A Multi-view Spectral-Spatial-Temporal Masked Autoencoder for Decoding Emotions with Self-supervised Learning

Rui Li
Shanghai Jiao Tong University
Shanghai, China
realee@sjtu.edu.cn

Yiting Wang
Shanghai Jiao Tong University
Shanghai, China
yw883@cornell.edu

Wei-Long Zheng
Shanghai Jiao Tong University
Shanghai, China
weilonglive@gmail.com

Bao-Liang Lu*
Shanghai Jiao Tong University
Shanghai, China
bllu@sjtu.edu.cn

ABSTRACT

Affective Brain-computer Interface has achieved considerable advances that researchers can successfully interpret labeled and flawless EEG data collected in laboratory settings. However, the annotation of EEG data is time-consuming and requires a vast workforce which limits the application in practical scenarios. Furthermore, daily collected EEG data may be partially damaged since EEG signals are sensitive to noise. In this paper, we propose a Multi-view Spectral-Spatial-Temporal Masked Autoencoder (MV-SSTMA) with self-supervised learning to tackle these challenges towards daily applications. The MV-SSTMA is based on a multi-view CNN-Transformer hybrid structure, interpreting the emotion-related knowledge of EEG signals from spectral, spatial, and temporal perspectives. Our model consists of three stages: 1) In the generalized pre-training stage, channels of unlabeled EEG data from all subjects are randomly masked and later reconstructed to learn the generic representations from EEG data; 2) In the personalized calibration stage, only few labeled data from a specific subject are used to calibrate the model; 3) In the personal test stage, our model can decode personal emotions from the sound EEG data as well as damaged ones with missing channels. Extensive experiments on two open emotional EEG datasets demonstrate that our proposed model achieves state-of-the-art performance on emotion recognition. In addition, under the abnormal circumstance of missing channels, the proposed model can still effectively recognize emotions.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; • **Computing methodologies** → **Artificial intelligence**; **Cognitive science**.

*Corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548243

KEYWORDS

affective computing; EEG-based emotion recognition; self-supervised learning; CNN-Transformer

ACM Reference Format:
Rui Li, Yiting Wang, Wei-Long Zheng, and Bao-Liang Lu. 2022. A Multi-view Spectral-Spatial-Temporal Masked Autoencoder for Decoding Emotions with Self-supervised Learning. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548243

1 INTRODUCTION

Affective Brain-computer Interfaces (aBCIs) allow machines to recognize and regulate human emotions. The aBCI technology has a vast potential not only for treating psychiatric disorders but also for serving as assessment tools for the general population in our daily lives [21]. Various modalities are utilized in aBCIs including functional magnetic resonance imaging (fMRI), stereo-electro-encephalography (SEEG), electroencephalography (EEG), etc., where EEG signal is non-intrusive and relatively easy to collect, especially with portable dry electrode EEG devices [24]. EEG can be used in a relatively convenient way while revealing delicate changes of emotion with high time resolution [8]. Among all underlying technologies, emotion recognition is the groundwork and milestone of aBCIs [4] making it a critical topic to investigate, and accurately evaluating people's emotion states can also contribute to the mental health assessment [3]. Thus, EEG-based emotion recognition has evolved rapidly in recent years with many excellent studies exploiting intact EEG data in a supervised manner [10, 14, 17, 31]. However, they usually demand annotated EEG data which requires a lot of workforce.

Imagine in the future brain-big-data center where EEG data from a massive group of people are sent back in real-time so that an abundance of unlabeled signals will be collected. Meanwhile, EEG data are usually sensitive to noise so they may be corrupted in real-world applications. Thus, many current outstanding supervised structures aiming for labeled and flawless data may perform inadequately. To get through those obstacles of practical applications, self-supervised learning is a data-efficient paradigm decoding representations with generalization ability, which draws growing attention in recent years [18].

There are several studies focused on contrastive learning to learn the representations of EEG data which can tackle the problems of

lacking labeled data [2, 11, 20]. However, if the data is also corrupted, it will be more applicable to utilize generative learning with auto-encoding structures to decode the representations of original data from the corrupted inputs. He *et al.* [9] proposed a masked autoencoder for self-supervised learning in the computer vision area achieving excellent performance. However, it cannot be directly employed well in EEG-based emotion recognition since the different properties between images and EEG signals.

To cope with these challenges, we propose a Multi-view Spectral-Spatial-Temporal Masked Autoencoder (MV-SSTMA) with self-supervised learning that can fully utilize the multi-view information of increasing unlabeled EEG data and also be able to recognize emotion states from damaged data. The MV-SSTMA model is based on a multi-view CNN-Transformer hybrid structure, consisting of the spectral embedding, multi-head spacial attention, and multi-scale casual convolution components for interpreting the spectral, spatial and temporal information of EEG signals. To be explicit, our model pre-trains a generalized model with unlabeled EEG data of all subjects by reconstructing masked data. The generalized model is shared by all subjects and will be stronger with the growth of the collected EEG data. Then only few labeled data of a specific subject are required to calibrate the model for personalization. Finally, the model has the ability to decode emotion states from both sound and damaged EEG data for the target individual in the test stage.

The main contributions of this paper are as follows:

- We proposed a Multi-view Spectral-Spatial-Temporal Masked Autoencoder for solving the problems of decoding emotions from few labeled and damaged EEG data towards daily applications.
- Our proposed model is based on a CNN-Transformer hybrid structure which is delicately designed to utilize the spectral, temporal and spacial properties of EEG signals, improving the performance of EEG-based emotion recognition.
- Extensive experiments demonstrate our proposed method can learn the generalized representations of EEG signals from abundant unlabeled EEG data and achieves excellent performance for recognizing emotion states from both flawless and impaired EEG data with only few labeled samples to calibrate.

2 RELATED WORK

In this section, we review the related work in two perspectives: EEG-based emotion recognition and self-supervised learning.

2.1 EEG-based Emotion Recognition

EEG-based emotion recognition draws growing attention currently while progressive studies are conducted to improve its performance. One important stage of EEG-based emotion recognition is feature extraction since EEG signals are complicated neural data. Before the deep learning methods are extensively adopted, spectral EEG features were commonly investigated such as power spectral density (PSD)[7], differential entropy (DE) [6] and differential asymmetry (DASM)[19], etc., from which DE feature is proved to be the most precise and stable one [30] in EEG-based emotion recognition tasks.

Along with the popularization of deep learning methods, complex computational models comprised of various processing units are authorized to perform further and deeper feature extraction [13], and progressive studies explore to further interpreting EEG features from different domains: the spectral domain, spacial domain, and temporal domain. Alhagry *et al.* [1] exploited temporal features with a two-layer long short-term memory network. Zhang *et al.* [27] practiced a recurrent neural network (RNN) to learn spacial-temporal representation from EEG signals. Zhong *et al.* [31] proposed regularized graph neural networks taking account of the topological structure of EEG channels for EEG-based emotion recognition. Li *et al.* [14] practiced a multi-domain adaptive graph convolutional network to exploit complementary knowledge between the temporal and spatial domains of EEG signals. Although those studies in a supervised manner successfully promoted the performance of EEG-based emotion recognition, they all need labeled and sound EEG data which are relatively difficult to deploy in daily applications.

2.2 Self-supervised Learning

Self-supervised learning can interpret feature representations from unlabeled data, driving advances of technologies in the big data era. Considerable models based on self-supervised learning emerge during the past decades, including the models with contrastive learning and generative learning, and so on [18].

Contrast learning is a technique of learning common attributes between data classes and distinguishing one data class from another by contrasting samples, which enhances the performance in many visual tasks. For example, Chen *et al.*[5] came up with SimCLR as simplified contrastive self-supervised learning of visual representations without requiring specialized architectures or a memory bank, achieving excellent performance in computer vision. For EEG data, Kostas *et al.* [12] adapted contrastive self-supervised learning to learn the compressed representation of EEG data. Banville *et al.* [2] investigated self-supervised learning to learn representations of EEG signals based on temporal context prediction as well as contrastive predictive coding on the problems of EEG-based sleep staging and pathology detection. Jiang *et al.*[11] proposed a self-supervised contrastive learning method of EEG signals for sleep stage classification. Furthermore, Mohsenvand *et al.* [20] extended SimCLR to time-series EEG signals for emotion recognition. Most of the existing methods are based on contrast learning, and various transformations in the proxy task are designed for time-series raw EEG signals.

Generative learning with auto-encoding models can reconstruct inputs from the original or corrupted inputs and construct the representation distribution at a point-wise level such as pixels in images and nodes in graphs, which is suitable and applicable in EEG-based emotion recognition for recognizing emotions from few labeled and damaged EEG data. The proxy task of reconstructing the masked EEG channels is not only suitable for all kinds of EEG data such as the pre-extracted spectral features but also can solve the problem that EEG channels are easily corrupted or contaminated during actual use. Masked autoencoder is a general denoising autoencoder and is robust to the introduction of noise [26]. He *et al.* [9] presented masked autoencoders (MAE) as scalable self-supervised learners for computer vision to reconstruct the missing patches in images. Nevertheless, MAE cannot be directly applied well with EEG data

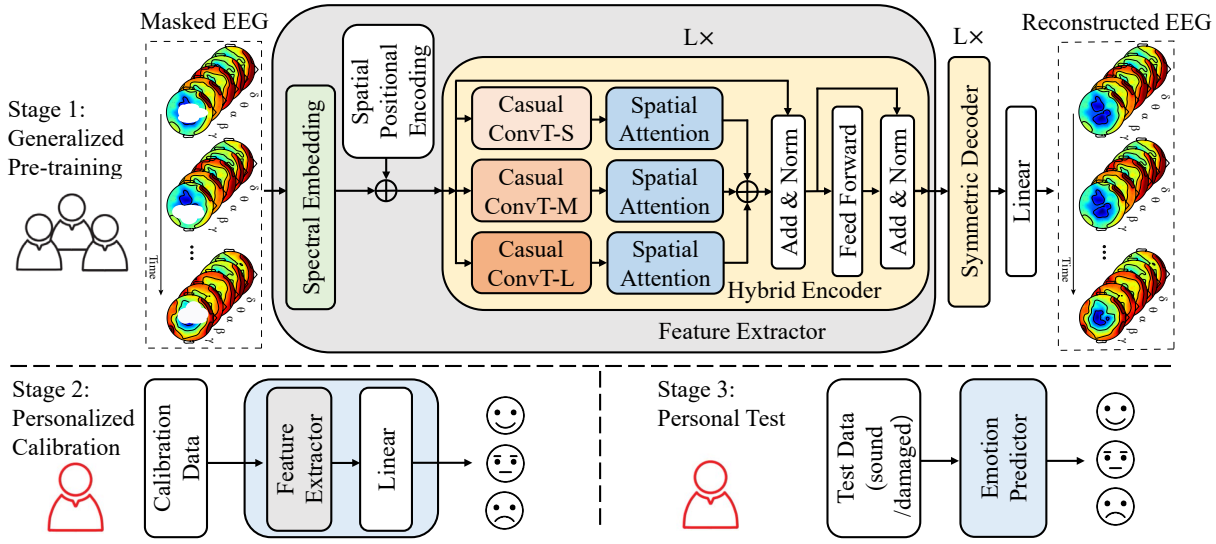


Figure 1: The overall process of our proposed Multi-view Spectral-Spatial-Temporal Masked Autoencoder based on a multi-view CNN-Transformer hybrid structure.

considering the complication of EEG signals, and it doesn't take temporal properties into account.

3 METHODOLOGY

3.1 Formulation

The pre-training dataset is concatenated by the unlabeled training data of all subjects, which is represented as $\mathrm{X} = \{\mathrm{X}_1, \ldots, \mathrm{X}_S\}$, where S denotes the number of the subjects. The concatenated EEG features are the extracted spectral features and can also be represented by a sequence $\mathrm{X} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_N) \in \mathbb{R}^{N \times C \times F}$, where N is the number of samples in time series, C denotes the number of EEG channels, and F represents the set of frequency bands (δ: 1-4 Hz, θ: 4-8 Hz, α: 8-14 Hz, β: 14-31 Hz, and γ: 31-50Hz) converted by a short-time Fourier transform (STFT) in the spectral domain. The pre-trained general feature extractor is denoted as E, and the calibrated emotion predictor for a specific subject s is represented as $\hat{E}_s$, where s denotes the s-th subject. X_s^C and Y_s^C represent the calibration data and label, respectively. The test data and label are denoted as X_s^T and Y_s^T for the subject s.

3.2 Overview

We design a Multi-view Spectral-Spatial-Temporal Masked Autoencoder (MV-SSTMA) based on a multi-view CNN-Transformer hybrid structure, which is depicted in Figure 1. The whole model can be divided into three stages: a generalized pre-training stage, a personalized calibration stage, and a personal test stage. In the pre-training stage, channels of unlabeled EEG data X from all subjects are randomly masked and later reconstructed to learn the general information extracted by the feature extractor E which is shared by all subjects. In the personalized calibration stage, only few labeled data X_s^C and Y_s^C from a specific subject s are used to calibrate the personal emotion predictor $\hat{E}_s$ from the pre-trained generalized feature extractor E. In the test stage, the sound EEG data as well as the damaged data X_s^T could be decoded to recognize the emotion states by $\hat{E}_s$. The algorithm of the overall process is shown as Algorithm 1.

3.3 Generalized Pre-training

In this stage, we pre-train the generalized feature extractor E, which learns the knowledge of unlabeled EEG data from all subjects aiming to better recognize emotion states for a specific subject later. To deal with the problems of decoding emotions from few and damaged EEG data, we choose the generative learning of reconstructing the masked EEG channels as the proxy task to learn the general representations of EEG data. Considering characteristics of EEG signals, we design the pre-training model based on a multi-view CNN-Transformer hybrid structure, which consists of a spectral embedding layer, a spatial positional encoding layer, L hybrid encoders, and L symmetric hybrid decoders. Each hybrid block includes a temporal multi-scale casual convolution layer and a spatial multi-head self-attention layer.

3.3.1 Spectral Embedding and Position Encoding. As the differential entropy (DE) feature has been proved to have excellent performance in EEG-based emotion recognition tasks [6], we use the DE feature extracted from the EEG signals in the spectral domain as the inputs of the model. The extracted DE feature $\mathrm{X} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_N) \in \mathbb{R}^{N \times C \times F}$ are transformed into samples $\tilde{\mathrm{X}} = (\tilde{\mathrm{x}}_1, \tilde{\mathrm{x}}_2, \ldots, \tilde{\mathrm{x}}_N) \in \mathbb{R}^{N \times C \times T \times F}$ with an overlapping window of T seconds. For each sample i, $\tilde{\mathrm{x}}_i \in \mathbb{R}^{C \times T \times F}$. In the spectral embedding layer, we first project $\tilde{\mathrm{x}}_i$ into a D- dimensional space via a linear layer to embed the spectral information of EEG signals. Thus $\tilde{\mathrm{x}}_i$ is embedded into

the shape of $C \times T \times D$, which is formulated as follows:

$$\tilde{\mathrm{x}}_i^{(s)} = \tilde{\mathrm{x}}_i \mathrm{w}^{(s)} + \mathrm{b}^{(s)}, \tag{1}$$

where the weight vector $\mathrm{w}^{(s)} \in \mathbb{R}^{F \times D}$ and bias $\mathrm{b}^{(s)} \in \mathbb{R}^{D}$.

For the spatial positional encoding layer, we divide EEG data into patches according to the EEG channels in the dimension C. One patch represents one EEG channel. To remember the position of each EEG channel and reconstruct it later, the sine-cosine positional encoding is added on C in the spatial dimension.

For the masking step, we randomly sample a visible subset $\tilde{\mathrm{x}}_i^v \in \mathbb{R}^{C_v \times T \times D}$ and mask subset $\tilde{\mathrm{x}}_i^m \in \mathbb{R}^{C_m \times T \times D}$, where $C_v \cup C_m = C$. Only the $\tilde{\mathrm{x}}_i^v$ is employed as the input of the hybrid encoder.

3.3.2 Temporal Multi-scale Casual Convolution. To capture the temporal information of the EEG signals, the multi-scale Casual Convolution layer is introduced to enable the model to learn dynamic temporal representations. We implement three branches of casual convolution layers with long, medium, and short kernel sizes, corresponding to the blocks of CasualConvT-L, CasualConvT-M, and CasualConvT-S in Figure 1. We aim to calculate a temporal brain summary for each EEG channel from the input spectral feature.

Multi-scale. The multi-scale Casual Convolution layers employ casual convolution with multiple lengths of convolutional kernels to capture different ranges of timesteps. The short temporal kernel aims to learn the short-term representations while the long temporal kernel is employed to extract long-term representations. From the multi-scale temporal kernels, the diverse representations of EEG data can be enriched and the emotion-related information can be learned fully. The dynamic long short-term temporal patterns are generated by applying the multi-scale temporal kernels in parallel on the input EEG samples. The temporal convolutional kernel size $k_t \times 1$ of CasualConvT-L, CasualConvT-M, CasualConvT-S are set as $k_l \times 1$, $k_m \times 1$, and $k_s \times 1$, respectively.

Channel-wise Casual Convolution. Unlike temporal images of videos, time series of EEG signals are represented as contiguous sequences of every single channel. Therefore, for each scale branch, we calculate a temporal brain summary $\widetilde{\mathrm{B}}_t$ for each channel $c \in \{1, \ldots, C\}$ where the embedding of c is updated by the adjacent frames of the same channel. More explicitly, the convolution operation with the kernel size of $K_t \times 1$ is performed on the temporal dimension T of the input $\mathrm{B}_{in} \in \mathbb{R}^{C_v \times T \times C}$ in each EEG channel. Here, K_t ensures the temporal information is encoded in the neighborhood.

Further, causal convolutions are used to force no information flow from the future to the past. As shown in Figure 2, the output at time t depends only on inputs from time t and earlier. The channel-wise temporal convolution implemented in our model does not change the shape of the vector, so the zero-padding with the length of $K_t - 1$ is added to keep the shape unchanged. The temporal convolution for three scale branches in our model can be formulated as :

$$\widetilde{\mathrm{B}}_t^s = \mathrm{BN}(\mathrm{CasualConvT}(\mathrm{B}_{in}, (k_s, 1))), \tag{2}$$

$$\widetilde{\mathrm{B}}_t^m = \mathrm{BN}(\mathrm{CasualConvT}(\mathrm{B}_{in}, (k_m, 1))), \tag{3}$$

$$\widetilde{\mathrm{B}}_t^l = \mathrm{BN}(\mathrm{CasualConvT}(\mathrm{B}_{in}, (k_l, 1))), \tag{4}$$

where $\mathrm{B}_{in} \in \mathbb{R}^{C_v \times T \times C}$ is the input spectral features and is set to $\tilde{\mathrm{x}}_i^v \in \mathbb{R}^{C_v \times T \times D}$ in the first layer. $\widetilde{\mathrm{B}}_t^s$, $\widetilde{\mathrm{B}}_t^m$ and $\widetilde{\mathrm{B}}_t^l$ are the temporal summary of different scales retaining the same shape of $C_v \times T \times D$, and BN is the batch normalization operation to maintain the stability of the model.

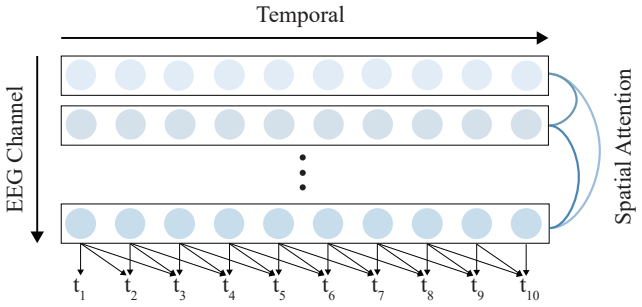


Figure 2: The process of the spatial attention between EEG channels and the channel-wise casual convolution when the temporal kernel size is set to 3×1.

3.3.3 Spatial Multi-head Self-Attention. Following the temporal convolutional layer, the spatial multi-head self-attention is employed to learn dynamics and inter-channel dependencies from all visible EEG channels, as shown in Figure 2. For the long scale branch, reshaping the $\widetilde{\mathrm{B}}_t^l$ into the shape of $C_v \times TD$, then the temporal brain embedding can be represent as $\widetilde{\mathrm{B}}_t^l = \left[b_1^l, \ldots, b_{C_v}^l\right] \in \mathbb{R}^{C_v \times TD}$. We utilize the scaled dot-product [25] to explicitly capture the topological relationships between EEG channels, which is formulated as:

$$Attention\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{TD}}\right), \tag{5}$$

where Q, K, and V denote the query vector, key vector and value vector, respectively and TD is the dimension of the key vector which is employed to scale the dot products.

The dot-product similarity is evaluated between the query representation Q of the channel of interest with K. If Q and K are similar, meaning high attention weight, then the corresponding value is assumed relevant. The Q, K and V vectors here are the projections of the input brain embeddings $\widetilde{\mathrm{B}}_t^l$. Specifically, the spatial brain summary for the long scale branch, denoted as B_s^l, is calculated as the attention weight between EEG channels by the multi-head attention [25]:

$$Q_{(i)}^l = \left[b_1^l, \ldots, b_{C_v}^l\right] W_{(Q,i)}^l, Q_{(i)}^l \in \mathbb{R}^{C_v \times M}, \tag{6}$$

$$K_{(i)}^l = \left[b_1^l, \ldots, b_{C_v}^l\right] W_{(K,i)}^l, K_{(i)}^l \in \mathbb{R}^{C_v \times M}, \tag{7}$$

$$V_{(i)}^l = \left[b_1^l, \ldots, b_{C_v}^l\right] W_{(V,i)}^l, V_{(i)}^l \in \mathbb{R}^{C_v \times M}, \tag{8}$$

$$head_i^l = Attention\left(Q_{(i)}^l, K_{(i)}^l, V_{(i)}^l\right) = A^l V_{(i)}^l, \tag{9}$$

$$B_s^l = Concat\left(head_1^l, \ldots, head_H^l\right) W_{(O)}^l, \tag{10}$$

where $head_i^l \in \mathbb{R}^{C_v \times M}, i \in \{1, \ldots, H\}$ and $M = TD/H$. $W_{(O)}^l \in \mathbb{R}^{HM \times TD}$ is the the weight matrix for concatenating and projecting the results from multi-head back into representation space. The

spatial attention matrix $A^l \in \mathbb{R}^{C_v \times C_v}$ denotes how much attention a channel pays to the other channel.

The other two branches are processed the same as the long scale branch. The spatial brain embedding in three scale branches are fused by the sum operation:

$$\widetilde{B}^{(C_v)} = B_s^s + B_s^m + B_s^l, \tag{11}$$

where B_s^s, B_s^m, and B_s^l denote the outputs of the spatial attention layers in the short scale branch, the median scale branch, and the long scale branch, respectively and $\widetilde{B}^{(C_v)}$ represents the overall spatial brain summary of three scale branches from all visible EEG channels. After the spatial attention, the layer normalization and the feed-forward network are following. There are L CNN-Transformer hybrid encoders stacked to update the embeddings and further extract the EEG features. The final embedding is denoted as $\tilde{\mathrm{x}}_i^{ve} \in \mathbb{R}^{C_v \times T \times D}$.

3.3.4 Decoder and Reconstruction. After the feature extractor, we apply the symmetric decoder to reconstruct the masked EEG channels, which consists of L similar CNN-Transformer hybrid blocks and a linear layer. The symmetric structure of the encoder-decoder is designed for a stronger decoder to reconstruct complicated EEG data. The input to the decoder is the full set consisting of the encoded visible channels $\tilde{\mathrm{x}}_i^{ve} \in \mathbb{R}^{C_v \times T \times D}$ and the masked channels $\tilde{\mathrm{x}}_i^{me} \in \mathbb{R}^{C_m \times T \times D}$. $\tilde{\mathrm{x}}_i^{me}$ is set as parameters that are randomly initialized and concatenated with the encoded visible channels. The decoder outputs the reconstructed EEG feature $\tilde{\mathrm{x}}_i^{rec} \in \mathbb{R}^{C \times T \times F}$. The reconstruction process predicts values for each masked EEG channel. The loss is only computed between reconstructed patches $\tilde{\mathrm{x}}_{im}^{rec} \in \mathbb{R}^{C_m \times T \times F}$ for the masked channels and the corresponding original EEG features by the mean squared error (MSE). Finally, we obtained a pre-trained general feature extractor E by minimizing the reconstruction loss.

3.4 Personalized Calibration & Test

For a specific subject s, the calibration data are composed of few labeled samples from each kind of emotion state in the original training dataset of the subject s, which are denoted as X_s^C and Y_s^C. Since the EEG data are chronologically recorded, it is reasonable to take the data from the very beginning of the training dataset as the calibration data. We obtain a personalized calibrated emotion predictor $\hat{E}_s$ by fine-tuning the generalized feature extractor E, followed by a linear layer to predict the emotion class. We measure the classification loss by cross-entropy.

In the test stage, our model accepts both sound and damaged EEG data. We use the test set of the subject s from the original test dataset, denoted as X_s^T and Y_s^T to verify the effectiveness of the personalized model $\hat{E}_s$. To simulate the impaired data, we masked channels the same way as the pre-training stage does.

Algorithm 1: The process of the Multi-view Spectral-Spatial-Temporal Masked Autoencoder.

Input:
The pre-training data $\mathrm{X} = \{\mathrm{X}_1, \ldots, \mathrm{X}_S\}$.
The calibration data X_s^C and label Y_s^C.
The test data X_s^T for the target subject s.
Output:
The generalized feature extractor E.
The personalized emotion predictor $\hat{E}_s$.
The predicted emotion class $\hat{\mathrm{Y}}_s^T$.
Generalized Pre-training Stage:
Randomly initialize E.
Mask the Pre-training data: $\mathrm{X}_{mask} = mask(\mathrm{X})$.
Reconstruct the input data: $\mathrm{X}_{rec} = reconstruct(\mathrm{X}_{mask})$.
Optimize E by minimizing the reconstruction MSE loss: $loss_{rec} = MSE(\mathrm{X}, \mathrm{X}_{rec})$.
return E.
Personalized Calibration Stage:
Initialize $\hat{E}_s$ with E.
Predict the Emotion class: $\hat{\mathrm{Y}}_s^C = \hat{E}_s(\mathrm{X}_s^C)$.
Fine-tune $\hat{E}_s$ by minimizing the classification loss: $loss_{cls} = CrossEntropy(\mathrm{Y}_s^C, \hat{\mathrm{Y}}_s^C)$.
return $\hat{E}_s$.
Test Stage :
Predict the emotion class: $\hat{\mathrm{Y}}_s^T = \hat{E}_s(\mathrm{X}_s^T)$.
return $\hat{\mathrm{Y}}_s^T$.

4 EXPERIMENTS

4.1 Datasets

Our proposed model is evaluated on two popular affective EEG datasets (SEED [29] and SEED-IV [28]). The stimuli materials of these datasets are all video clips.

SEED Dataset contains EEG signals of 15 participants divided into three emotion states including positive, neutral, and negative. Every subject conducted three sessions of 15 trials each at different times. In each session, the first 9 trials are usually used as training data and the remaining 6 trials are used as test data [29].

SEED-IV Dataset is collected for four emotion states: happy, sad, fear, and neutral emotions. 15 subjects participated in three sessions on different days with 24 trials each. In general, the first 16 trials are training data and the remaining 8 ones are the test data for each session [28].

4.2 Implementation Details

To make our results comparable, we follow the same common experimental settings as the prior studies [14, 15, 17, 23, 27, 29, 31] on two datasets, whose performance is evaluated by averaged accuracy and standard deviation over the sessions, as the classes in the datasets are balanced. For each experiment, our pre-training data X are concatenated from unlabeled original training data of all subjects, composed of the 9 trials for the SEED dataset and 16 trials for the SEED-IV dataset. Few labeled data with the number of 10,20,30 for each emotion state from the beginning of the training dataset of the target subject are used for calibration.

The pre-training data $\mathrm{X} \in \mathbb{R}^{N \times F \times C}$ are transformed into $\tilde{\mathrm{X}} \in \mathbb{R}^{N \times F \times T \times C}$ by an overlapping window with the size of T to keep the same sample size N as the compared experiments, where T is set to 10 samples and C is the number of the EEG channels, which is equal to 62. The experiments utilized PyTorch [22] deep learning

Table 1: Average accuracies and standard deviations (acc/std %) of our model and baselines in SEED and SEED-IV datasets when using all labeled training data.

Model	Dataset			
	SEED		SEED-IV	
	Acc.	Std.	Acc.	Std.
STRNN	89.50	7.63	-	-
DGCNN	90.40	8.49	69.88	16.29
BiDANN	92.38	7.04	70.29	12.63
BiHDM	93.12	6.06	74.35	14.09
R2G-STNN	93.34	5.96	-	-
RGNN	94.24	5.95	79.37	10.54
MD-AGCN	94.81	4.52	87.63	5.77
MAE	92.27	5.19	87.81	5.36
MV-SSTMA	**95.32**	**3.05**	**92.82**	**5.03**

Table 2: Average accuracies and standard deviations (acc/std %) of our model and baselines in SEED and SEED-IV datasets when using few labeled training data.

Model	# Labeled Data	Dataset			
		SEED		SEED-IV	
		Acc.	Std.	Acc.	Std.
MD-AGCN	10	63.02	4.59	62.25	3.15
	20	65.07	4.49	64.84	3.07
	30	65.51	5.23	65.28	3.40
MAE	10	76.50	9.16	73.73	8.77
	20	77.18	8.29	75.24	9.40
	30	77.54	8.53	76.11	7.74
MV-SSTMA	10	**81.40**	7.67	**80.22**	7.66
	20	**81.94**	6.90	**80.94**	7.06
	30	**83.49**	6.35	**82.55**	6.44

framework. The learning rate of our model range from 0.001 to 0.00001 for each experiment. Moreover, the spectral embedding size D is set to 16, and the number of the hybrid block L equals to 6. The multi-head dimension H is set to 6. The compared MAE[9] method implemented in this paper follows a similar design as our MV-SSTMA to fit the EEG data, calculating the spatial attention between EEG channels but removing the multi-view and temporal aspects of MV-SSTMA.

4.3 Baseline Models

- STRNN[27]: The spatial-temporal recurrent neural network is based on a unified spatial-temporal dependency model that learns the information from both spatial and temporal domains of EEG signals.
- DGCNN[23]: Dynamical graph convolutional neural network learns the representations of EEG signals by graph convolution in a dynamic way for EEG-based emotion recognition.
- BiDANN[16]: Bi-hemispheres domain adversarial neural network focuses on discriminative features of EEG signals from both the right and left sides of the hemispheres of the brain for EEG-based emotion recognition.
- BiHDM[15]: Bi-Hemispheres discrepancy model investigates the asymmetric differences of the right and left hemispheres of the brain.
- R2G-STNN[17]: A region to global spatial-temporal neural network model learns the global and regional EEG representations in both spatial and temporal aspects of EEG signals.
- RGNN[31]: Regularized graph neural network explores the topology of EEG channels with graph convolution.
- MD-AGCN[14]: A multi-domain adaptive graph convolutional network taking full advantages of features on different domains.
- MAE[9]: Masked autoencoders as scalable self-supervised learners by reconstructing the missing patches in images for computer vision.

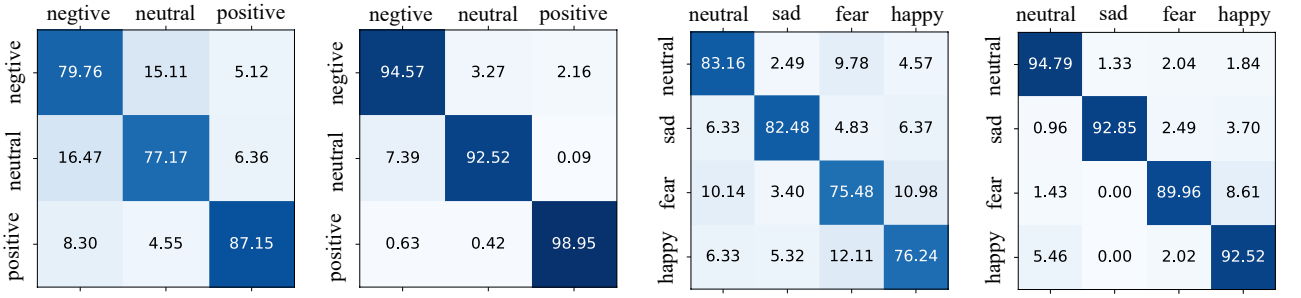
4.4 Results Analysis and Comparison

4.4.1 Results for Different Numbers of Calibration Data. We present the results of comparison between our model and advanced baseline models using all labeled and few (with numbers of 10, 20, and 30 for each kind of emotion state) labeled training data on SEED and SEED-IV datasets in Table 1 and Table 2, respectively. For each kind of emotions, the 10, 20, and 30 labeled data are from the beginning of the trial in the same period. To be noted, our results are only compared with the advanced models that follow the same common experimental settings.

Table 1 shows that our model achieves state-of-the-art results in comparison with supervised methods on both SEED and SEED-IV datasets, which indicates pre-training process can improve the generalization and efficiency of the model, especially on the problems with more emotion classes. Specifically, the recognition accuracy of our model reaches 95.32% with a standard deviation of 3.05% on the SEED dataset. On the SEED-IV dataset, our model achieves significant improvement with the highest accuracy of 92.82% and the lowest standard deviation of 5.03%. Furthermore, the MAE method also exceeds baseline methods on SEED-IV, whereas performs worse than some supervised models on SEED. The reason might be that those supervised models take temporal information of EEG into account.

In the case of only few labeled data for calibration, the proposed MV-SSTMA is evaluated with the self-supervised method MAE and the supervised method MD-AGCN. The column of # Labeled Data in Table 2 means the number of labeled training data for each kind of emotion state we use to calibrate the model. The accuracy increases while more labeled data are used for all models. The increment is minor might because labeled data of different numbers are adjacent and token from the same period implying the lack of diversity. In addition, our model outperforms MAE and MD-ADCN in every scenario.

The Paired t-test is also conducted on the performance of MV-SSTMA and MAE for all subjects in all situations above, as well as performance of MV-SSTMA and MD-AGCN. The significance levels are much lower than 1% in all cases, indicating the significant differences between them.



a) 10 Labeled Data - SEED b) All Labeled Data - SEED c) 10 Labeled Data - SEED-IV d) All Labeled Data - SEED-IV

Figure 3: The confusion matrices on SEED and SEED-IV datasets with 10 and all labeled training data to calibrate for MV-SSTMA. Each column represents the predicted class that our model outputs and each row serves as the target class.

Table 3: Average accuracies and standard deviations (acc/std %) of our model and baselines on SEED-IV for different rates of impaired channels with 10 labeled training data.

Model	Impaired rates		
	30%	50%	70%
MD-AGCN	57.32/4.05	56.91/4.11	56.26/4.25
MAE	66.86/7.99	64.73/8.01	59.93/7.38
MV-SSTMA	**73.68**/7.58	**70.81**/6.58	**65.11**/5.64

In general, two components may contribute to the improvement in the performance: 1) The pre-training stage captures the generalization representation of EEG signals while the calibration process specifies the model to individuals overcoming subject domain shift. 2) Our proposed model thoroughly utilizes EEG signals on the spectral, temporal, and spatial domains.

4.4.2 Results for Impaired Channels. We demonstrate the results of different rates of impaired channels in test data with 10 labeled calibration data on the SEED-IV dataset in Table 3. Each column stands for the percentage of impaired channels in test data. The reason for employing the SEED-IV dataset is that the four emotion categories in SEED-IV include all three emotion states in SEED. From Table 3, it is apparent that when 30% of channels in test data are corrupted, emotion states can be well recognized with MV-SSTMA, achieving the accuracy of 73.68% and 7.58% standard deviations with only 10 labeled calibration data. In addition, even with more damaged EEG channels, our model can still distinguish emotion states well.

4.5 Ablation Study

To demonstrate the effect of the channel-wise casual convolutional layer in the hybrid encoder block, we implement the ablation study by replacing the channel-wise casual convolutional layer with a temporal embedding, namely NoHybrid. In the NoHybrid model, the temporal information is still considered by adding temporal embedding in the original spectral embedding layer, but it cannot be viewed interchangeably with spatial information in L encoder blocks. We also implement the ablation study by reducting the multi-scale temporal branches of MV-SSTMA, that only employ one single scale branch in the model, named SingleScale. The casual convolution in the SingleScale model is also replaced by normal convolution operation to evaluate the contribution of the casual convolution.

Table 4: Ablation Study for the classification performance (acc/std %) with different numbers of labeled calibration data on SEED and SEED-IV datasets.

Model	# Labeled Data	Dataset			
		SEED		SEED-IV	
		Acc.	Std.	Acc.	Std.
NoHybrid	10	76.60	6.24	75.95	8.01
	20	78.09	7.51	76.50	5.62
	30	78.21	7.44	77.05	7.54
	All	94.44	4.67	88.44	8.13
SingleScale	10	80.21	6.19	78.97	5.99
	20	80.24	6.60	79.44	7.34
	30	81.03	5.75	79.55	7.90
	All	95.01	3.32	91.44	5.88
MV-SSTMA	10	**81.40**	7.67	**80.22**	7.66
	20	**81.94**	6.90	**80.94**	7.06
	30	**83.49**	6.35	**82.55**	6.44
	All	**95.32**	3.05	**92.82**	5.03

Table 4 exhibits the performance of our MV-SSTMA, the model NoHybrid, and the model SingleScale with different numbers of calibration labeled data for each kind of emotion state on SEED and SEED-IV datasets. The fact that our model always surpasses NoHybrid model and SingleScale model indicates the importance of the channel-wise casual convolutional layer in the hybrid encoder block and the multi-scale branches with casual convolution. Moreover, as the temporal information is also considered in the NoHybrid model and SingleScale model, their performance are still better than MAE.

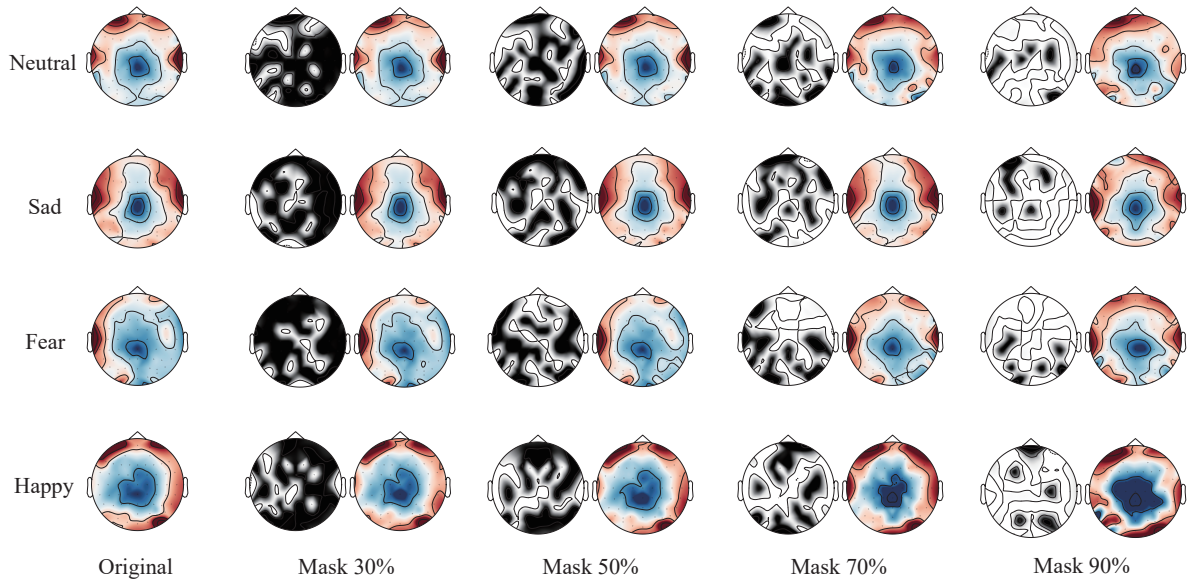


Figure 4: The topographic maps of the reconstructed test data with different mask rates for four emotion states in the γ band. The row denotes different emotion states. The first column is the original data, while the rest columns stand for the position layout of the masked EEG channels and reconstructed data with different mask rates. Color white and black in position layout mean the masked and visible areas, respectively.

4.6 Visualization

Figure 3 presents the confusion matrices of MV-SSTMA with 10 and all labeled training data to calibrate on both SEED and SEED-IV datasets, illustrating the ability to discriminate each emotion state.

For the SEED dataset, our model can recognize positive emotion state best and hardest to recognize the neutral emotion state on both 10 and all labeled training data to calibrate. For the SEED-IV dataset, fear is most difficult emotion state to recognize with 10 labeled calibration data, while the neutral state is the easiest one to recognize. Moreover, when all labeled training data are adopted to calibrate, our model still decodes the neutral state better than all other three emotion states and the fear state is also the most difficult one to discriminate.

We further investigate the ability of our model for reconstructing the impaired EEG channels from the test data. Figure 4 illustrates the reconstructed test data, which was manually damaged by masking the EEG channels randomly with different mask rates. Here we take the four emotion states in the γ band as an example, as the γ band is proved to be the most effective frequency band for emotion recognition [29]. From Figure 4, we can see that when the mask rate is 30% and 50%, the EEG features can be well reconstructed. With the 70% mask rates, the features can also be reconstructed in general, but some details might be lost. Nevertheless, when the mask rate is 90%, the EEG features are much more difficult to recover.

5 CONCLUSIONS

In this paper, we propose a Multi-view Spectral-Spatial-Temporal Masked Autoencoder with self-supervised learning to solve the problems of decoding emotions from few labeled and damaged EEG data. Our model takes full advantage of EEG signals by exploring spectral, spatial, and temporal properties of EEG data through the multi-view CNN-Transformer hybrid structure. Three stages of pre-training, calibrating, and testing ensure the generalization, personalization, and efficiency characteristics of the overall framework.

Extensive experiments on the SEED and SEED-IV datasets demonstrate the outstanding performance of our model compared with various advanced baseline models. Results for few labeled and impaired EEG data demonstrate the proposed MV-SSTMA model can learn the EEG representations from abundant unlabeled data and effectively decode emotion states from few labeled and even damaged EEG data. The visualization of reconstructing the damaged EEG channels on the test data demonstrates the effectiveness and the ability of our model to recover the missing channels of emotional EEG data. In general, our model promotes the performance of EEG-based emotion recognition in a self-supervised manner.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Natural Science Foundation of China (No. 61976135), MOST 2030 Brain Project (No. 2022ZD0208500), Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX), SJTU Global Strategic Partnership Fund (2021 SJTU-HKUST), Shanghai Marine Equipment Foresight Technology Research Institute 2022 Fund (No. GC3270001/012), and GuangCi Professorship Program of RuiJin Hospital Shanghai Jiao Tong University School of Medicine.

REFERENCES

[1] Salma Alhagry, Aly Aly Fahmy, and Reda A El-Khoribi. 2017. Emotion recognition based on EEG using LSTM recurrent neural network. *Emotion* 8, 10 (2017), 355–358.

[2] Hubert Banville, Omar Chehab, Aapo Hyvärinen, Denis-Alexander Engemann, and Alexandre Gramfort. 2021. Uncovering the structure of clinical EEG signals with self-supervised learning. *Journal of Neural Engineering* 18, 4 (2021), 046020.

[3] Andrey V Bocharov, Gennady G Knyazev, and Alexander N Savostyanov. 2017. Depression and implicit emotion processing: An EEG study. *Neurophysiologie Clinique/Clinical Neurophysiology* 47, 3 (2017), 225–230.

[4] Clemens Brunner, Niels Birbaumer, Benjamin Blankertz, Christoph Guger, Andrea Kübler, Donatella Mattia, José del R Millán, Felip Miralles, Anton Nijholt, Eloy Opisso, et al. 2015. BNCI Horizon 2020: towards a roadmap for the BCI community. *Brain-Computer Interfaces* 2, 1 (2015), 1–10.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*. PMLR, 1597–1607.

[6] Ruo-Nan Duan, Jia-Yi Zhu, and Bao-Liang Lu. 2013. Differential entropy feature for EEG-based emotion classification. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 81–84.

[7] Lester I Goldfischer. 1965. Autocorrelation function and power spectral density of laser-produced speckle patterns. *Josa* 55, 3 (1965), 247–253.

[8] Matti Hämäläinen, Riitta Hari, Risto J Ilmoniemi, Jukka Knuutila, and Olli V Lounasmaa. 1993. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics* 65, 2 (1993), 413.

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377* (2021).

[10] Ziyu Jia, Youfang Lin, Xiyang Cai, Haobin Chen, Haijun Gou, and Jing Wang. 2020. SST-EmotionNet: Spatial-spectral-temporal based attention 3d dense network for EEG emotion recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2909–2917.

[11] Xue Jiang, Jianhui Zhao, Bo Du, and Zhiyong Yuan. 2021. Self-supervised Contrastive Learning for EEG-based Sleep Staging. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[12] Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. 2021. BENDR: using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *arXiv preprint arXiv:2101.12037* (2021).

[13] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[14] Rui Li, Yiting Wang, and Bao-Liang Lu. 2021. A Multi-Domain Adaptive Graph Convolutional Network for EEG-based Emotion Recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5565–5573.

[15] Yang Li, Lei Wang, Wenming Zheng, Yuan Zong, Lei Qi, Zhen Cui, Tong Zhang, and Tengfei Song. 2020. A novel bi-hemispheric discrepancy model for EEG emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems* 13, 2 (2020), 354–367.

[16] Yang Li, Wenming Zheng, Zhen Cui, Tong Zhang, and Yuan Zong. 2018. A Novel Neural Network Model based on Cerebral Hemispheric Asymmetry for EEG Emotion Recognition.. In *IJCAI*. 1561–1567.

[17] Yang Li, Wenming Zheng, Lei Wang, Yuan Zong, and Zhen Cui. 2019. From regional to global brain: A novel hierarchical spatial-temporal neural network model for EEG emotion recognition. *IEEE Transactions on Affective Computing* (2019).

[18] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. 2021. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[19] Yisi Liu and Olga Sourina. 2013. Real-time fractal-based valence level recognition from EEG. In *Transactions on computational science XVIII*. Springer, 101–120.

[20] Mostafa Neo Mohsenvand, Mohammad Rasool Izadi, and Pattie Maes. 2020. Contrastive representation learning for electroencephalogram classification. In *Machine Learning for Health*. PMLR, 238–253.

[21] Femke Nijboer, Fabrice O Morin, Stefan P Carmien, Randal A Koene, Enrique Leon, and Ulrich Hoffmann. 2009. Affective brain-computer interfaces: Psychophysiological markers of emotion in healthy persons and in persons with amyotrophic lateral sclerosis. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 1–11.

[22] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic Differentiation in Pytorch. (2017).

[23] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. 2018. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing* 11, 3 (2018), 532–541.

[24] Babak A Taheri, Robert T Knight, and Rosemary L Smith. 1994. A dry electrode for EEG recording. *Electroencephalography and Clinical Neurophysiology* 90, 5 (1994), 376–383.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[26] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*. 1096–1103.

[27] Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li. 2018. Spatial–temporal recurrent neural network for emotion recognition. *IEEE Transactions on Cybernetics* 49, 3 (2018), 839–847.

[28] Wei-Long Zheng, Wei Liu, Yifei Lu, Bao-Liang Lu, and Andrzej Cichocki. 2018. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics* 49, 3 (2018), 1110–1122.

[29] Wei-Long Zheng and Bao-Liang Lu. 2015. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development* 7, 3 (2015), 162–175.

[30] Wei-Long Zheng, Jia-Yi Zhu, and Bao-Liang Lu. 2017. Identifying stable patterns over time for emotion recognition from EEG. *IEEE Transactions on Affective Computing* 10, 3 (2017), 417–429.

[31] Peixiang Zhong, Di Wang, and Chunyan Miao. 2020. EEG-based emotion recognition using regularized graph neural networks. *IEEE Transactions on Affective Computing* (2020).