MetaEmotionNet: Spatial-Spectral-Temporal based Attention 3D Dense Network with Meta-learning for EEG Emotion Recognition

Xiaojun Ning, Jing Wang, *Member, IEEE*, Youfang Lin, Xiyang Cai, Haobin Chen, Haijun Gou, Xiaoli Li, *Senior Member, IEEE*, and Ziyu Jia

***Abstract*—Emotion recognition has become an important area in affective computing. Emotion recognition based on multi-channel electroencephalogram (EEG) signals has gradually become popular in recent years. However, on the one hand, how to make full use of different EEG features and the discriminative local patterns among the features for various emotions is challenging. Existing methods ignore the complementarity among the spatial-spectral-temporal features and discriminative local patterns in all features, which limits the classification performance. On the other hand, when dealing with cross-subject emotion recognition, existing transfer learning methods need a lot of training data. At the same time, it is extremely expensive and time-consuming to collect the labeled EEG data, which is not conducive to the wide application of emotion recognition models for new subjects. To solve the above challenges, we propose a novel spatial-spectral-temporal based attention 3D dense network with meta-learning, named MetaEmotionNet, for emotion recognition. Specifically, MetaEmotionNet integrates the spatial-spectral-temporal features simultaneously in a unified network framework through two-stream fusion. At the same time, the 3D attention mechanism can adaptively explore discriminative local patterns. In addition, a meta-learning algorithm is applied to reduce dependence on training data. Experiments demonstrate that the MetaEmotionNet is superior to the baseline models on two benchmark datasets.**

***Index Terms*—Electroencephalogram, Emotion recognition, Meta-learning, Attention mechanism, Affective computing.**

I. Introduction

EMOTION has an important influence on human behavior and psychology. Meanwhile, many mental diseases like autism and depression are related to emotions [1], [2]. Hence, emotion is often adopted as a reference for assessing patients' mental diseases [3]. More and more researchers focus on the analysis of EEG for different emotions induced by specific stimulation patterns, and develop emotional artificial intelligence in human-computer interaction (HCI) [4], [5]. Generally speaking, different emotions can be recognized in various ways, such as human facial expressions [6], [7], language expressions [8], [9], body movements [10], [11], and EEG signals [12], [13]. However, the first three ways are susceptible to subjective influences of the participants, that is, participants can deliberately disguise their emotions. On the contrary, EEG signals can objectively reflect the state of the human body and become a reliable approach to recognize specific emotions [14]. Therefore, emotion recognition based on EEG signals has gradually attracted the attention of researchers.

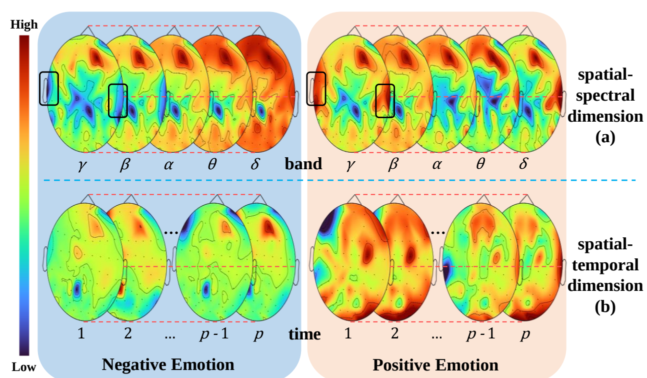


Fig. 1. The distribution of brain activation in the (a) spatial-spectral dimension and (b) spatial-temporal dimension. Different emotion states show different levels of activation in brain regions, frequency bands, and time stamps. Moreover, like the example shown in the black boxes, there exist discriminative local patterns for different emotion states, which presents that the β and γ bands in the temporal lobe of positive emotions are more activated than negative emotions.

Although previous attempts at EEG emotion recognition have achieved high classification performance [15]–[19], there are still some challenges remain open:

On the one hand, the existing approaches have not made full use of different EEG features and the discriminative local patterns among the features for different emotions simultaneously. As illustrated in Fig. 1, the spatial-spectral-temporal features of EEG signals are different for different emotions. For example, in the spatial-spectral dimension (top row of Fig. 1), when positive emotion is produced, the higher frequency band (β and γ) is more activated than the lower frequency band (δ and θ), while negative emotions are just the opposite.

Manuscript received September 21, 2023. This work was partially supported by the National Natural Science Foundation of China (Grant No. 62306317 and 61603029) and China Postdoctoral Science Foundation (Grant No. 2023M733738). *(Corresponding author: Ziyu Jia.)*

Xiaojun Ning, Jing Wang, Youfang Lin, Haobin Chen, and Haijun Gou are with the Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: ningxj@bjtu.edu.cn; wj@bjtu.edu.cn; yflin@bjtu.edu.cn; hbchen@bjtu.edu.cn; haijungou@bjtu.edu.cn).

Xiyang Cai is with the Samueli School Of Engineering, University of California, Los Angeles, CA 90095, USA (e-mail: caixiyang@ucla.edu).

Xiaoli Li is with the Institute for Infocomm Research, A*STAR, 138632, Singapore, and the School of Computer Science and Engineering, Nanyang Technological University, 308232, Singapore (e-mail: xlli@i2r.a-star.edu.sg).

Ziyu Jia is with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China. Ziyu Jia is also with Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jia.ziyu@outlook.com).

Meanwhile, in the spatial-temporal dimension (bottom row of Fig. 1), the overall activation of the brain in negative emotion is low, while positive emotion is on the contrary. Moreover, the discriminative local patterns are also important to recognize different emotions. For example, when a person is in a positive emotional state, the degree of activation in the β and γ bands of temporal lobe regions (locate behind the ears and extends to both sides of the brain, as marked in Fig. 1 by black boxes) is higher than the negative emotion [14]. Inspired by these phenomena, there are some attempts to identify different emotions based on EEG features. For example, recurrent neural network [17], [20] is adopted to capture temporal features. Convolutional neural network [21], [22] or graph neural network [23], [24] is utilized to capture spatial-spectral features. However, it can be noticed that most existing models only use a single feature or a combination of two features (such as spatial and spectral features) [25], [26], which does not make full use of the complementarity among the spatial-spectral-temporal features. Therefore, the classification performance of these models is limited to a certain extent. Furthermore, how to capture local patterns in different dimensional features also makes EEG emotion recognition challenging.

On the other hand, existing deep learning-based emotion recognition methods rely on a large amount of labeled data [26]. Specifically, the EEG signal patterns of different subjects are not consistent due to subject differences. This makes the classification accuracy of the model on new subjects terrible [27], which is not conducive to the wide application of emotion recognition models. To attack this problem, many transfer learning methods are used in EEG emotion recognition to improve the performance of the model on new subjects. For example, domain adaptation [28], a typical transfer learning method, needs sufficient training on the source subjects, and then uses the data distribution of the target subjects to adapt. However, existing transfer learning methods for emotion recognition still require a large amount of training data and calibration data [27], [29]. This reduces the efficiency of the models and is not suitable for the actual scene. How to reduce data dependence while achieving fast adaptation to new subjects has become another challenge.

To tackle the above challenges, we propose a Spatial-Spectral-Temporal based Attention 3D Dense Network with Meta-learning named *MetaEmotionNet*. The MetaEmotionNet includes the feature extraction network and meta-learning method. Specifically, for feature extraction, we design a Spatial-Spectral-Temporal based Attention 3D Dense Network (SST-Net) which consists of the spatial-spectral stream and spatial-temporal stream. Each stream is made up of several Attention based 3D Dense Blocks (A3DB) and Transition Layers. Meanwhile, we apply the meta-learning method for low data dependence and fast adaption. Fig. 2 presents the whole process of the experiment. When subjects are stimulated by multimedia material, their EEG signals could be constructed as 3D representations for input into MetaEmotionNet. In the MetaEmotionNet algorithm, the SST-Net for 3D representations extracts spatial-spectral-temporal features, and the meta-learning strategy ensures low data dependency and fast adaptation. Finally, MetaEmotionNet outputs the probability values of various emotions to achieve classification. The main contributions of this paper are summarized as follows:

- We propose a novel two-stream 3D Dense network, which fuses the spatial-spectral-temporal information of EEG signals in a unified network framework based on the constructed 3D EEG representation.
- We develop a parallel Spatial-Spectral/Temporal attention mechanism to adaptively capture discriminative patterns in brain regions, frequency bands, and time stamps.
- We propose a meta-learning method for EEG emotion recognition, which reduces the dependence on training data and achieves fast adaptation to new subjects.
- Extensive experiments are conducted on two benchmark datasets and the experimental results show that our MetaEmotionNet outperforms all state-of-the-art models.

Compared to the SST-EmotionNet published in our preliminary work [25], MetaEmotionNet has the following important improvements:

- We employ a meta-learning method to reduce the need for training samples and achieve fast adaptation for unknown subjects.
- We adopt cross-subject emotion recognition experiments to evaluate the performance of MetaEmotionNet.
- We conduct more ablation experiments to evaluate the impact of each component in MetaEmotionNet on the performance.
- We explore and discuss the interpretability of the attention mechanism in MetaEmotionNet.

II. Related Work

More and more researchers are devoted to the analysis of human physiological signals [30]–[35]. EEG data is a typical time series data that can objectively reflect our real emotions and is gradually applied to emotion recognition tasks [36], [37]. At the beginning, traditional machine learning classifiers with manually extracted features are designed for emotion recognition. For instance, Wang *et al.* [38] adopt support vector machine to identify specific emotions. Bahari *et al.* [39] adopt recurrence plot analysis and non-linear k nearest neighbor classifier (kNN) to recognize various emotions. However, the design and selection of hand-crafted features are built on a wealth of expert knowledge, which limits the performance and extension of traditional machine learning technology.

Driven by the achievements of deep learning in computer vision, speech recognition, and natural language translation [40], researchers also use deep learning for EEG emotion recognition. Specifically, researchers intend to extract discriminative features from different dimensions of EEG signals, such as frequency features, temporal features, and spatial features, to improve the performance of emotion recognition. For **frequency features**, power spectral density (PSD) [41], [42] and differential entropy (DE) [43], [44] features have been shown to be effective and stable in emotion recognition. Al-Nafjan *et al.* [4] employ fast Fourier transform based PSD features to recognize different emotions. Zheng *et al.* [44] present a deep belief network to recognize emotions applying

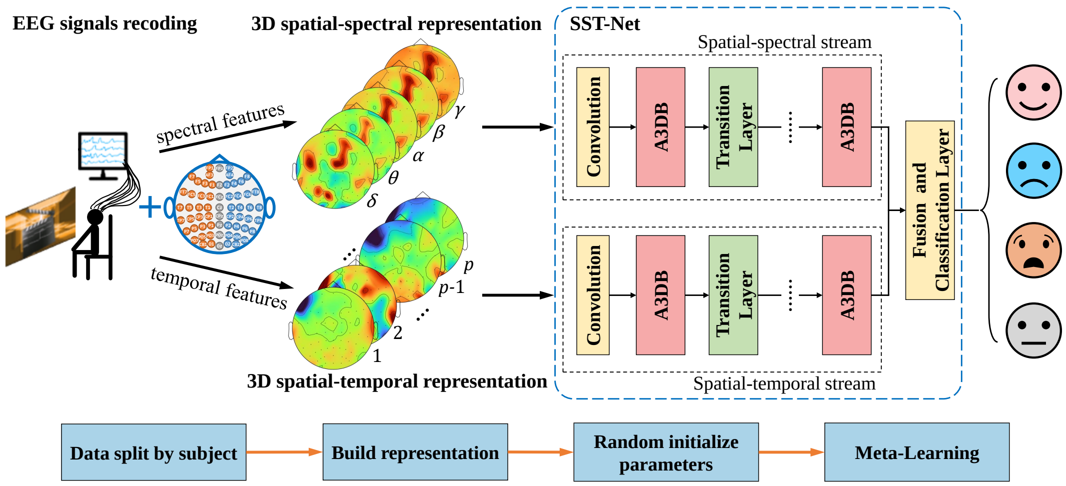


Fig. 2. The entire process of EEG-based emotion recognition under multimedia stimuli. When subjects were stimulated by multimedia material, their EEG signals were constructed as 3D representations for input into MetaEmotionNet. MetaEmotionNet consists of spatial-spectral-temporal feature extraction and meta-learning strategy. SST-Net consists of spatial-spectral stream and spatial-temporal stream with the same structure. At the end of the SST-Net, spatial-spectral stream and spatial-temporal stream are fused for classification. Meanwhile, in order to reduce the usage of training samples, meta learning is used to update the weight of SST-Net.

DE features from multiple frequency bands. Yang *et al.* [45] present a hierarchical network with subnetwork nodes based on DE features. Gao *et al.* [18] utilize DE features matrices with a Channel-fused Dense Convolutional Network (CDCN) for emotion recognition and achieve stable performance. For **temporal features** extraction, some efforts have tried to use deep neural networks to extract valuable information from raw EEG. For example, Fourati *et al.* [46] propose an Echo State Network, which apply recursive layer encodes the EEG signals to extract deep temporal features. Alhagry *et al.* [20] input the EEG signal into two LSTM layers and achieve satisfactory recognition performance. Combining LSTM layer with residual structure, Ma *et al.* [17] present a multimodal residual LSTM model (MMResLSTM), which can efficiently learn emotion-related high-level features. To extract **spatial features**, Li *et al.* [21] construct 2D EEG maps and utilize a hierarchical neural network to learn spatial features. Zhang *et al.* [15] present spatial-temporal recurrent neural network (STRNN). STRNN utilizes a multidirectional RNN layer to extract spatial clues by traversing EEG channels in different directions. Jung *et al.* [22] use the image-based EEG representation to capture the brain interaction information. Li *et al.* [47] design a bi-hemispheric discrepancy method to model the asymmetric differences of different emotions. Zhang *et al.* [24] propose a Graph Convolutional Broad Network for exploring the graph-structured brain data. Although the above attempts to recognize emotions have achieved satisfactory results, few methods are comprehensively considering frequency features, temporal features, and spatial features. Most methods only consider one or two features to model EEG signals, ignoring the complementary of the spatial-spectral-temporal features.

In addition, due to the subject differences, the performance of the emotion recognition models is greatly challenged in the cross-subject setting. Some researchers discover that despite the differences existing objectively, the emotion features among different subjects are still similar. Hence, some researchers employ **transfer learning** method to enhance the generalization performance of the model. Chai *et al.* [48] propose adaptive subspace feature matching (ASFM) to improve cross-subject performance, which uses principal component analysis (PCA) and iterative pseudo-label refinement strategy to reduce the difference of data distribution between different subject. Lin and Jung [49] adopt a conditional TL (cTL) framework to promote positive transfer among different subjects by conditional control and feature selection. **Deep transfer learning** also is also applied to cross-subject emotion recognition. Chai *et al.* [50] present a subspace alignment autoencoder (SAAE) which uses stacked auto encoders to map features to domain invariant subspaces. Graph regularization, kernel PCA, and maximum mean discrepancy are also employed to reduce the difference of feature distribution. Fahimi *et al.* [51] employ cross-subject attention to classify the emotion. They use source subjects to train a CNN and then use some calibration data of the target subject to fine-tune the network. Li *et al.* [52] propose bi-hemisphere domain adversarial neural network for cross-subject tasks. It uses two local domain discriminators to extract the features of two hemispheres adversarially, and a global discriminators with Gradient Reversal Layer (GRL) to reduce subject differences. Recently, Zhao *et al.* [53] perform a plug-and-play domain adaptation method for cross-subject emotion recognition. It uses few amount of unlabeled EEG signals from unknown subjects to adjust the model. However, most transfer learning methods still need a large amount of source data, and they may lead to negative migration [54]. In recent years, **meta-learning** is proposed to solve few-shot learning difficulties [55]–[57]. Only a small number of physiological signal-based work attempts to apply the meta-learning method. Banluesombatkul *et al.* [58] propose a MetaSleepLearner method, which applies Model-Agnostic Meta-Learning (MAML) algorithm to sleep stage classification and realizes the fast adaptation to new subjects. However, the vanilla MAML method is hard to

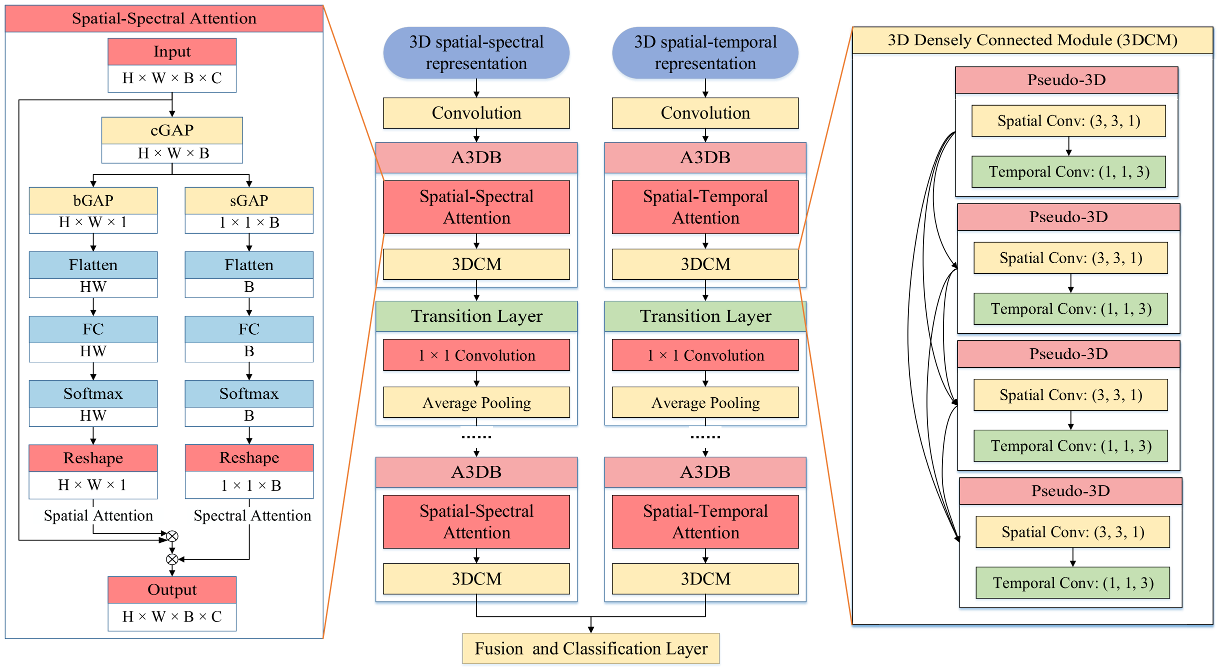


Fig. 3. Overall architecture of Spatial-Spectral-Temporal based Attention 3D Dense Network (SST-Net). SST-Net consists of two streams: Spatial-Spectral stream and Spatial-Temporal stream.

optimize and consumes a lot of computation [59]. Inspired by them, we design a more efficient meta-learning method for emotion recognition to improve the ability of the model to adapt to unknown subjects.

III. Preliminaries

In this paper, we first denote several necessary notations that can be helpful to understand the following equations in Section IV, and then present the EEG emotion recognition problem statement.

We define $\mathbf{X}^T = (\mathbf{S}_1^T, \mathbf{S}_2^T, \ldots, \mathbf{S}_P^T) \in \mathbb{R}^{E \times P}$ as an EEG signal sample containing P time stamps, where E is the number of electrodes, $\mathbf{S}_p^T = (s_p^1, s_p^2, \ldots, s_p^E) \in \mathbb{R}^E (p \in \{1, 2, \ldots, P\})$ represents the EEG signals of all E electrodes collected at time stamp p. Through some transformation operations (see Section 4.1), the 3D spatial-temporal representation $\mathbf{X}^{T'} = (\mathbf{M}_1^T, \mathbf{M}_2^T, \ldots, \mathbf{M}_P^T) \in \mathbb{R}^{H \times W \times P}$ of EEG signals is constructed, where H and W denote the height and width of an EEG 2D map.

We define $\mathbf{X}^S = (\mathbf{S}_1^S, \mathbf{S}_2^S, \ldots, \mathbf{S}_B^S) \in \mathbb{R}^{E \times B}$ as the spectral features containing B frequency bands extracted from EEG signals, and $\mathbf{S}_b^S = (s_b^1, s_b^2, \ldots, s_b^E) \in \mathbb{R}^B (b \in \{1, 2, \ldots, B\})$ represents the EEG signals of all E electrodes collected in frequency band b. Similar to spatial-temporal representation, we also transform the spectral features to 3D spatial-spectral representation $\mathbf{X}^{S'} = (\mathbf{M}_1^S, \mathbf{M}_2^S, \ldots, \mathbf{M}_B^S) \in \mathbb{R}^{H \times W \times B}$.

The emotion recognition problem based on two EEG 3D representations can be defined as:

$$Y_{classification} = F(\mathbf{X}^{T'}, \mathbf{X}^{S'}) \tag{1}$$

where F represents the mapping function, and $Y_{classification}$ is the prediction of emotion. In addition, more detailed notations in this paper are defined in Table I.

TABLE I
Notation to be Used in the Proposed Method

Notation	Meaning
$\mathbb{R}$	the sets of real numbers.
$E; e$	E denotes the number of electrodes, and e denotes the index of electrodes.
$H, W; h, w$	H and W is the height and width of the 2D maps, and h, w is the pixel index in 2D maps.
$C; c$	C denotes the number of channels, and c denotes the index of channel.
$P; p$	P is the length of EEG signals, and p is the index of EEG signal series.
$B; b$	B is the number of frequency bands, and b denotes the index of frequency bands.
$\mathbf{X}^T$	EEG signals from one sample.
$\mathbf{X}^S$	spectral features of EEG signals from one sample.
$\mathbf{S}_p$	EEG signals from E electrodes at time stamp p.
$\mathbf{S}_b$	EEG signals from E electrodes at frequency band b.
$\mathbf{X}^{T'}$	3D spatial-temporal representation of EEG signals.
$\mathbf{X}^{S'}$	3D spatial-spectral representation of EEG signals.
$\mathbf{M}^{avg}$	the feature map calculated by channel-wise global average pooling.
$\mathbf{M}^{spa}$	the feature map for describing the distribution of EEG signals on 2D map.
$\mathbf{M}^{spe}$	the feature map for describing the distribution of EEG signals on different time stamps or frequency bands.
$\mathbf{A}^{spa}$	the normalized 2D spatial attention matrix.
$\mathbf{A}^{spe}$	the normalized 2D spectral attention matrix.
θ_m	the meta weight of the neural network.
θ'_t	the updated meta weights during meta training.
θ''_m	the meta weight after fast adaptation.

IV. Methodology

Fig. 3 presents the main architecture of MetaEmotionNet, which serves as the base spatial-spectral-temporal learner in meta-learning. The MetaEmotionNet contains two independent streams (spatial-spectral stream and spatial-temporal stream)

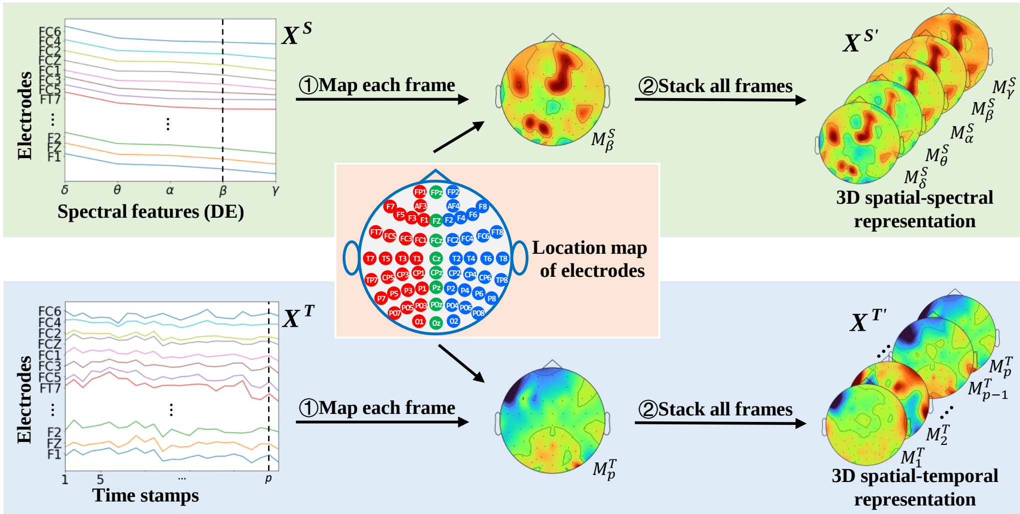


Fig. 4. The construction process of 3D map representation. Map each frame: map each spectral feature or time stamp based on the electrode position to build 2D EEG map. Stack all frame: stack 2D EEG maps by spectral or temporal dimension.

with the same network structure. Specifically, each stream consists of several Attention based 3D Dense Blocks (A3DB) and Transition Layers. At the end of MetaEmotionNet, the spatial-spectral-temporal features are fused from two streams. In addition, the meta-learning method is used to update the weight of the spatial-spectral-temporal learner. The spatial-spectral-temporal learner and the meta-learning method together form the MetaEmotionNet.

We summarize the key points of MetaEmotionNet as four points:

- Integrate the spatial-spectral-temporal features in a unified network architecture based on the 3D EEG representation.
- Develop a Spatial-Spectral/Temporal Attention module to adaptively capture some valuable brain regions and frequency bands or time stamps.
- Design the A3DB and Transition Layer to help feature reuse, enhance feature propagation, and have better parameter efficiency. Meanwhile, the Pseudo-3D is adopted to improve the memory efficiency of traditional 3D convolutions.
- Utilize meta-learning method to reduce the dependence on training data, so as to achieve fast adaptation to unknown subjects.

A. 3D Representation

To comprehensively model the spatial-spectral-temporal pattern of EEG signals, we design two representations: 3D spatial-temporal representation and 3D spatial-spectral representation as the input of MetaEmotionNet. Fig. 4 shows the generation of these two 3D representations.

Firstly, we construct 3D spatial-temporal representation from raw EEG signals. At a certain time p, the E channels EEG signals $\mathbf{S}_p^T = (s_p^1, s_p^2, \ldots, s_p^E) \in \mathbb{R}^E$ can be transformed into a 2D map $\mathbf{M}_p^T \in \mathbb{R}^{H\times W}$ according to the relative position of the electrodes on the head, where $p \in \{1, 2, \ldots, P\}$. Therefore, for a multi-channel EEG segment with P time stamps, P 2D maps can be obtained.

Secondly, we construct 3D spatial-spectral representation from the DE features of EEG signals. From raw EEG signals, the DE features of 5 frequency bands (δ, θ, α, β, γ) of each EEG channel can be calculated. Similar to the 3D spatial-temporal representation, for a certain frequency band b, the DE feature $\mathbf{S}_b^S = \left(\mathrm{s}_b^1, \mathrm{s}_b^2, \ldots, \mathrm{s}_b^E\right) \in \mathbb{R}^E$ can also be mapped to a 2D map $\mathbf{M}_b^S \in \mathbb{R}^{H\times W}$, where $b \in \{\delta, \theta, \alpha, \beta, \gamma\}$. Therefore, the multi-channel DE feature matrix of 5 frequency bands can be transformed into five 2D maps.

Finally, by stacking these 2D temporal maps and 2D spectral maps respectively, the 3D spatial-temporal representation and 3D spatial-spectral representation of EEG signals are obtained. In addition, larger size input helps to expand (deepen) the network, which improves the representation learning ability of the model. However, the spatial resolution of the collected signal depends on the number of electrodes is limited. To singularize the details of the feature maps, we need to improve their resolution. Thus, transposed convolution is employed on these maps to form each sample.

B. Spatial-Spectral-Temporal based Attention 3D Dense Network

The Spatial-Spectral-Temporal based Attention 3D Dense Network (SST-Net) is composed of multiple modules: Spatial-Spectral/Temporal Attention, 3D Densely Connected Module, Transition Layer, and Fusion Layer.

1) Spatial-Spectral/Temporal Attention: Different emotions have different discrimination patterns in brain regions, frequency bands, and time stamps [60]. To adaptively focus on the most valuable spatial-spectral-temporal information, we propose a parallel Spatial-Spectral/Temporal Attention (SST-Attention) module. In our model, the structures of Spatial-Spectral Attention and Spatial-Temporal Attention employed in the two streams are the same. We take Spatial-Spectral Attention as an example to introduce its structure as presented

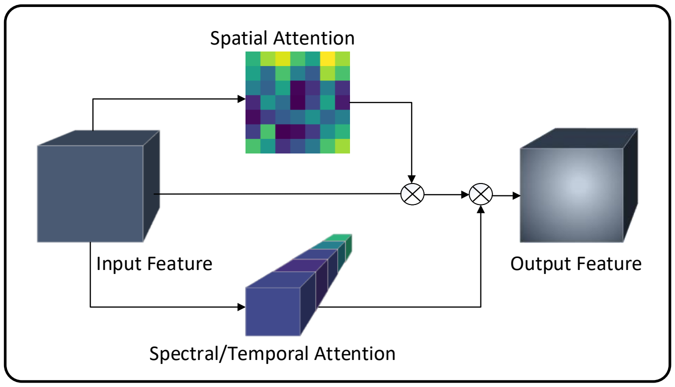


Fig. 5. The structure of spatial-spectral/temporal attention.

in Fig. 5. Specifically, for the input $\mathbf{X} \in \mathbb{R}^{H\times W\times B\times C}$ of attention layer, we utilize channel-wise global average pooling (cGAP) to shrink the dimension of feature map:

$$\mathbf{M}^{avg}_{h,w,b} = \mathbf{F_{cGAP}}(\mathbf{X}_{h,w,b}) = \frac{1}{C}\sum_{c=1}^{C}\mathbf{X}_{h,w,b}(c) \tag{2}$$

where $\mathbf{M}^{avg} \in \mathbb{R}^{H\times W\times B}$ denotes the compressed feature map along the channel dimension. $\mathbf{F_{cGAP}}$ is the cGAP function. $\mathbf{X}_{h,w,b}$ represents a vector containing cross-channel data at h, w, b of $\mathbf{X}$.

We calculate the attention value from different dimensions. First, the band-wise global average pooling (bGAP) and spatial-wise global average pooling (sGAP) are used to gather the features into the different dimensions in Spatial-Spectral Attention.

$$\mathbf{M}^{spa}_{h,w} = \mathbf{F_{bGAP}}(\mathbf{M}^{avg}_{h,w}) = \frac{1}{B}\sum_{b=1}^{B}\mathbf{M}^{avg}_{h,w}(b) \tag{3}$$

$$\mathbf{M}^{spe}_{b} = \mathbf{F_{sGAP}}(\mathbf{M}^{avg}_{b}) = \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}\mathbf{M}^{avg}_{b}(h,w) \tag{4}$$

where $\mathbf{M}^{spa} \in \mathbb{R}^{H\times W}$ represents the distribution of spectral feature on the brain generated by shrinking $\mathbf{M}^{avg}$ through band dimension B. $\mathbf{M}^{spe}_{b}$ represents the element in the b-th band of $\mathbf{M}^{spe}$. $\mathbf{F_{bGAP}}$ and $\mathbf{F_{sGAP}}$ represent the bGAP function and the sGAP function. $\mathbf{M}^{avg}_{h,w}$ represents a vector containing cross-band data at h, w of $\mathbf{M}^{avg}$. $\mathbf{M}^{avg}_{b}$ represents a matrix containing a 2D map in the b-th band of $\mathbf{M}^{avg}$.

Second, the fully-connected layer with softmax activation function is used to obtain the spatial attention matrix $\mathbf{A}^{spa}$ and spectral attention matrix $\mathbf{A}^{spe}$:

$$\mathbf{A}^{spa} = softmax(\mathbf{W}^{spa}\mathbf{M}^{spa} + \mathbf{b}^{spa}) \tag{5}$$

$$\mathbf{A}^{spe} = softmax(\mathbf{W}^{spe}\mathbf{M}^{spe} + \mathbf{b}^{spe}) \tag{6}$$

where $\mathbf{W}^{spa}$, $\mathbf{b}^{spa}$, $\mathbf{W}^{spe}$ and $\mathbf{b}^{spe}$ are learnable parameters.

Finally, we apply spatial attention and spectral attention on the feature maps:

$$\mathbf{X}^{spa-spe} = \mathbf{A}^{spe} \otimes (\mathbf{A}^{spa} \otimes \mathbf{X}) \tag{7}$$

where $\otimes$ is element-wise multiplication operating. In this operation, the spectral attention value is broadcasted along the spatial dimension, and the spatial attention value is broadcasted along the spectral dimension.

2) 3D Densely Connected Module: To make the deep spatial-spectral-temporal features better propagate, inspired by 2D DenseNet [61], we propose a 3D Densely Connected Module (3DCM). Each 3DCM module is composed of several Pseudo-3D convolutional layers with dense connectivity.

Dense Connectivity. The classic operation of DenseNet is the dense connection between convolutional layers, which means a direct connection from any layer to all subsequent layers. This connection approach is more conducive to the propagation of features and gradients [61]. As shown in Fig. 3, taking the l-th convolutional layer as an example, it takes the feature maps of the previous layers as input:

$$\mathbf{X}_l = H_l([\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_{l-1}]) \tag{8}$$

where $[\cdot]$ is a concatenate operation, and H_l represents the Pseudo-3D convolution.

Pseudo-3D. As is known to all, traditional 3D convolution takes up a lot of computing and storage resources. To solve this problem, we replace the traditional 3D convolution with Pseudo-3D convolution [62], which is defined as:

$$H_l(\mathbf{X}) = f^{1\times1\times d}(f^{k\times k\times 1}(\mathbf{X})) \tag{9}$$

3) Transition Layer: In DenseNet [61], the transition layer is used to reduce the number of feature maps and make the model more concise. To prevent over-fitting and reduce model complexity, our proposed model also utilizes this structure. Specifically, first we apply the batch normalization layer to normalize the feature maps. Then 1×1 convolution is employed to reduce the number of feature maps. For example, if m feature maps are input, $\lfloor\theta m\rfloor$ feature maps will be output after 1×1 convolution, where $\theta \in (0,1]$ represents the compression factor. Finally, a $2\times2\times1$ average pooling layer is used to further reduce The size of each feature map to avoid information redundancy and excessive calculation.

$$\mathbf{X}' = f^{1\times1\times1}(\mathbf{X}) \tag{10}$$

$$\mathbf{X}'' = \mathbf{F}^{2\times2\times1}_{\mathbf{AP}}(\mathbf{X}') \tag{11}$$

where $f^{1\times1\times1}$ represents the 1×1 convolution, and $\mathbf{F}^{2\times2\times1}_{\mathbf{AP}}$ represents the $2\times2\times1$ average pooling function.

4) Fusion Layer: To fuse the spatial-spectral-temporal features from two streams, we design a fusion layer for emotion recognition. The fusion layer first concatenates the outputs of the two streams, and then realizes multi-emotion classification through a fully connected layer with softmax activation. Naturally, we employ the cross-entropy loss commonly used in multi-classification tasks:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R} y_{i,r}\log\hat{y}_{i,r} \tag{12}$$

where R is the number of emotion classes, N is the number of samples, y represents the true label and $\hat{y}$ represents the value predicted by the model.

As far as we know, none of the existing methods consider the complementarity of the EEG multi-dimensional features at the same time. The proposed method simultaneously captures the spatial-spectral-temporal features of EEG through dual-stream fusion.

C. Meta-learning

Because of the subject differences among different subjects, the EEG signals collected from different subjects are not completely consistent. In order to overcome the subject differences, traditional transfer learning methods still need a lot of data from source domain and target domain to ensure the effectiveness of transfer learning, which is not suitable for the actual situation of users. Meta-learning method, a special transfer learning method [55], can learn task-related knowledge in the source domain through a small number of samples. Specifically, for training subjects, only part of their data needs to be sampled to participate in the training. For unknown subjects, only a small number of samples can quickly adapt to them. This is more in line with the practical application of brain-computer interfaces. Therefore, we propose a meta-learning method based on reptile [63] for emotion recognition to reduce the dependence on a large number of data.

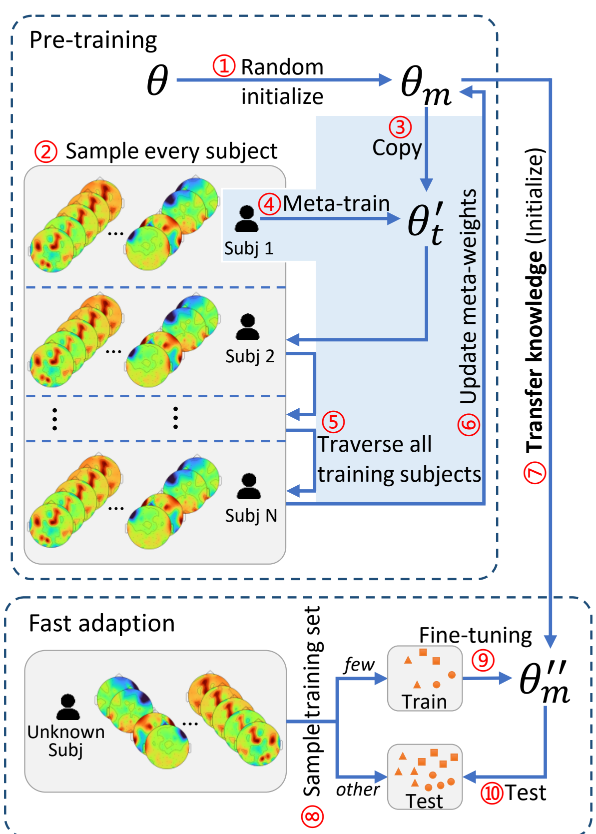


Fig. 6. The detailed process of our meta-learning framework, which includes pre-training stage and fast adaptation stage.

As shown in the Fig. 6, our meta-learning method is divided into two stages: pre-training (steps ①-⑥) and fast adaptation (steps ⑧-⑩).

Specifically, step ①, initialize the weight θ of SST-Net to θ_m randomly. Step ②, sample the data of each training subject. Each subject is treated as one meta-task. For example, the EEG data of a certain subject is used as the data distribution of a meta-task, and the meta-task takes minimizing cross entropy as its optimization goal. Of course, in actual experiments, we only need to randomly select a small number of samples for meta-training. Steps ③-⑥ are the meta-training process. First, take out a weight θ'_t, and then use the data of each subject in turn to optimize the weight θ'_t using the gradient descent method:

Algorithm 1 Algorithm for meta training

Require: Meta-tasks τ_{train}

Initialize the weight θ of SST-Net to θ_m
for meta training iteration **do**
  Copy: $\theta'_t \leftarrow \theta_m$
  for meta-task τ_b in τ_{train} **do**
    Randomly select few samples $\mathbf{X}_t = \{\mathbf{X}_t^T, \mathbf{X}_t^S, y_t\}$ from τ_b
    Forward: $\hat{y}_t = \text{SST-Net}(\theta'_t; \mathbf{X}_t^T, \mathbf{X}_t^S)$
    Calculate loss $\mathcal{L}_t = \mathcal{L}(\hat{y}_t, y_t)$ by Eq.(12)
    Backward: update θ'_t with several steps of SGD or Adam optimizer: $\theta'_t \leftarrow \theta'_t - \alpha \nabla_{\theta'_t} \mathcal{L}_t$
  end for
  Update meta weight θ_m: $\theta_m \leftarrow \theta_m + \beta\left(\theta'_t - \theta_m\right)$
end for

$$\theta'_t \leftarrow \theta'_t - \alpha \nabla_{\theta'_t} \mathcal{L}\left(F_{\theta'_t}\right) \tag{13}$$

where α denotes the learning rate of internal circulation. When the training subjects complete a traversal, as in step ⑥, the meta weight is updated as follows:

$$\theta_m \leftarrow \theta_m + \beta\left(\theta'_t - \theta_m\right) \tag{14}$$

where θ_m denotes the meta weight. After repeated meta training, the meta weight θ_m is updated multiple times. In this process, the model can gradually learn the relevant knowledge of the training subjects. Hence, as in step ⑦, transfer the meta-weight θ_m as learned task-related knowledge:

$$\theta''_m \leftarrow \theta_m \tag{15}$$

After that, we make fast adaptation to the unknown subjects. As in step ⑧, divide the unknown subjects' data into training set and test set based on cross-trial principle, and the number of samples in the training set is less than the test set (only 10% of the samples are used as training set). Then, as in step ⑨, utilize the sampled training set to fine-tune the weight θ''_m. Finally, as in step ⑩, use the test set to evaluate the performance of the model.

V. Experiments

In order to evaluate the performance of all models, we conducted a lot of experiments on two real emotion recognition datasets.

A. Datasets

We evaluate our model on SEED [14] and SEED-IV [36] datasets, which are two representative benchmark datasets. The details of the datasets are shown in Table II.

SEED dataset contains the EEG data of different emotions of 15 subjects (7 males and 8 females; average age: 23.27 ± 2.37). The different emotions (negative, positive, and neutral emotions) of the subjects were stimulated by watching 15 video clips. The duration of each video clip is about 240 seconds. 7 males and 8 females participated in the viewing of the video clip. Each subject has 3 sessions, each session contains 15 movie clips. SEED contains 15 subjects contained 62 EEG channels at a sampling frequency of 1000 Hz. These EEG signals are divided into 200 milliseconds as a sample, and then down-sampled to 200 Hz. The Short Time Fourier Transform is used to extract the DE features of different frequency bands for each channel.

SEED-IV dataset contains the EEG data of different emotions (happy, sad, fear, and neutral) of 15 subjects (7 males and 8 females; age range: 20 to 24 years old). These different emotions of the subjects were stimulated by watching 72 video clips. The duration of each video clip is about 120 seconds. 7 males and 8 females participated in the viewing of the video clip. Each subject needs to have three sessions. SEED-IV contains 15 subjects contained 62 EEG channels at a sampling frequency of 1000 Hz. These EEG signals are divided into 4-second segments as a sample, and then down-sampled to 128 Hz. Similar to SEED, the Short Time Fourier Transform is used to extract the DE features of five frequency bands for each channel.

TABLE II
THE SUMMARY TABLE OF SEED AND SEED-IV DATASET

Character	Details	
Dataset	SEED	SEED-IV
EEG Electrodes	62	62
Frequency band	δ , θ , α , β , γ	δ , θ , α , β , γ
Emotions	positive, neutral, negative	positive, neutral, negative, fear
Subjects	15	15
Sessions	3	3
Trials	15	24

B. Settings

We implement the MetaEmotionNet based on the Tensorflow framework. We utilize the cross entropy as the loss function while training the MetaEmotionNet on NVIDIA RTX 2080 GPU. We choose Adam optimizer to minimize the loss function. The inner learning rate is set to 0.001, and the meta-learning rate is set to 0.01. The number of the A3DB is set to 2 and the number of layers in each 3DCM is set to 6. For the Transition Layer, we set its compression factor $\theta = 0.5$. The entire model architecture is shown in Table V.

We adopt the leave-one-subject-out (LOSO) cross-validation strategy to partition the dataset. To guarantee a fair comparison, we use the same experimental settings for all baselines and the proposed method. Specifically, we utilize all the data from 14 subjects as the training set for each fold. The remaining one subject, regarded as an unknown subject, uses a small number of trials (3 trials per session for SEED, and 4 trials per session for SEED-IV) for fine-tuning, and the remaining trials for testing. In addition, in order to evaluate the performance of our model when the data is reduced, the MetaEotionNet trained with *few* samples is called MetaEotionNet(f). In this case, for training subjects, we only randomly sample 10% from each subject's data for training; for unknown subject, we also use about 10% of the data from several trials for fast adaption. All experiments follow the principle of trial independence. Table III shows the number of samples used for each model. Table IV shows the number of samples in each emotion category for each subject.

To evaluate the classification performance of the model and compare it with the other baseline models, we employ Accuracy (ACC), F1-score, and Kappa as the evaluation metrics, which are defined as follows:

$$\text{ACC} = \frac{TP+TN}{TP+TN+FP+FN} \in [0,1] \tag{16}$$

$$\text{F1-score} = \frac{2TP}{2TP+FP+FN} \in [0,1] \tag{17}$$

$$\text{Kappa} = \frac{p_0 - p_e}{1 - p_e} \in [-1,1] \tag{18}$$

where TP denotes true positive, TN denotes true negative, FP denotes false positive, FN denotes false negative, p_o denotes the observed agreement, and p_e denotes the expected agreement. Larger values indicate better classification performance.

TABLE III
THE NUMBER OF SAMPLES USED FOR EACH MODEL. SEED HAS 10182 SAMPLES PER SUBJECT, AND SEED-IV HAS 2505 SAMPLES PER SUBJECT.

Dataset	SEED	SEED-IV
Samples per training subject	10182 (100%), 3 sessions × 15 trials	2505 (100%), 3 sessions × 24 trials
Samples per training subject for MetaEmotionNet(f)	1018 (10%), sample randomly	250 (10%), sample randomly
Calibration samples of target subject	2022 (20%), 3 sessions × 3 trials	446 (18%), 3 sessions × 4 trials
Test samples of target subject	8160 (80%), 3 sessions × 12 trials	2059 (82%), 3 sessions × 20 trials

TABLE IV
THE NUMBER OF SAMPLES PER SUBJECT FOR EACH EMOTION CATEGORY IN THE SEED AND SEED-IV DATASETS.

SEED	Emotion	Negative	Neutral		Positive	All
	Samples	3510	3312		3360	10182
	Ratio	34.5%	32.5%		33.0%	100%
SEED-IV	Emotion	Neutral	Sad	Fear	Happy	All
	Samples	678	683	615	529	2505
	Ratio	27.1%	27.3%	24.5%	21.1%	100%

C. Baseline Models

- SVM [64]: Support vector machine utilizes the least squares to perform classification.
- RF [65]: Random forest, a classifier composed of multiple decision trees.

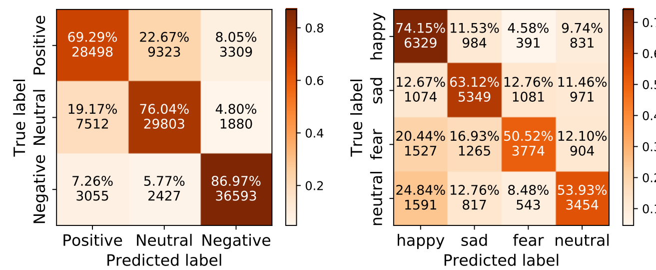
TABLE V
MODEL ARCHITECTURE.

Layers		
Transposed Convolution	$8 \times 8 \times 1$ conv	$\times 1$
Convolution	$3 \times 3 \times 1$ conv	$\times 1$
A3DB(1)	$3 \times 3 \times 1$ conv	$\times 6$
	$1 \times 1 \times 3$ conv	
Transition Layer	1×1 conv	$\times 1$
	2×2 average pooling	
A3DB(2)	$3 \times 3 \times 1$ conv	$\times 6$
	$1 \times 1 \times 3$ conv	
Classification Layer	Concatenate Layer	$\times 1$
	3 Dense, softmax	

- MLP [66]: Multilayer perceptron, a basic feedforward neural network.
- STRNN [15]: Spatial–temporal recurrent neural network. It integrates the feature learning from both spatial and temporal information of signal into a unified model.
- 3D-CNN [16]: A 3D convolutional neural network designed to capture discriminative deep features of EEG signals.
- MMResLSTM [17]: Multimodal Residual LSTM Network, employs the LSTM neural network combined with the residual structure to efficiently learning emotion-related high-level features.
- CDCN [18]: Channel-fused Dense Convolutional Network utilizes convolution and dense structures handling temporal dependencies and electrode correlations of EEG signals.
- ACRNN [19]: Attention-based Convolutional Recurrent Neural Network comprehensively considers the spatial information, temporal information and attention information of the EEG signals.
- STFFNN [67]: Spatial-Temporal Feature Fusion Neural Network, which extracts discriminative spatial-temporal features and integrates complementary information.
- TSception [68]: A multi-scale convolutional neural network for EEG emotion recognition, which captures temporal dynamics and spatial asymmetry of EEG.

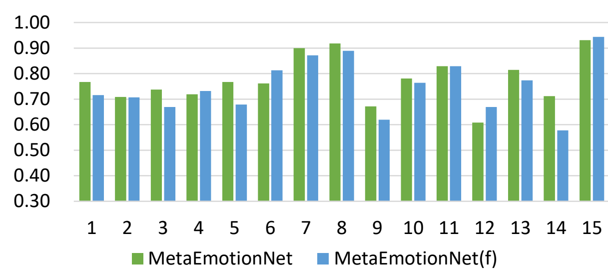
D. Overall Comparison

We compare MetaEmotionNet with the state-of-the-art baselines on emotion recognition datasets. As shown in Table VI, the traditional machine learning methods SVM and RF only have the worst performance because it is difficult to capture the abundant information of EEG signals. CNN and RNN can better extract deep temporal or spatial features. Hence, the final performances of STRNN, 3D-CNN, CDCN, and MMResLSTM are better than traditional machine learning methods. The recently proposed ACRNN, STFFNN, and TSception all model the temporal and spatial dimensions of EEG signals. Meanwhile, ACRNN utilizes attentional mechanisms to focus on valuable temporal and spatial information, STFFNN integrates complementary information between discriminative features, and TSception captures the temporal dynamics and spatial asymmetry of EEG. These methods improve classification stability and achieve better results on the SEED or SEED-IV datasets. The proposed method MetaEmotionNet uses two streams of spatial-temporal and spatial-spectral as well as the attention mechanism to comprehensively extract the spatial-spectral-temporal features of EEG signals. Thus, the proposed method achieves the best performance among metrics. At the same time, the meta-learning method effectively improves the adaptability of the model, which is more effective than traditional methods. It is worth noting that our meta-learning method achieves competitive performance even with only a small amount of data. Especially on the SEED-IV dataset, our method is still superior to baseline methods which use more data. For more detailed classification results, the confusion matrices of proposed MetaEmotionNet are shown in Fig. 7. The classification accuracy of each subject is shown in Fig. 8.

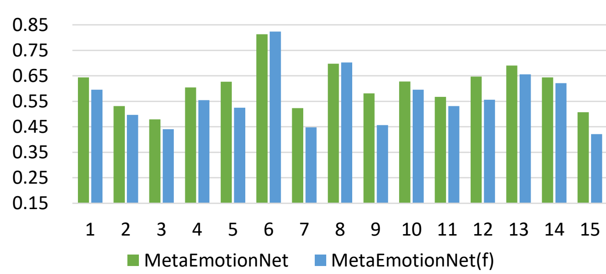


(a) SEED dataset (b) SEED-IV dataset

Fig. 7. The confusion matrices of MetaEmotionNet on SEED and SEED-IV datasets.



(a) SEED dataset



(b) SEED-IV dataset

Fig. 8. The classification ACC (accuracy) of each subject using MetaEmotionNet. The maximum expected value of the ACC is 1 (i.e., 100%).

E. Ablation Studies and Analysis

We conduct ablation experiments to further evaluate the effectiveness of different components in MetaEmotionNet and analyze the results of the experiment on the SEED dataset.

TABLE VI
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART MODELS ON THE SEED AND SEED-IV DATASETS. METAEMOTIONNET AND METAEMOTIONNET(f) ARE OUR PROPOSED MODEL, IN WHICH METAEMOTIONNET(f) ONLY USES A SMALL NUMBER OF SAMPLES.

Model	#ATD	SEED			SEED-IV		
		ACC / STD	F1-score / STD	Kappa / STD	ACC / STD	F1-score / STD	Kappa / STD
RF [65]	All	0.533 / 0.087	0.487 / 0.115	0.299 / 0.129	0.347 / 0.066	0.297 / 0.083	0.134 / 0.083
SVM [64]	All	0.531 / 0.109	0.468 / 0.137	0.296 / 0.163	0.411 / 0.074	0.348 / 0.075	0.209 / 0.094
MLP [66]	All	0.701 / 0.094	0.674 / 0.123	0.550 / 0.142	0.507 / 0.063	0.397 / 0.083	0.328 / 0.089
STRNN [15]	All	0.732 / 0.111	0.723 / 0.116	0.598 / 0.166	0.532 / 0.074	0.517 / 0.076	0.372 / 0.100
3D-CNN [16]	All	0.742 / 0.078	0.738 / 0.079	0.613 / 0.116	0.541 / 0.109	0.507 / 0.133	0.384 / 0.149
MMResLSTM [17]	All	0.744 / 0.104	0.713 / 0.106	0.616 / 0.157	0.511 / 0.114	0.461 / 0.112	0.348 / 0.149
CDCN [18]	All	0.693 / 0.094	0.668 / 0.116	0.540 / 0.142	0.545 / 0.131	0.521 / 0.150	0.393 / 0.172
ACRNN [19]	All	0.763 / 0.081	0.739 / 0.093	0.644 / 0.121	0.492 / 0.092	0.462 / 0.082	0.296 / 0.110
STFFNN [67]	All	0.720 / 0.090	0.710 / 0.095	0.579 / 0.136	0.567 / 0.073	0.550 / 0.075	0.418 / 0.099
TSception [68]	All	0.643 / 0.098	0.636 / 0.107	0.465 / 0.148	0.562 / 0.095	0.558 / 0.096	0.414 / 0.128
MetaEmotionNet	All	**0.775 / 0.088**	**0.772 / 0.087**	**0.663 / 0.132**	**0.612 / 0.083**	**0.589 / 0.102**	**0.479 / 0.114**
MetaEmotionNet(f)	Few	0.750 / 0.100	0.737 / 0.118	0.625 / 0.151	0.562 / 0.106	0.535 / 0.121	0.412 / 0.143

* #ATD is the amount of training data. "Few" means that only 10% of the training data is used for training.
* ACC, F1-score $\in [0, 1]$, and Kappa $\in$ [-1,1]. Larger values indicate better classification performance.

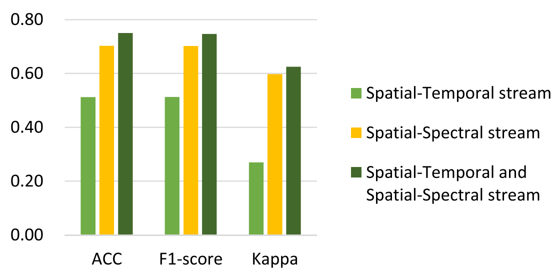


Fig. 9. Ablation studies on two stream fusion. The maximum expected value for all metrics is 1.

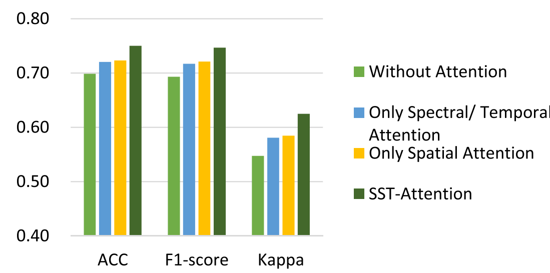


Fig. 10. Ablation studies on attention modules. The maximum expected value for all metrics is 1.

Ablation on two stream structure. We conduct ablation experiments on the fusion of spatial-temporal stream and spatial-spectral stream. As shown in Fig. 9, it is obvious that the performances of the single stream models, only spatial-spectral or only spatial-temporal stream, are not as good as the two stream structure. Specifically, the accuracy of spatial-spectral stream and spatial-temporal stream is 4.75% and 23.81% lower than the accuracy of two stream structure respectively. The results show that two stream structure effectively leverages the spatial-spectral-temporal features, and the features from different dimensions are complementary to improve the classification accuracy. Moreover, spectral features may be more suitable for emotion recognition than temporal features. This is manifested in that the model only with spatial-spectral stream has a better performance than the one only with spatial-temporal stream.

Ablation on Spatial-Spectral/Temporal Attention. In order to explore the effectiveness of Spatial-Spectral/Temporal Attention mechanism, we compared the performance of four variants: without attention, only Spectral/Temporal Attention, only Spatial Attention, and Spatial-Spectral/Temporal Attention. Fig. 10 presents that Spatial-Spectral/Temporal Attention contributes to the optimal performance of the model, which is better than other variants in all indicators including accuracy, F1-score, and Kappa. When part of SST-Attention is ablated, the evaluation metrics decrease. For example, the classification accuracy of SST-Attention increases by 2.70% and 2.98% than the model only adopted Spatial Attention and the model only adopted Spectral/Temporal Attention respectively. After further ablation, when the whole SST-Attention module is ablated, the obtained classification accuracy is the worst among the variants used for comparison.

In conclusion, the results present that both Spatial Attention and Spectral/Temporal Attention can effectively capture valuable information and give higher attention on attentive brain regions, frequency bands, and time stamps. The Spatial-Spectral/Temporal Attention combines the two, it considers more comprehensive attention, which is more conducive to the classification.

TABLE VII
ABLATION STUDIES ON META-LEARNING.

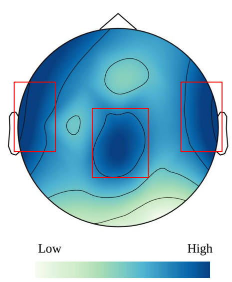
Model	ACC / STD	F1-score / STD	Kappa / STD
without Meta-learning	0.690 / 0.077	0.687 / 0.086	0.534 / 0.116
MetaEmotionNet	0.775 / 0.088	0.772 / 0.087	0.663 / 0.132
MetaEmotionNet(f)	0.750 / 0.100	0.737 / 0.118	0.623 / 0.151

Ablation on Meta-learning. In order to explore the effectiveness of meta-learning, we compared the performance differences with and without meta-learning. For the experiment without meta-learning, we use the same data partitioning strategy as MetaEmotionNet. As shown in Table VII, the model without meta-learning only achieves an accuracy of 68.96%. When the meta-learning method is added, the performance of the model has been greatly improved. Even if the amount of data used is smaller, its accuracy rate is still 6.07% higher than that without meta-learning.

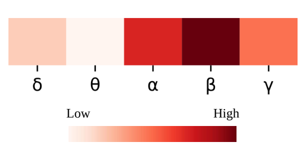
VI. Discussions

In this section, we explore the interpretability of the proposed MetaEmotionNet and the advantages of the meta-learning strategy, and discuss the limitations.

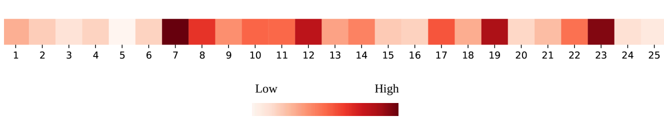
First, we visualize the Spatial-Spectral/Temporal Attention in MetaEmotionNet. Fig. 11 (a) indicates that Spatial Attention shows more interest in specific brain regions (such as temporal lobe and parietal lobe are marked by the red boxes). This is consistent with some existing physiological conclusions. For example, the temporal lobe is the key to distinguishing between positive emotion and negative emotion. The proposed Spatial Attention can adaptively recognize discriminative local patterns for emotion states in specific brain regions. Fig. 11 (b) indicates that specific frequency bands play a more important role than other bands in the recognition of emotion state. It is in compliance with existing researches on the single band that the higher frequency bands (α, β, and γ) make more contributions to the classification [14], [36]. Fig. 11 (c) indicates that Temporal Attention module pays more attention on $7^{th}, 19^{th}, 23^{rd}$ of 25 timestamps in a sample.



(a) Spatial Attention weights on the brain.



(b) Spectral Attention weights across different frequency bands.



(c) Temporal Attention weights across the time stamps.

Fig. 11. The visualization of Spatial-Spectral/Temporal attention weights. The darker the color, the higher the attention weight.

Then, for the meta-learning strategy, in order to explore its effect on the model fitting ability, we compared the convergence speed of the model with and without the meta-learning strategy. As shown in Fig. 12, MetaEmotionNet converges faster than the model without meta-learning. MetaEmotionNet only needs about 3 epochs to achieve better performance, while the model without meta-learning requires more epochs.

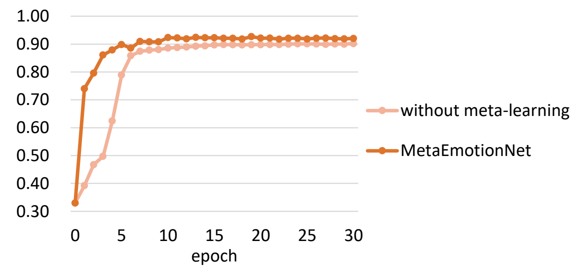


Fig. 12. Accuracy curves of different variant models when adapting to unknown subject.

Limitations. On the one hand, each person's emotion is affected by various factors such as age, gender, and cultural differences, etc. The datasets used in this study are all collected from a group of Chinese youths, and do not take into account the issue of age and cultural differences among different subjects, which may produce bias in the practical application of emotion modeling. Although the meta-learning strategy in this paper can quickly adapt to new subjects, relevant experiments are needed to verify the effect of emotion recognition in cross-age and cross-cultural scenarios. On the other hand, in real life, emotions when watching videos are not static, but change continuously over time, and laughter and tears in video clips can induce sudden changes in emotions, but most of the current emotion datasets do not take this temporal dynamics into account, and many of the datasets have video clips with specific emotional tendencies, which poses a challenge for developing a more intelligent, accurate, and close-to-reality emotion recognition system.

VII. Conclusion

In this paper, we propose a novel MetaEmotionNet model for cross-subject emotion recognition. MetaEmotionNet effectively uses the complementarity of different features through the two-stream 3D neural network. At the same time, the attention mechanism is developed to adaptively focus on the attentive spatial-spectral-temporal features of the brain. The meta-learning approach reduces the influence of subject differences through meta updating and fast adaptation. Even if the data dependence is reduced, it still achieves high classification performance. Experiments on two real-world datasets show MetaEmotionNet is better than the state-of-the-art baselines. It is noteworthy that the proposed SST-Attention module has a certain degree of physiological interpretability. In addition, our MetaEmotionNet provides a general deep learning framework for physiological time series, which can be further applied to other fields, such as motor imagery and mental workload assessment.

Acknowledgments

Dr. Paul Streli for the assistance in proofreading the initial manuscript.

REFERENCES

[1] A. M. Al-Kaysi, A. Al-Ani, C. K. Loo, T. Y. Powell, D. M. Martin, M. Breakspear, and T. W. Boonstra, "Predicting tdcs treatment outcomes of patients with major depressive disorder using automated eeg classification," *Journal of Affective Disorders*, vol. 208, pp. 597–603, 2017.

[2] A. V. Bocharov, G. G. Knyazev, and A. N. Savostyanov, "Depression and implicit emotion processing: An eeg study," *Neurophysiologie Clinique/Clinical Neurophysiology*, vol. 47, no. 3, pp. 225–230, 2017.

[3] P. Zhong, D. Wang, and C. Miao, "Eeg-based emotion recognition using regularized graph neural networks," *arXiv preprint arXiv:1907.07835*, 2019.

[4] A. Alnafjan, M. Hosny, A. Al-Wabil, and Y. Al-Ohali, "Classification of human emotions from electroencephalogram (eeg) signal using deep neural network," *International Journal of Advanced Computer Science and Applications*, vol. 8, 01 2017.

[5] C. Dai, J. Wu, D. Pi, L. Cui, B. Johnson, and S. I. Becker, "Electroencephalogram signal clustering with convex cooperative games," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.

[6] K. Anderson and P. W. McOwan, "A real-time automated system for the recognition of human facial expressions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 1, pp. 96–105, 2006.

[7] H. Gao, M. Wu, Z. Chen, Y. Li, X. Wang, S. An, J. Li, and C. Liu, "Ssa-icl: Multi-domain adaptive attention with intra-dataset continual learning for facial expression recognition," *Neural Networks*, vol. 158, pp. 228–238, 2023.

[8] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Proceedings of artificial neural networks in engineering*, vol. 710, 1999, p. 22.

[9] X. Xu, J. Deng, E. Coutinho, C. Wu, L. Zhao, and B. W. Schuller, "Connecting subspace learning and extreme learning machine in speech emotion recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 795–808, 2019.

[10] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE transactions on affective computing*, vol. 3, no. 2, pp. 211–223, 2011.

[11] G. Du, Y. Zeng, K. Su, C. Li, X. Wang, S. Teng, D. Li, and P. X. Liu, "A novel emotion-aware method based on the fusion of textual description of speech, body movements and facial expressions," *IEEE Transactions on Instrumentation and Measurement*, pp. 1–1, 2022.

[12] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang, "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Computer methods and programs in biomedicine*, vol. 140, pp. 93–110, 2017.

[13] P. Sarma and S. Barma, "Emotion recognition by discriminating eeg segments with high affective content from automatically selected relevant channels," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[14] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on Autonomous Mental Development*, vol. 7, no. 3, pp. 162–175, 2015.

[15] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial–temporal recurrent neural network for emotion recognition," *IEEE transactions on cybernetics*, vol. 49, no. 3, pp. 839–847, 2018.

[16] Y. Zhao, J. Yang, J. Lin, D. Yu, and X. Cao, "A 3d convolutional neural network for emotion recognition based on eeg signals," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–6.

[17] J. Ma, H. Tang, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal residual lstm network," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 176–183.

[18] Z. Gao, X. Wang, Y. Yang, Y. Li, K. Ma, and G. Chen, "A channel-fused dense convolutional network for eeg-based emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, 2020.

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

[20] S. Alhagry, A. A. Fahmy, and R. A. El-Khoribi, "Emotion recognition based on eeg using lstm recurrent neural network," *Emotion*, vol. 8, no. 10, pp. 355–358, 2017.

[21] J. Li, Z. Zhang, and H. He, "Hierarchical convolutional neural networks for eeg-based emotion recognition," *Cognitive Computation*, vol. 10, no. 2, pp. 368–380, 2018.

[22] S. Siddharth, T.-P. Jung, and T. J. Sejnowski, "Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing," *IEEE Transactions on Affective Computing*, 2019.

[23] T. Song, W. Zheng, P. Song, and Z. Cui, "Eeg emotion recognition using dynamical graph convolutional neural networks," *IEEE Transactions on Affective Computing*, 2018.

[24] T. Zhang, X. Wang, X. Xu, and C. P. Chen, "Gcb-net: Graph convolutional broad network and its application in emotion recognition," *IEEE Transactions on Affective Computing*, 2019.

[25] Z. Jia, Y. Lin, X. Cai, H. Chen, H. Gou, and J. Wang, "Sst-emotionnet: Spatial-spectral-temporal based attention 3d dense network for eeg emotion recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2909–2917.

[26] X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar, and P. Marttinen, "Eeg based emotion recognition: A tutorial and review," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–57, 2022.

[27] D. Wu, Y. Xu, and B.-L. Lu, "Transfer learning for eeg-based brain–computer interfaces: A review of progress made since 2016," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 1, pp. 4–19, 2020.

[28] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[29] S. Aggarwal and N. Chugh, "Review of machine learning techniques for eeg based brain computer interface," *Archives of Computational Methods in Engineering*, pp. 1–20, 2022.

[30] Z. Jia, Y. Lin, Y. Liu, Z. Jiao, and J. Wang, "Refined nonuniform embedding for coupling detection in multivariate time series," *Physical Review E*, vol. 101, no. 6, p. 062113, 2020.

[31] Z. Jia, Y. Lin, Z. Jiao, Y. Ma, and J. Wang, "Detecting causality in multivariate time series via non-uniform embedding," *Entropy*, vol. 21, no. 12, p. 1233, 2019.

[32] Z. Jia, Y. Lin, J. Wang, R. Zhou, X. Ning, Y. He, and Y. Zhao, "Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, C. Bessiere, Ed. ijcai.org, 2020, pp. 1324–1330. [Online]. Available: https://doi.org/10.24963/ijcai.2020/184

[33] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, "Salientsleepnet: Multimodal salient wave detection network for sleep staging," *arXiv preprint arXiv:2105.13864*, 2021.

[34] Z. Jia, Y. Lin, J. Wang, K. Yang, T. Liu, and X. Zhang, "Mmcnn: A multi-branch multi-scale convolutional neural network for motor imagery classification," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*. Springer International Publishing, 2021, pp. 736–751.

[35] Z. Jia, X. Cai, G. Zheng, J. Wang, and Y. Lin, "Sleepprintnet: A multivariate multimodal neural network based on physiological time-series for automatic sleep staging," *IEEE Transactions on Artificial Intelligence*, 2021.

[36] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE Transactions on Cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2019.

[37] X. Zhang, J. Liu, J. Shen, S. Li, K. Hou, B. Hu, J. Gao, and T. Zhang, "Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine," *IEEE transactions on cybernetics*, 2020.

[38] X.-W. Wang, D. Nie, and B.-L. Lu, "Eeg-based emotion recognition using frequency domain features and support vector machines," in *International conference on neural information processing*. Springer, 2011, pp. 734–743.

[39] F. Bahari and A. Janghorbani, "Eeg-based emotion recognition using recurrence plot analysis and k nearest neighbor classifier," in *2013 20th Iranian Conference on Biomedical Engineering (ICBME)*. IEEE, 2013, pp. 228–233.

[40] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[41] C. A. Frantzidis, C. Bratsas, C. L. Papadelis, E. Konstantinidis, C. Pappas, and P. D. Bamidis, "Toward emotion aware computing: an integrated approach using multichannel neurophysiological recordings and affective visual stimuli," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 589–597, 2010.

[42] L. I. Goldfischer, "Autocorrelation function and power spectral density of laser-produced speckle patterns," *JOSA*, vol. 55, no. 3, pp. 247–253, 1965.

[43] R.-N. Duan, J.-Y. Zhu, and B.-L. Lu, “Differential entropy feature for eeg-based emotion classification,” in *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2013, pp. 81–84.

[44] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu, “Eeg-based emotion classification using deep belief networks,” in *2014 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2014, pp. 1–6.

[45] Y. Yang, Q. J. Wu, W.-L. Zheng, and B.-L. Lu, “Eeg-based emotion recognition using hierarchical network with subnetwork nodes,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 2, pp. 408–419, 2017.

[46] R. Fourati, B. Ammar, C. Aouiti, J. Sanchez-Medina, and A. M. Alimi, “Optimized echo state network with intrinsic plasticity for eeg-based emotion recognition,” in *International Conference on Neural Information Processing*. Springer, 2017, pp. 718–727.

[47] Y. Li, W. Zheng, L. Wang, Y. Zong, L. Qi, Z. Cui, T. Zhang, and T. Song, “A novel bi-hemispheric discrepancy model for eeg emotion recognition,” *arXiv preprint arXiv:1906.01704*, 2019.

[48] X. Chai, Q. Wang, Y. Zhao, Y. Li, D. Liu, X. Liu, and O. Bai, “A fast, efficient domain adaptation technique for cross-domain electroencephalography (eeg)-based emotion recognition,” *Sensors*, vol. 17, no. 5, p. 1014, 2017.

[49] Y.-P. Lin and T.-P. Jung, “Improving eeg-based emotion classification using conditional transfer learning,” *Frontiers in human neuroscience*, vol. 11, p. 334, 2017.

[50] X. Chai, Q. Wang, Y. Zhao, X. Liu, O. Bai, and Y. Li, “Unsupervised domain adaptation techniques based on auto-encoder for non-stationary eeg-based emotion recognition,” *Computers in biology and medicine*, vol. 79, pp. 205–214, 2016.

[51] F. Fahimi, Z. Zhang, W. B. Goh, T.-S. Lee, K. K. Ang, and C. Guan, “Inter-subject transfer learning with an end-to-end deep convolutional neural network for eeg-based bci,” *Journal of neural engineering*, vol. 16, no. 2, p. 026007, 2019.

[52] 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 Transactions on Affective Computing*, 2019.

[53] L.-M. Zhao, X. Yan, and B.-L. Lu, “Plug-and-play domain adaptation for cross-subject eeg-based emotion recognition,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, 2021, pp. 863–870.

[54] W. Zhang, L. Deng, and D. Wu, “Overcoming negative transfer: A survey,” *arXiv preprint arXiv:2009.00909*, 2020.

[55] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[56] H. Shao, D. Zhong, X. Du, S. Du, and R. N. J. Veldhuis, “Few-shot learning for palmprint recognition via meta-siamese network,” *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[57] J. Long, R. Zhang, Z. Yang, Y. Huang, Y. Liu, and C. Li, “Self-adaptation graph attention network via meta-learning for machinery fault diagnosis with few labeled data,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[58] N. Banluesombatkul, P. Ouppaphan, P. Leelaarporn, P. Lakhan, B. Chaitusaney, N. Jaimchariyatam, E. Chuangsuwanich, W. Chen, H. Phan, N. Dilokthanakul, and T. Wilaiprasitporn, “Metasleeplearner: A pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 6, pp. 1949–1963, 2021.

[59] A. Antoniou, H. Edwards, and A. Storkey, “How to train your maml,” *arXiv preprint arXiv:1810.09502*, 2018.

[60] T.-H. Li, W. Liu, W.-L. Zheng, and B.-L. Lu, “Classification of five emotions from eeg and eye movement signals: Discrimination ability and stability over time,” in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2019, pp. 607–610.

[61] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[62] Z. Qiu, T. Yao, and T. Mei, “Learning spatio-temporal representation with pseudo-3d residual networks,” in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[63] A. Nichol, J. Achiam, and J. Schulman, “On first-order meta-learning algorithms,” *arXiv preprint arXiv:1803.02999*, 2018.

[64] J. A. Suykens and J. Vandewalle, “Least squares support vector machine classifiers,” *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[65] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[66] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[67] Z. Wang, Y. Wang, J. Zhang, C. Hu, Z. Yin, and Y. Song, “Spatial-temporal feature fusion neural network for eeg-based emotion recognition,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[68] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, and C. Guan, “Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition,” *IEEE Transactions on Affective Computing*, pp. 1–1, 2022.