neural network (RNN) to capture spatial-temporal representations from EEG signals. Zhong et al. [8] introduced a regularized graph neural network that considers the topological structure of EEG channels. Although these supervised approaches have successfully enhanced emotion recognition performance based on EEG signals, they require well-annotated and robust EEG data, which is relatively challenging to obtain in practical applications. Additionally, they often ignore the influence of session differences, such as the variations in the duration and content of the elicitation videos across different experiments, which introduce emotional biases.

B. Transfer Learning

Transfer learning seeks to enhance the performance of a new task by leveraging knowledge from a source task. DA, a subset

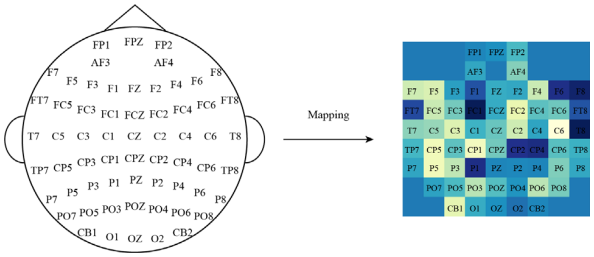

Fig. 1. Mapping EEG electrode distribution map to a two-dimensional plane. The left illustration depicts the spatial arrangement of channels on the brain cap, while the right is the 2D converted feature matrix format. The missing channels are filled with 0.

of transfer learning, has been extensively applied in EEG-based emotion recognition, demonstrating promising results. Chen et al. [25] introduced a multi-source marginal distribution adaptation method that captures domain-invariant and domain-specific features for emotion recognition. Li et al. [26] developed an innovative domain adaptation method for emotion recognition, which extracts generalized features across different subjects and sessions by simultaneously adapting both the marginal and conditional distributions to approximate the joint distribution. However, these DA methods require access to target domains with data distributions. Unlike DA, DG aims to generate domain-invariant representations from the source domains without utilizing data from the target domain. Ma et al. [27] developed a domain residual network that facilitated the separate learning of domain-specific and domain-shared weights, with the latter being used to classify emotion in unknown domains. Ozdenizci et al. [10] proposed an adversarial inference approach to extend deep learning models for EEG-based person identification, aiming to learn session-invariant person-discriminative representations. However, this requirement becomes impractical when only one source domain of labeled data is available. Recently, Li et al. [28] utilized self-supervised learning for initial model pre-training and subsequently fine-tuned the model on new data, demonstrating notable performance in emotion recognition tasks, including scenarios where data may be incomplete or corrupted. However, this model cannot handle complex tasks such as cross-session analysis. Conducting cross-session emotion recognition with limited training data still poses significant challenges.

III. Method

A. Formulation

We transform the EEG channels into a two-dimensional plane using the EEG electrode distribution map to improve spatial information consistency among adjacent channels, as depicted in Fig.1. Specifically, each channel is repositioned onto a two-dimensional electrode topology, with a size of 9×9, and zero-padding is performed for missing electrodes. We apply this transformation to frequency-domain features, resulting in the EEG image $\mathbf{x} \in \mathbb{R}^{9\times9\times C_f}$, where C_f represents the number of frequency bands. The pre-training dataset consists of unlabeled data from various subjects and sessions, represented as $\mathbf{X}_{Pre} = \{\mathbf{x}_{Pre}^{(i)}\}_{i=1}^{N_{Pre}} \in \mathbb{R}^{N_{Pre}\times9\times9\times C_f}$, with N_{Pre} being the number of samples in this dataset. The fine-tuning data contains a limited amount of labeled data from a specific subject s, represented as $\mathbf{X}_F^s = \{\mathbf{x}_F^{(i)}\}_{i=1}^{N_F} \in \mathbb{R}^{N_F\times9\times9\times C_f}$ and $\mathbf{Y}_F^s = \{\mathbf{y}_F^{(i)}\}_{i=1}^{N_F}$, where N_F is the number of samples in this dataset. The test data and labels for the specific subject s are denoted as $\mathbf{X}_T^s = \{\mathbf{x}_T^{(i)}\}_{i=1}^{N_T} \in \mathbb{R}^{N_T\times9\times9\times C_f}$ and $\mathbf{Y}_T^s = \{\mathbf{y}_T^{(i)}\}_{i=1}^{N_T}$, with N_T representing the number of samples in the test dataset.

B. Overview

We propose a Multi-scale pre-training model based on mask autoencoder (MAE) [29], as shown in Fig.2. The framework consists of a multi-scale pre-training stage, a personalized fine-tuning stage, and a personal testing stage.

In the multi-scale pre-training stage, both the channel-level feature extractor E_{Pre_1} and the spatial-level feature extractor E_{Pre_3} are employed to extract general information, which is shared by all subjects. Specifically, the unlabeled EEG data $\mathbf{x}_{Pre}$ is initially convolved with different scales of kernels (1×1 and 3×3), which are represented by $Conv_1$ and $Conv_3$, resulting in channel-level representation $\tilde{\mathbf{x}}_{Pre\_1}$ and spatial-level representation $\tilde{\mathbf{x}}_{Pre\_3}$. For channel-level representation $\tilde{\mathbf{x}}_{Pre\_1}$, considering the presence of missing data in some channels, we avoid encoding these channels with missing data to preserve complete information and prevent the introduction of noise. We reconstruct the masked portions to learn the encoder E_1 and obtain $\mathbf{z}_{Pre\_1}$. For spatial-level representation $\tilde{\mathbf{x}}_{Pre\_3}$, which include multiple channel information, we apply the attention feature extractor, denoted by $Attn$, to align the features of pre-training data and fine-tuning data based on brain region correlations, resulting in the aligned feature $\bar{\mathbf{x}}_{Pre\_3}$. We subsequently employ masking and reconstruction on $\bar{\mathbf{x}}_{Pre\_3}$ to learn the encoder E_3 and obtain $\mathbf{z}_{Pre\_3}$. The formulas are as follows:

$$\mathbf{z}_{Pre\_1} = E_{Pre\_1}(\mathbf{x}_{Pre}) = E_1(Conv_1(\mathbf{x}_{Pre})) \quad (1)$$

$$\mathbf{z}_{Pre\_3} = E_{Pre\_3}(\mathbf{x}_{Pre}) = E_3(Attn(Conv_3(\mathbf{x}_{Pre}))) \quad (2)$$

In the fine-tuning calibration stage, only a limited amount of labeled data from a specific subject is employed to fine-tune channel-level feature extractor E_{Pre_1} for the personal emotion predictor. Simultaneously, we freeze the parameters of the pre-trained spatial-level feature extractor E_{Pre_3} for the generalized emotion predictor. Finally, we fuse the channel-level representation with the spatial-level representation to perform the final emotion classification. Through this comprehensive training, the subject-specific model demonstrates an exceptional capability to decode emotional states from a different session of EEG data during the test phase. We elaborate on each stage as follows.

C. Multi-Scale Pre-Training

To use more corrupted EEG data and enhance the learning capacity of the model, we adopt the MAE framework with a transformer-based backbone network [30]. The model splits images into equal blocks and uses transformer encoders to extract features, with an asymmetric encoder-decoder design for image reconstruction. It leverages transformers for global information, masking for robustness, and self-supervised training for generalizability.

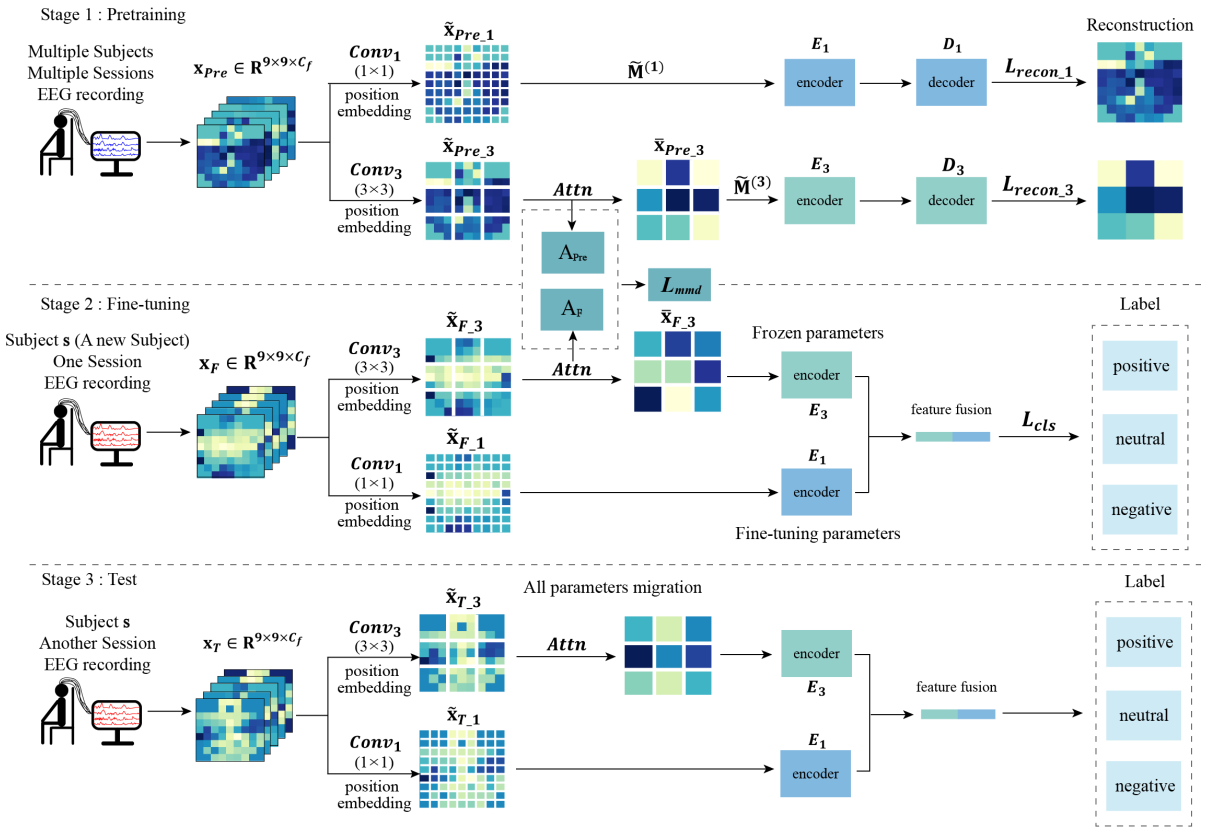

Fig. 2. Overall structure of MSMAE. The framework consists of a multi-scale pre-training stage, a personalized fine-turning stage, and a personal testing stage.

In our study, we employ convolutional kernels for patch embedding. The size of the convolutional kernel offers different interpretations for partitioning in two-dimensional EEG images, where 1×1 convolutions partition individual electrodes to learn inter-channel relationships, and 3×3 convolutions are utilized to learn about broad spatial features. We conduct multi-scale feature fusion to enhance data utilization and model representation capacity, enabling the extraction of deeper emotional representations from the frequency domain channel features and spatial features of the EEG.

1) Channel-Level Representation: By employing 1×1 convolution, we map each EEG electrode to a patch, enabling the vision-transformers framework to encode channel relationships and capture specific feature information. However, the challenge of partially missing channels and zero-padding, combined with random masking, risks losing valuable data. To address this, we have improved our approach by ensuring all zero-padded patches are masked, preserving meaningful channel information in our feature extraction process. More specifically, given the input pre-training data $\mathbf{x}_{Pre}$, we embed patches using C_1 convolutional kernels of size 1×1 with added positional embeddings, obtaining $\tilde{\mathbf{x}}_{Pre\_1} \in \mathbb{R}^{9\times9\times C_1}$:

$$\tilde{\mathbf{x}}_{Pre\_1} = Conv_1(\mathbf{x}_{Pre}, kernel\_size = (1,1), stride = 1) \tag{3}$$

where $Conv_1$ represents a convolution operation. Assuming that out of the 81 (9×9) patches, there are p non-zero padded patches (e.g., $p = 62$ as illustrated in Fig.1). To ensure the effectiveness of subsequent feature encoding, we randomly mask these p non-zero padded patches in addition to masking all zero-padded patches. The formula is as follows:

$$\mathrm{M}_{i,j}^{(1)} = \begin{cases} 0, & if\ position(i,j) should\ be\ masked \\ 1, & otherwise \end{cases} \tag{4}$$

$$q_{i,j} = \sum_{k=1}^{C_f} x_{i,j,k} \tag{5}$$

$$Q_{i,j} = \begin{cases} 1, & if\ q_{i,j} \neq 0 \\ 0, & otherwise \end{cases} \tag{6}$$

$$\tilde{\mathbf{M}}^{(1)} = \mathbf{M}^{(1)} \circ \mathbf{Q} \tag{7}$$

where $x_{i,j,k} \in \mathbf{x}_{Pre}$, $\mathbf{M}^{(1)} = \left\{\mathrm{M}_{i,j}^{(1)}\right\}_{i=1,j=1}^{9} \in \mathbb{R}^{9\times9}$ represents the original random mask, $\mathbf{Q} = \left\{Q_{i,j}\right\}_{i=1,j=1}^{9} \in \mathbb{R}^{9\times9}$ represents the matrix corresponding to the 2D EEG images with missing channels, $\tilde{\mathbf{M}}^{(1)} = \left\{\tilde{\mathrm{M}}_{i,j}^{(1)}\right\}_{i=1,j=1}^{9}$ represents the updated mask, and $\circ$ denotes the element-wise multiplication.

2) Spatial-Level Representation: When using a 3×3 convolution for partitioning, each patch contains more electrode channel information. The neighboring channels in EEG signals influence each other and reflect the corresponding brain region's signal characteristics. The connectivity between these brain regions is closely related to their spatial positions. The spatial features of EEG signals reflect the coordination and interaction among different areas of the brain, which is crucial for analyzing the spatial distribution and temporal variations of neural activity. In cross-session emotion recognition experiments, factors, like induced emotional stimuli, external environments, and physiological expressions contribute to

variability. However, the regional influence of EEG signals remains more objective and stable. Therefore, we consider further encoding and decoding the spatial features. By using large-scale convolutions for weighted average partitioning, we not only incorporate spatial features of brain regions to some extent but also exhibit universality across all EEG data with missing channels, reducing the workload of data preprocessing. Specifically, given the input pre-training data $\mathbf{x}_{Pre}$, C_3 convolutional kernels of size 3×3 are applied to obtain $\tilde{\mathbf{x}}_{Pre\_3} \in \mathbb{R}^{3\times3\times C_3}$:

$$\tilde{\mathbf{x}}_{Pre\_3} = Conv_3(\mathbf{x}_{Pre}, kernel\_size = (3,3), stride = 3) \tag{8}$$

$$\tilde{\mathbf{x}}_{Pre\_3}(i,j) \leftarrow \frac{\tilde{\mathbf{x}}_{Pre\_3}(i,j)}{\sum_{a=1}^{3}\sum_{b=1}^{3}\mathbb{I}(\mathbf{x}_{Pre}((i-1)\times 3+a,(j-1)\times 3+b)\neq \mathbf{0})} \tag{9}$$

Here, $Conv_3$ represents a convolution operation, and $\mathbb{I}(\cdot)$ denotes the indicator function that returns 1 if the condition is true and 0 otherwise. $\tilde{\mathbf{x}}_{Pre\_3}(i,j) \in \mathbb{R}^{C_3}$(for i = 1,2,3 and j = 1,2,3) represents the patch obtained through 3×3 convolution, and it is then normalized by dividing by the number of existing channels in each corresponding patch.

3) Invariance Learning for Region Correlation: We align pre-training and fine-tuning data features based on brain region correlations to obtain subject-invariant and temporal-invariant features. Considering that each individual's emotional fluctuations are unique and represent their distinct characteristics, we choose to align the shared features based on brain region correlations instead of directly aligning the pre-training spatial-level representation data $\tilde{\mathbf{x}}_{Pre\_3}$ and fine-tuning spatial-level representation data $\tilde{\mathbf{x}}_{F\_3}$.This approach partially attenuates the differences in data distribution while preserving the unique characteristics of EEG signals. Specifically, $\tilde{\mathbf{x}}_{Pre\_3} \in \mathbb{R}^{3\times3\times C_3}$ and $\tilde{\mathbf{x}}_{F\_3} \in \mathbb{R}^{3\times3\times C_3}$ are first rearranged into $\tilde{\mathbf{x}}^R_{Pre\_3} \in \mathbb{R}^{9\times C_3}$ and $\tilde{\mathbf{x}}^R_{F\_3} \in \mathbb{R}^{9\times C_3}$, respectively. Subsequently, an attention mechanism is employed to capture the correlations between patches:

$$\boldsymbol{A}_{Pre} = \frac{\boldsymbol{Q}_{Pre}\boldsymbol{K}^T_{Pre}}{\sqrt{d_k}} \in \mathbb{R}^{9\times9}, \boldsymbol{A}_F = \frac{\boldsymbol{Q}_F\boldsymbol{K}^T_F}{\sqrt{d_k}} \in \mathbb{R}^{9\times9} \tag{10}$$

Here, $\boldsymbol{Q}_{Pre} \in \mathbb{R}^{9\times d_k}$ and $\boldsymbol{K}_{Pre} \in \mathbb{R}^{9\times d_k}$ refer to the queries and keys for $\tilde{\boldsymbol{x}}^R_{Pre\_3}$, respectively, obtained by performing linear transformations on $\tilde{\boldsymbol{x}}^R_{Pre\_3}$, while $\boldsymbol{Q}_F$ and $\boldsymbol{K}_F$ are the corresponding queries and keys for $\tilde{\boldsymbol{x}}^R_{F\_3}$; the dimension of the keys (queries), denoted as d_k, is used for scaling the dot product.

Then, the similarity between $\boldsymbol{A}_{Pre}$ in the pre-training data and $\boldsymbol{A}_F$ in the fine-tuning data is measured using Maximum Mean Discrepancy (MMD):

$$L_{mmd} = \left\| \frac{1}{B}\sum_{i=1}^{B}\emptyset(\boldsymbol{A}^{(i)}_{Pre}) - \frac{1}{B}\sum_{j=1}^{B}\emptyset(\boldsymbol{A}^{(j)}_{F}) \right\|^2_H \tag{11}$$

where B stands for the number of samples in a training mini-batch, i and j are the indexes within the batch, $\boldsymbol{A}^{(i)}_{Pre}$ represents the correlation matrix of the i-th pre-training sample, $\boldsymbol{A}^{(j)}_{F}$ represents the spatial correlation matrix of the j-th fine-turning sample, and $\emptyset(\cdot)$ denotes the mapping function.

By doing so, we can quantify the distribution differences in attention representations between the impaired pre-training data and the fine-tuning data. Introducing this type of loss mitigates the feature disparities between different subjects while preserving the emotional characteristics inherent to the subject, thereby enhancing the model's classification performance and generalization ability. The attention mechanism is further used to obtain the aligned feature $\bar{\mathbf{x}}_{Pre\_3}$:

$$\bar{\mathbf{x}}^R_{Pre\_3} = Attn(\tilde{\mathbf{x}}^R_{Pre\_3}) = Softmax(\boldsymbol{A}_{Pre})\boldsymbol{V}_{Pre} \tag{12}$$

where $\boldsymbol{V}_{Pre}$ is the values obtained by performing a linear transformation on $\bar{\mathbf{x}}^R_{Pre\_3}$, and $Attn$ is the attention feature extractor. Finally, $\bar{\mathbf{x}}^R_{Pre\_3}$ is rearranged into $\bar{\mathbf{x}}_{Pre\_3} \in \mathbb{R}^{3\times3\times C_3}$ for the subsequent $2D$ masking with the size of 3×3.

At this point, $\bar{\mathbf{x}}_{Pre\_3}$ has better spatial features and prior knowledge compared to the initial data, and it also has some complementary relationship with the channel-level representation.

4) Encoder, Decoder, and Reconstruction: Based on the aforementioned embedding using different scales, we obtain the data $\tilde{\mathbf{x}}_{Pre\_1}$ and $\bar{\mathbf{x}}_{Pre\_3}$. We apply masks to these data based on different meanings of scale features. Then, we utilize a multi-layer Transformer encoder to extract features, followed by a decoder to reconstruct the images. The formula is shown below:

$$\mathbf{z}_{Pre\_1} = E_1(\tilde{\mathbf{x}}_{Pre\_1} \odot \tilde{\mathbf{M}}^{(1)}), \mathbf{x}'_{Pre\_1} = D_1(\mathbf{z}_{Pre\_1}) \tag{13}$$

$$\mathbf{z}_{Pre\_3} = E_3(\bar{\mathbf{x}}_{Pre\_3} \odot \tilde{\mathbf{M}}^{(3)}), \mathbf{x}'_{Pre\_3} = D_3(\mathbf{z}_{Pre\_3}) \tag{14}$$

where $\tilde{\mathbf{M}}^{(3)} \in \mathbb{R}^{3\times3}$ represents the random mask for $\bar{\mathbf{x}}_{Pre\_3}$, $\odot$ denotes the element-wise masking operation, and the mask values are broadcasted correspondingly. $\mathbf{z}_{Pre\_1}$ and $\mathbf{z}_{Pre\_3}$ are the masked data obtained through the encoder E_1 and E_3. $\mathbf{x}'_{Pre\_1}$ and $\mathbf{x}'_{Pre\_3}$ are the reconstructed data obtained through the decoders D_1, D_3. Then, we use the mean squared error (MSE) to measure the quality of the masked reconstruction. The reconstruction loss is computed only from masked non-zero patches to avoid introducing noise. Specifically, the formulas are as follows:

$$L_{recon\_1} = \frac{1}{|\Omega^1| \times C_f}\sum\nolimits_{(i,j)\in\Omega^1}\left\|\mathbf{x}_{Pre}(i,j) - \mathbf{x}'_{Pre\_1}(i,j)\right\|^2 \tag{15}$$

$$L_{recon\_3} = \frac{1}{|\Omega^3| \times C_3}\sum\nolimits_{(i,j)\in\Omega^3}\left\|\bar{\mathbf{x}}_{Pre\_3}(i,j) - \mathbf{x}'_{Pre\_3}(i,j)\right\|^2 \tag{16}$$

where Ω^1 represents the index set of the masked non-zero patches for $\mathbf{x}_{Pre}$, Ω^3 represents the index set of masked patches for $\bar{\mathbf{x}}_{Pre\_3}$, $|\cdot|$ denotes the number of elements in the set, (i, j) is the index of the masked patches, $\mathbf{x}'_{Pre\_1}(i,j) \in \mathbb{R}^{C_f}$, and $\mathbf{x}'_{Pre\_3}(i,j) \in \mathbb{R}^{C_3}$. We obtain the reconstruction losses, L_{recon_1} and L_{recon_3}, for two segments of different scales. For a mini-batch training dataset, the reconstruction losses can be expressed as $L^B_{recon\_1}$ and $L^B_{recon\_3}$.

D. Fine-Tuning Stage and Test Stage

After pre-training, generalized feature extractors E_{Pre_1} and E_{Pre_3} are obtained, which can be fine-tuned to obtain a personalized feature extractor $\hat{E}$, adapting it to a new task. However, certain modifications have been made when it comes to model transfer for EEG data.

In channel-level representation learning, in order to address the issue of zero-padding when mapping EEG data to two-dimensional brain images, we use channel masking during the pre-training stage to minimize the impact of zero-padding on the pre-training data. Similarly, during the fine-tuning stage, in the self-attention mechanism, a masking matrix can be used to assign a weight of 0 to the contribution of the padded regions in the attention weights. This effectively removes the influence of missing data on the attention weights and prevents the padded regions from interfering with the results.

Specifically, given the input $\mathbf{x}_F$, we obtain $\tilde{\mathbf{x}}_{F\_1}$ through the patch and positional embedding. Afterward, we calculate the corresponding attention matrix $\boldsymbol{A}_{chan} \in \mathbb{R}^{81\times 81}$ within the encoder E_1, where each element $A_{i,j}^{chan}$ is defined as:

$$A_{i,j}^{chan} = \frac{exp(e_{i,j} + M_{i,j}^{(F)})}{\sum_{k=1}^{n} exp(e_{i,k} + M_{i,k}^{(F)})} \tag{17}$$

where n represents the number of patches, $e_{i,j}$ represents the similarity score between the i-th and j-th patches, determined through the dot product of two vectors, and $M_{i,j}^{(F)}$ serves as a padding patch indicator. If either the value of the i-th or the j-th patch is missing (i.e., represented by a padding value), then $M_{i,j}^{(F)}$ is set to -∞ to eliminate their contribution to the attention matrix (i.e., $\lim_{M_{i,j}^{(F)}\to-\infty} exp(e_{i,j}+M_{i,j}^{(F)}) = 0$); otherwise, $M_{i,j}^{(F)}$ is set to 0.

In the pre-training phase for spatial-level representation, spatial feature alignment has already been performed through fine-tuning data. Therefore, the aligned data can be directly used for feature extraction in the encoder. To reduce the number of tuning parameters and enhance model stability, in this stage, we choose to freeze the pre-trained parameters of spatial-level representation without adjustments. This approach allows us to effectively leverage the previous pre-training results while avoiding issues such as overfitting during fine-tuning, thus improving the model's generalization ability. Finally, the features extracted from the channel-level representations and spatial-level representations, denoted as $\mathbf{z}_{F\_1}$ and $\mathbf{z}_{F\_3}$, respectively, are concatenated together and passed through a Batch Normalization layer to enhance the model's robustness and generalization ability. A classification layer is then incorporated into the fused features $\mathbf{z}_F$ for emotion classification, and we compute the classification loss using cross-entropy.

In the test stage, we employ a new session of EEG data from the specific subject, denoted as $\mathbf{x}_T^s$ and $\mathbf{y}_T^s$, to validate the effectiveness of the subject-specific model. The details of our proposed method are shown in Algorithm 1.

Algorithm 1 Multi-scale Masked Autoencoders

Input: Pre-training data $\mathbf{X}_{Pre} = \{\mathbf{x}_{Pre}^{(i)}\}_{i=1}^{N_{Pre}}$, fine-tuning data $\mathbf{X}_F^s = \{\mathbf{x}_F^{(i)}\}_{i=1}^{N_F}$ and labels $\mathbf{Y}_F^s = \{\mathbf{y}_F^{(i)}\}_{i=1}^{N_F}$ for a specific subject s; test data $\mathbf{X}_T^s = \{\mathbf{x}_T^{(i)}\}_{i=1}^{N_T}$ for the specific subject s; the number of epochs $Epoch$ and the batch size B.
Output: The generalized feature extractors E_{Pre_1} and E_{Pre_3} (include $Conv_3, Attn, and E_3$); the personalized emotion predictor $\hat{E}$; and the predicted emotion class $\hat{\mathbf{Y}}_T^s = \{\hat{\mathbf{y}}_T^{(i)}\}_{i=1}^{N_T}$.
Pre-training Stage for Channel-level Representation:
Randomly initialize E_{Pre_1}.
for $i = 1$: $Epoch$ **do**
  repeat
    Draw one batch of pre-training data $\mathbf{x}_{Pre}^B$.
    Embed the pre-training data $\mathbf{x}_{Pre}^B$ to obtain $\tilde{\mathbf{x}}_{Pre\_1}^B$.
    Mask the pre-training data and encode: $\mathbf{z}_{Pre\_1}^B = E_1(\tilde{\mathbf{x}}_{Pre\_1}^B \odot \tilde{\mathbf{M}}^{(1)})$.
    Reconstruct the input data: $\mathbf{x'}_{Pre\_1}^B = D_1(\mathbf{z}_{Pre\_1}^B)$.
    Optimize E_{Pre_1} by minimizing the reconstruction loss $L_{recon\_1}^B$.
  until all samples in $\mathbf{X}_{Pre}$ have been drawn.
Return E_{Pre_1}.
Pre-training Stage for Spatial-level Representation:
Randomly initialize E_{Pre_3}.
for $i = 1$: $Epoch$ **do**
  repeat
    Draw one batch of pre-training data $\mathbf{x}_{Pre}^B$ and one batch of fine-tuning data $\mathbf{x}_F^B$.
    Embed the input data $\mathbf{x}_{Pre}^B$ and $\mathbf{x}_F^B$ to obtain $\tilde{\mathbf{x}}_{Pre\_3}^B$ and $\tilde{\mathbf{x}}_{F\_3}^B$, and calculate their corresponding spatial correlation matrix $\boldsymbol{A}_{Pre}^B$ and $\boldsymbol{A}_F^B$.
    Align the space feature by minimizing the MMD loss $L_{mmd} = MMD(\boldsymbol{A}_{Pre}^B, \boldsymbol{A}_F^B)$.
    Optimize $Conv_3 and Attn$ by minimizing the reconstruction loss L_{mmd}.
  until all samples in $\mathbf{X}_{Pre}$ have been drawn.
Return $Conv_3 and Attn$.
Calculate $\bar{\mathbf{x}}_{Pre\_3}^B$ based on $Attn$: $\bar{\mathbf{x}}_{Pre\_3}^B = Attn(\tilde{\mathbf{x}}_{Pre\_3}^B)$.
for $i = 1$: $Epoch$ **do**
  repeat
    Mask the pre-training data and encode: $\mathbf{z}_{Pre\_3}^B = E_3\left(\bar{\mathbf{x}}_{Pre\_3}^B \odot \tilde{\mathbf{M}}^{(3)}\right)$.
    Reconstruct the input data: $\mathbf{x'}_{Pre\_3}^B = D_3(\mathbf{z}_{Pre\_3}^B)$.
    Optimize E_3 by minimizing the reconstruction loss $L_{recon\_3}^B$.
  until all samples in $\mathbf{X}_{Pre}$ have been drawn.
Return E_{Pre_3}.
Personalized Calibration Stage:
Initialize $\hat{E}$ with E_{Pre_1}, E_{Pre_3}, and frozen E_{Pre_3}.
for $i = 1$: $Epoch$ **do**
  repeat
    Draw one batch of fine-tuning data $\mathbf{x}_F^B$.
    Calculate $\mathbf{z}_{F\_1}^B$, $\mathbf{z}_{F\_3}^B$ based on E_{Pre_1} and E_{Pre_3}.
    Calculate the fused feature: $\mathbf{z}_F^B = \text{BatchNorm}([\mathbf{z}_{F\_1}^B, \mathbf{z}_{F\_3}^B])$.
    Predict the emotion class $\hat{\mathbf{y}}_F^B$ based on $\mathbf{z}_F^B$.
    Optimize $\hat{E}$ by minimizing the classification loss: $L_{cls} = CrossEntropy(\hat{\mathbf{y}}_F^B, \mathbf{y}_F^B)$.
  until all samples in $\mathbf{X}_F^s$ have been drawn.
Return $\hat{E}$.
Test stage:
Predict the emotion class: $\hat{\mathbf{Y}}_T^s = \hat{E}(\mathbf{X}_T^s)$.
Return $\hat{\mathbf{Y}}_T^s$.

TABLE I
PARAMETER DETAILS OF THE MODEL

Architecture Hyperparameters	Value		Optimization Hyperparameters	Value
	Channel-level	Spatial-level		
Encoder dim	256	512	Learning rate	0.001
Decoder dim	256	512	Batch size	32
Layer	2	2	Pre-train epoch	50
Heads	4	4	Fine-tune epoch	50
Mask	0.75	0.5	Optimizer	SGD

E. Implementation

Due to the different number of blocks in the channel-level and spatial-level representation learning stages, 9×9 and 3×3, respectively, different mask rates are established for each stage. Specifically, the mask rate for the channel-level representation learning stage is set as 0.75, in accordance with the original MAE [29], and for the spatial-level representation learning stage, it is adjusted to 0.5, due to the limited number of blocks.

The encoder and decoder parameters for channel-level and spatial-level representation learning are set identically for simplicity. Specifically, the dimensions for the encoder and decoder are chosen from {128, 256, 512, 1024}, with the number of layers selected from {1, 2, 3, 4}, and the number of self-attention heads from {2, 4, 6, 8}. MSMAE is optimized using the SGD optimizer with a learning rate of 0.001, 50 epochs, and a batch size of 32. The parameter settings are detailed in Table I.

IV. EXPERIMENT

A. Datasets

Experiments are performed on two publicly available datasets, namely SEED and SEED-IV. The SEED dataset includes EEG signals from 15 subjects, which are recorded using an ESI NeuroScan System with 62 channels [31]. Each subject participates in three sessions, with an interval of approximately one week between sessions. During these sessions, the subjects' data are collected while watching emotion-eliciting movies designed to evoke three different emotional states: negative, positive, or neutral. The signals are initially recorded at a sampling rate of 1000Hz and are subsequently downsampled to 200Hz for analysis. They are further segmented into non-overlapping 1-second segments, with each segment treated as a sample. Consequently, for each subject and each session, there is a total of 3,394 samples.

The SEED-IV dataset consists of EEG signals of 15 subjects recorded using the same recording device as SEED [32]. Similar to the SEED dataset, each subject participates in three separate sessions with intervals between them. In this case, four emotional states are collected: happiness, sadness, fear, and neutral. The signals are divided into 4-second non-overlapping segments, and each segment is regarded as an individual sample. Consequently, for Sessions I, II, and III, there are 851, 832, and 822 samples per subject, respectively.

B. Data Preprocessing

To construct a unified pre-training model, it is necessary to preprocess all the data in a consistent manner. Firstly, based on the structure of the EEG cap, the EEG channels of each frame are mapped into a two-dimensional EEG image to preserve the spatial location of the electrodes, as shown in Fig. 1. This transformation is applied to frequency-domain features for each sample. We employ Differential Entropy (DE) for the frequency-domain feature, which is widely used in emotion recognition [31]. Specifically, DE features are derived from five predefined frequency bands, which include Delta (1-3 Hz), Theta (4-7 Hz), Alpha (8-13 Hz), Beta (14-30 Hz), and Gamma (31-50 Hz). Additionally, min-max normalization is performed at the sample level to address the issue of varying feature ranges, improve the convergence performance of the model, and eliminate the dimensional differences between different features.

TABLE II
AVERAGE ACCURACIES (%) AND STANDARD DEVIATIONS (%) ON SEED AND SEED-IV DATASETS USING WITHIN-SUBJECT CROSS-SESSION CROSS-VALIDATION

Model	Acc±Std	
	SEED	SEED-IV
Vit [37]	70.10±7.17	47.43±12.22
SimpleVit [33]	73.60±7.27	45.82±11.53
FBCCNN [34]	74.79±7.69	49.90±13.90
STNET [35]	75.62±7.12	58.07±11.89
DGCNN [36]	78.02±7.51	56.59±12.86
Ours	**80.86±6.21**	**59.33±12.61**

C. Cross-Session Evaluation

Compared to other datasets, the SEED and SEED-IV datasets possess unique characteristics in that each subject completed the experiment in three different sessions. We utilize this distinctive feature to investigate the generalization of models across sessions, specifically assessing whether the models can consistently deliver satisfactory performance when training and testing data come from different sessions. When receiving the same stimuli, the recognition accuracy of various methods for predicting the emotions of the same subject at different times will show temporal stability variations. However, up to the present, there have been limited studies on cross-session experiments, most of which involve the acquisition of test data to minimize the data distribution discrepancy with the training data during the training process. In contrast, our experimental setups do not require the inclusion of test data during training. This approach, while more challenging, offers enhanced practical value. Specifically, we use one session's EEG data as training data and another as testing data. The pairs of sessions used for validation encompass session1-session3, session2-session1, session3-session2, session1-session2, session2-session3, and session3- session1.Through a comprehensive six-fold cross-validation, we calculate the average recognition accuracy, along with the standard deviation, for all 15 subjects.

D. Method Comparisons

We compare the proposed MSMAE with several relevant models on the SEED and SEED-IV datasets to demonstrate

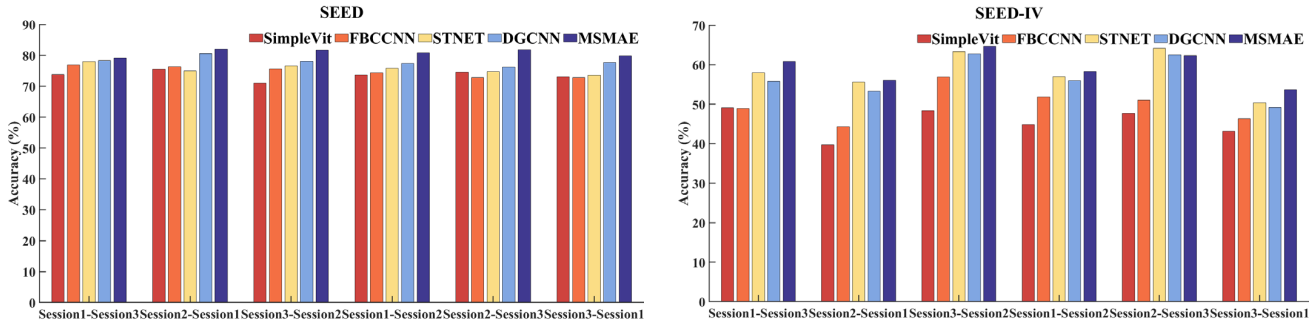

Fig. 3. Comparison between MSMAE and other algorithms in various cross-session scenarios within SEED and SEED-IV.

TABLE III
ABLATION STUDY OF OUR MODEL WITHIN SEED AND SEED-IV

Method	Acc±Std	
	SEED	SEED-IV
Vit (1×1)	70.10±7.17	47.43±12.22
Vit (3×3)	71.82±7.63	48.95±12.32
MAE (1×1)	71.73±6.65	44.02±11.79
MAE (3×3)	73.46±7.80	50.45±11.93
Vit (1×1&3×3)	71.54±7.73	54.52±11.55
MAE(1×1&3×3)	77.41±6.65	56.55±11.79
Ours	**80.86±6.21**	**59.33±12.61**

its effectiveness. We select models focusing on spatial features to ensure a more meaningful comparison, including Vit [37], SimpleVit [33], FBCCNN [34], STNET [35], and DGCNN [36]. Furthermore, we implement these models using TorchEEG, a PyTorch-based library for EEG signal analysis. We search the parameter space of these compared models following the descriptions outlined in their respective papers. The average accuracies (± standard deviation) for each method are reported in Table II. The experimental results demonstrate that our method significantly outperforms existing methods. Specifically, as shown in Table II, our model achieves a recognition accuracy of 80.86% on the SEED dataset with a standard deviation of only 6.21%. On the SEED-IV dataset, our model achieves a recognition accuracy of 59.33%, accompanied by a standard deviation of 12.61%. Furthermore, according to the results in Fig. 3, our method exhibits performance improvement across different session-to-session transfers. Even without utilizing the target domain, our model can reduce the influence between different domains by aligning regional features. Additionally, as illustrated in Fig. 4, our model demonstrates advantages for each subject, indicating the generalization and stability of our model.

E. Ablation Study

To evaluate the effectiveness of each module in the MSMAE model, we conduct ablation experiments with the MAE model and the Vit model at different scales, namely (1 × 1) and (3 × 3). We also compare the results with the feature fusion of both methods at scale (1 × 1&3 × 3), which are listed in Table III. By comparing the experimental results at different scales, it is observed that the results at the 3 × 3 scale consistently outperform those at the 1 × 1 scale, indicating the advantage of spatial frequency features in EEG emotion recognition tasks. Furthermore, by comparing the results of the Vit and MAE models on 1 × 1 scale on the SEED-IV dataset with relatively limited pre-training and fine-tuning data, we find that using MAE for pre-training a large model tends to lead to overfitting and a decrease in accuracy compared to Vit without pre-training. However, such pre-trained models are highly dependent on the volume of data, as the performance of the model largely relies on the quality and diversity of the data used during the training process. Building upon this foundation, we further enhance the model's performance by fusing multi-scale features and conducting pre-training. Importantly, our model achieves higher stability and generalization performance by aligning the region correlations between the pre-training and fine-tuning data. Through these ablation experiments, we validate the importance of scale selection, pre-training, and multi-scale feature fusion in our model. These results provide strong support for our research and application in complex EEG emotion recognition tasks and offer valuable directions for future improvements and optimizations.

We randomly select one subject from the SEED dataset for visualization. The t-SNE visualization of different methods is presented in Fig.5. In comparison to other methods, MSMAE demonstrates a reduction in data distribution discrepancy to some extent, even without utilizing target domain information.

F. Interpretability

To validate the interpretability of our proposed method, we conduct EEG topographic visualization using adjacency matrices at a scale of 1 × 1 learned from MSMAE. Followed by [38] and [39], we visualize the degree centrality of each scalp EEG electrode based on the adjacency matrices. Suppose $\tilde{\boldsymbol{A}} = \{\tilde{A}_{i,j}\}_{i,j=1}^{p}$ is the submatrix of $\boldsymbol{A}_{chan} \in \mathbb{R}^{81\times 81}$, where p represents the number of non-zero padded patches in channel-level representation (with $p = 62$ in the SEED dataset). In this matrix, the i-th row and i-th column values correspond to the connection weights associated with the i-th electrode. The degree centrality of the i-th EEG electrode, denoted as DC_i, can be derived by

$$DC_i = \sum_{n=1}^{p} \tilde{A}_{i,n} + \sum_{m=1}^{p} \tilde{A}_{m,i} - 2\tilde{A}_{i,i} \ (i = 1, \ldots, p) \tag{18}$$

Fig. 6 presents the EEG topographic maps of positive, neutral, and negative emotions in the SEED dataset. The values of DC are scaled to the interval of [0, 1]. Through scalp mapping visualization, we can gain a direct and intuitive understanding

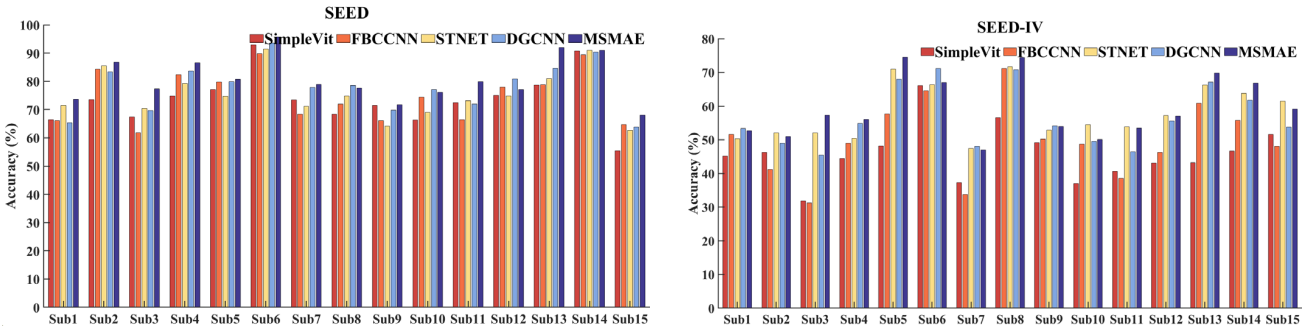

Fig. 4. Comparison of MSMAE and other algorithms on different subjects within SEED and SEED-IV.

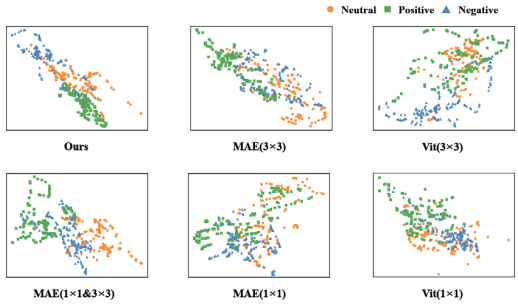

Fig. 5. Feature visualization by different methods and at different scales within SEED dataset.

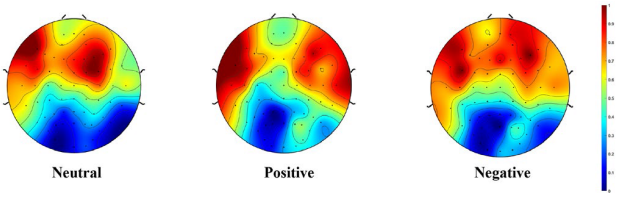

Fig. 6. Topographic maps learned from the MSMAE model within the SEED dataset.

of the spatial distribution of the emotion recognition task, which reflects the intercorrelation analysis of EEG signals between electrodes in our method. By examining Fig. 6, we observe that the regions of emotional activity are primarily concentrated in the frontal and temporal areas. These findings from saliency maps have been validated and are consistent with existing research on emotions [40], [41], [42]. Furthermore, we note that in neutral emotions, the neural patterns are relatively smoother compared to positive and negative emotions. Positive emotions are more readily activated in the lateral temporal areas compared to negative and neutral emotions, consistent with the finding in [31]. In addition, we observe that the activation range of negative emotions is larger in the frontal regions.

G. Cross-Dataset Generalization

We perform cross-dataset experiments to assess the generalization ability of our model. We chose the unlabeled data from the latest publicly available dataset, FACED [43], as the pre-training data. This dataset contains EEG signals from 123 subjects with 32 channels. Given that the SEED and SEED-IV datasets lack the A1 and A2 electrodes, we exclude these channels and remain 30 channels for our analysis. We fine-tune the model with data from one session of a specific subject from the SEED or SEED-IV dataset and test the model on another session of the same subject. The challenge of cross-dataset experiments is that pre-training is conducted using

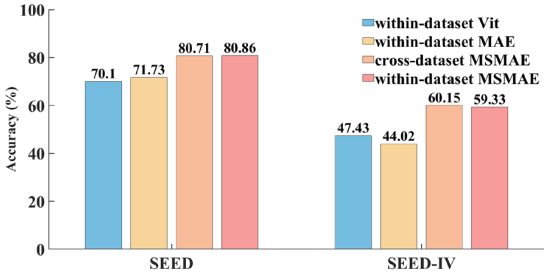

Fig. 7. Comparison of cross-dataset and within-dataset accuracy.

unlabeled data with 30 channels, whereas fine-tuning used 62-channel data from the SEED or SEED-IV dataset, which resulted in missing channels and differences between devices. Notably, in our cross-dataset and with-dataset settings, the only difference lies in whether the pre-training data originates from the same dataset as the fine-tuning data.

We compare the performance of MSMAE under the cross-dataset and within-dataset settings. Additionally, Vit (1 × 1) and MAE (1 × 1) under the within-dataset setting are also included for comparison, as depicted in Fig. 7. Based on the experimental results, our model demonstrates consistent and stable generalization ability in the cross-dataset setting. Furthermore, it confirms our model's capability to address the issue of missing channels, validating the robustness and portability of our model.

V. Conclusion

This paper introduces a unified, multi-scale pre-training framework to overcome challenges related to missing EEG channels and limited labeled data in emotion recognition. We propose a novel multi-scale fusion approach combining channel-level and spatial-level representation learning with an improved masking mechanism to preserve electrode relationships and invariance learning for regional correlations. Compared to the Vit (1 × 1) without pre-training, MSMAE significantly improves accuracy by 10.76% on the SEED dataset and 11.9% on the SEED-IV dataset. Moreover, MSMAE surpasses the original MAE (1 × 1) in accuracy by 9.13% on the SEED dataset and by 15.31% on the SEED-IV dataset. MSMAE also demonstrates superiority over current state-of-the-art methods, outperforming the second-best method by 2.84% and 1.26% on the SEED and SEED-IV datasets.

In summary, the proposed model significantly elevates the performance of cross-session emotion recognition in a self-supervised fashion. MSMAE is a general framework that can

be easily extended to other EEG-based learning tasks, offering promising directions for future research. However, the current implementation of MSMAE relies on handcrafted features as input, potentially resulting in the loss of valuable information in the original signals. Consequently, our future efforts will explore MSMAE's potential for directly extracting information from raw signals, addressing this constraint, and enhancing the framework's utility.

REFERENCES

[1] U. Tariq, J. Yang, and T. S. Huang, "Supervised super-vector encoding for facial expression recognition," *Pattern Recognit. Lett.*, vol. 46, pp. 89–95, Sep. 2014.

[2] D. Lottridge, M. Chignell, and A. Jovicic, "Affective interaction: Understanding, evaluating, and designing for human emotion," *Rev. Human Factors Ergonom.*, vol. 7, no. 1, pp. 197–217, Sep. 2011.

[3] B. A. Taheri, R. T. Knight, and R. L. Smith, "A dry electrode for EEG recording," *Electroencephalogr. Clin. Neurophysiology*, vol. 90, no. 5, pp. 376–383, May 1994.

[4] S. M. Alarc ao and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 374–393, Jul. 2019.

[5] V. Jayaram, M. Alamgir, Y. Altun, B. Scholkopf, and M. Grosse-Wentrup, "Transfer learning in brain-computer interfaces," *IEEE Comput. Intell. Mag.*, vol. 11, no. 1, pp. 20–31, Feb. 2016.

[6] J. Li, S. Qiu, Y.-Y. Shen, C.-L. Liu, and H. He, "Multisource transfer learning for cross-subject EEG emotion recognition," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3281–3293, Jul. 2020.

[7] Y. Li, W. Zheng, L. Wang, Y. Zong, and Z. Cui, "From regional to global brain: A novel hierarchical spatial–temporal neural network model for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 568–578, Apr. 2022.

[8] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1290–1301, Jul. 2022.

[9] C. Chen, C.-M. Vong, S. Wang, H. Wang, and M. Pang, "Easy domain adaptation for cross-subject multi-view emotion recognition," *Knowl.-Based Syst.*, vol. 239, Mar. 2022, Art. no. 107982.

[10] O. Ozdenizci, Y. Wang, T. Koike-Akino, and D. Erdogmus, "Adversarial deep learning in EEG biometrics," *IEEE Signal Process. Lett.*, vol. 26, no. 9, pp. 710–714, May 2019.

[11] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdogmus, "Learning invariant representations from EEG via adversarial inference," *IEEE Access*, vol. 8, pp. 27074–27085, 2020.

[12] D. Bethge et al., "Domain-invariant representation learning from EEG with private encoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1236–1240.

[13] E. T. Bullmore et al., "Fractal analysis of electroencephalographic signals intracerebrally recorded during 35 epileptic seizures: Evaluation of a new method for synoptic visualisation of ictal events," *Electroencephalogr. Clin. Neurophysiology*, vol. 91, no. 5, pp. 337–345, Nov. 1994.

[14] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 186–197, Mar. 2010.

[15] L. I. Goldfischer, "Autocorrelation function and power spectral density of laser-produced speckle patterns," *J. Opt. Soc. Amer.*, vol. 55, no. 3, pp. 247–253, Mar. 1965.

[16] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. 6th Int. IEEE/EMBS Conf. Neural Eng. (NER)*, Nov. 2013, pp. 81–84.

[17] V. Gupta, M. D. Chopda, and R. B. Pachori, "Cross-subject emotion recognition using flexible analytic wavelet transform from EEG signals," *IEEE Sensors J.*, vol. 19, no. 6, pp. 2266–2274, Mar. 2019.

[18] A. Anuragi, D. S. Sisodia, and R. B. Pachori, "EEG-based cross-subject emotion recognition using Fourier-bessel series expansion based empirical wavelet transform and NCA feature selection method," *Inf. Sci.*, vol. 610, pp. 508–524, Sep. 2022.

[19] A. Nalwaya and R. B. Pachori, "Fourier–Bessel domain adaptive wavelet transform-based method for emotion identification from EEG signals," *IEEE Sensors Lett.*, vol. 8, no. 2, pp. 1–4, Feb. 2024.

[20] A. Bhattacharyya, L. Singh, and R. B. Pachori, "Fourier–Bessel series expansion based empirical wavelet transform for analysis of non-stationary signals," *Digit. Signal Process.*, vol. 78, pp. 185–196, Jul. 2018.

[21] X. Liu et al., "Emotion recognition and dynamic functional connectivity analysis based on EEG," *IEEE Access*, vol. 7, pp. 143293–143302, 2019.

[22] C. Chen, Z. Li, F. Wan, L. Xu, A. Bezerianos, and H. Wang, "Fusing frequency-domain features and brain connectivity features for cross-subject emotion recognition," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022.

[23] S. Alhagry, A. Aly, and R. A. El-Khoribi, "Emotion recognition based on EEG using LSTM recurrent neural network," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 10, pp. 355–358, 2017.

[24] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 839–847, Mar. 2019.

[25] H. Chen, M. Jin, Z. Li, C. Fan, J. Li, and H. He, "MS-MDA: Multisource marginal distribution adaptation for cross-subject and cross-session EEG emotion recognition," *Frontiers Neurosci.*, vol. 15, Dec. 2021, Art. no. 778488.

[26] J. Li, S. Qiu, C. Du, Y. Wang, and H. He, "Domain adaptation for EEG emotion recognition based on latent representation similarity," *IEEE Trans. Cognit. Develop. Syst.*, vol. 12, no. 2, pp. 344–353, Jun. 2020.

[27] B.-Q. Ma, H. Li, W.-L. Zheng, and B.-L. Lu, "Reducing the subject variability of EEG signals with adversarial domain generalization," in *Proc. Int. Conf. Neural Inf. Process.*, 2019, pp. 30–42.

[28] R. Li, Y. Wang, W.-L. Zheng, and B.-L. Lu, "A multi-view spectral-spatial-temporal masked autoencoder for decoding emotions with self-supervised learning," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 6–14.

[29] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15979–15988.

[30] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[31] W. Zheng, J. Zhu, and B. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 417–429, Jul. 2019.

[32] W. Zheng, W. Liu, Y. Lu, B. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.

[33] L. Beyer, X. Zhai, and A. Kolesnikov, "Better plain ViT baselines for ImageNet-1k," 2022, *arXiv:2205.01580*.

[34] B. Pan and W. Zheng, "Emotion recognition based on EEG using generative adversarial nets and convolutional neural network," *Comput. Math. Methods Med.*, vol. 2021, pp. 1–11, Oct. 2021.

[35] Z. Zhang, S.-H. Zhong, and Y. Liu, "GANSER: A self-supervised data augmentation framework for EEG-based emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2048–2063, Jul./Sep. 2023.

[36] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, Jul. 2020.

[37] A. Dosovitskiy et al., *An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale*. Ithaca, NY, USA: Cornell Univ., 2020.

[38] T. Song, S. Liu, W. Zheng, Y. Zong, and Z. Cui, "Instance adaptive graph for EEG emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 3, 2020, pp. 2701–2708.

[39] M. Ye, C. L. P. Chen, and T. Zhang, "Hierarchical dynamic graph convolutional network with interpretability for EEG-based emotion recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 9, 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9979692, doi: 10.1109/TNNLS.2022.3225855.

[40] S. K. Hadjidimitriou and L. J. Hadjileontiadis, "Toward an EEG-based recognition of music liking using time-frequency analysis," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 12, pp. 3498–3510, Dec. 2012.

[41] P. Li et al., "EEG based emotion recognition by combining functional connectivity network and local activations," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 10, pp. 2869–2881, Oct. 2019.

[42] M. Sun, W. Cui, S. Yu, H. Han, B. Hu, and Y. Li, "A dual-branch dynamic graph convolution based adaptive TransFormer feature fusion network for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2218–2228, Oct. 2022.

[43] J. Chen, X. Wang, C. Huang, X. Hu, X. Shen, and D. Zhang, "A large finer-grained affective computing EEG dataset," *Sci. Data*, vol. 10, no. 1, Oct. 2023, Art. no. 740.