Copyright © 2024 Elsevier Ltd. All rights reserved. This manuscript version is made available under the CC-BY-NC-ND 4.0 license https://creativecommons.org/licenses/by-nc-nd/4.0/ (see: https://www.elsevier.com/about/policies/sharing).

S^3T-Net: A Novel Electroencephalogram Signals-Oriented Emotion Recognition Model

Weilong Tan[1], Hongyi Zhang[1], Zidong Wang, Han Li, Xingen Gao and Nianyin Zeng*

***Abstract*—In this paper, a novel skipping spatial-spectral-temporal network (S^3T-Net) is developed to handle intra-individual differences in electroencephalogram (EEG) signals for accurate, robust, and generalized emotion recognition. In particular, aiming at the 4D features extracted from the raw EEG signals, a multi-branch architecture is proposed to learn spatial-spectral cross-domain representations, which benefits enhancing the model generalization ability. Time dependency among different spatial-spectral features is further captured via a bi-directional long-short term memory module, which employs an attention mechanism to integrate context information. Moreover, a skip-change unit is designed to add another auxiliary pathway for updating model parameters, which alleviates the vanishing gradient problem in complex spatial-temporal network. Evaluation results show that the proposed S^3T-Net outperforms other advanced models in terms of the emotion recognition accuracy, which yields an performance improvement of 0.23% , 0.13%, and 0.43% as compared to the sub-optimal model in three test scenes, respectively. In addition, the effectiveness and superiority of the key components of S^3T-Net are demonstrated from various experiments. As a reliable and competent emotion recognition model, the proposed S^3T-Net contributes to the development of intelligent sentiment analysis in human-computer interaction (HCI) realm.**

***Index Terms*—Human-computer interaction (HCI); EEG signals; emotion recognition; spatial-temporal network; skip-change unit**

I. Introduction

In the field of human-computer interaction (HCI) [1], it is of vital significance to develop the artificial intelligence (AI)-based intelligent sentiment analysis models, which benefits estimating the human emotional states and understanding their cognitive/behavioral processes. In practice, the subjective feedback, the external expressions, and the physiological signals of the user can greatly influence the performance of the deep learning (DL)-based models, which mainly owes to the objective conditions such as environments and subjectivity in awareness [2]–[7]. Therefore, how to use unstable physiological signals to accurately identify human emotional states has become one of the main challenges posed to current sentiment analysis techniques.

As an effective emotion analysis technique, developing the electroencephalogram (EEG)-oriented DL approaches has gradually becoming a hot spot in emotion recognition areas. In [8], a multivariate time-frequency Granger causality analysis method based on a nonlinear vector auto-regressive model and coiflet wavelet packet decomposition has been proposed, which successfully characterizes the functional connectivity of different frequency bands of cerebral electromyography synchrony in the time-frequency domain. Moreover, the frequency, spatial and temporal information of multi-channel EEG signals have been integrated in [9], and the proposed model is able to learn temporal dependency among different representations via a long-short term memory (LSTM) structure. A bi-hemispheric bias model for emotion recognition has been developed in [10], where the data representations of each left and right hemispheres are learned by presetting the range of the brain electrode maps, which enhances the model stability by learning hemisphere differences in left and right brains.

To take full advantages of the topology of EEG channels, a method to read multi-channel differential entropy features for training hierarchical convolutional neural networks (HCNNs) has been developed in [11], which yields higher emotion recognition accuracy as compared to traditional classifiers. In [12], a regularized graph neural network (RGNN) for EEG-based emotion recognition has been proposed, which simulates the inter-channel relationships among EEG signals by the adjacency matrix in GNN to capture the different channel-wise local and global features. Similarly, a parallel spatial-spectral-temporal-attention mechanism module has been employed to learn 3D spatial-spectral representation of EEG signals, which realizes the adaptive capture of important brain regions, frequency bands, and temporal domains, thereby yielding higher recognition accuracy.

With regard to above successful applications of the EEG signals-oriented emotion recognition methods, it can be concluded that a great many of efforts have been carried out to promoting thorough EEG feature extraction, which endows the DL models with considerable reliability to perceive, understand and regulate emotions. While it should be pointed out that the EEG signal data are relatively lacked in real-world scenes as compared to the great demand of data-driven DL models, which draws attention on the over-fitting phenomenon. Simultaneously, it is also necessary to take sufficient exploration on characteristics and correlations of EEG data in the spatial-spectral-temporal domains to enhance the model gener-

This work was supported in part by the Natural Science Foundation of China under Grant 62073271, the Fundamental Research Funds for the Central Universities of China under Grant 20720220076, the Natural Science Foundation for Distinguished Young Scholars of the Fujian Province under Grant 2023J06010, and the National Science and Technology Major Project under Grant J2019-I-0013-0013.

W. Tan and H. Zhang are with the School of Opto-Electronic and Communication Engineering, Xiamen University of Technology, Fujian 361024, China.

H. Li and N. Zeng are with the Department of Instrumental and Electrical Engineering, Xiamen University, Fujian 361105, China.

Corresponding authors: N. Zeng. Email: zny@xmu.edu.cn

[1] : These authors contributed equally to the work.

alization ability, which benefits the popularization of applying the DL models in different experimental environments.

Based on the above discussion, a novel 4D EEG signals-oriented skipping spatial-spectral-temporal network (S^3T-Net) is developed for emotion recognition in this paper. Particularly, to comprehensively capture the intrinsic correlations among various EEG features so as to enhance the model generalization ability, a multi-branch (MB) architecture and a bi-directional LSTM (Bi-LSTM) module with an attention mechanism are built to simultaneously extract relevant features in the temporal, spectral, and spatial domains from the raw EEG data. When extracting the multi-scale features of brain maps, the small reception field branch in MB focuses on dealing with specific details, and the other large reception field branch is responsible for learning the global information and, by doing so, different features can be fused to generate more robust spectral-spatial representations. Then, the temporal dependence among above strong representations across different time slices can be captured by the Bi-LSTM module, where the attention mechanism makes it able to automatically learn important information. Moreover, a skip-change unit is designed to address the over-fitting phenomenon in training stage, which alleviates the gradient vanishing issue in MB architecture during training, thereby improving the model generalization ability. More information can be found in Section III

The main contributions of this paper are outlined as follows:

1. A novel emotion recognition model S^3T-Net is proposed, which can take comprehensive and in-depth explorations on the EEG signals to pursue strong robustness and generalization.

2. Spatial and spectral-domain information can be effectively captured via the designed multi-branch architecture, which benefits establishing multi-view feature representations of EEG signals.

3. Temporal dependencies between spatial-spectral representations of EEG signals can be learned by the bi-directional LSTM structure, which promotes the information interaction between context.

4. The designed skip-change unit provides a feasible way to address the problem of gradient decay in the back-propagation process when training the spatial-temporal networks.

The remainder of this paper is organized as follows. Related work is reviewed in Section II, and the proposed S^3T-Net is introduced in Section III. Experimental results and discussions are presented in Section IV, and finally, the conclusions are drawn in Section V.

II. Related work

Since it is required to fully capture the correlation between the physiological signals of an individual (e.g., EEG) and the emotional state to achieve accurate emotion recognition, the DL models have played important roles in related domain due to the strong feature extraction ability [13]–[19].

In [20], a 1D-CNN has been employed to select high-quality features from specific channels, which achieves an average accuracy of 95.27% and 89.02% on the SEED and DEAP dataset, respectively, and it also reduces the computational burden. LSTM along with a firefly optimization algorithm for feature selection has been applied in [21], which achieves considerable recognition accuracy on SEED dataset. In [22], a temporal-difference minimizing neural network (TDMNN) has been developed, which applies a time-check assessment module and a multi-branching strategy to analyze the temporal stability of emotion activity; and in [23], a spatial-temporal hyper-graph CNN has been proposed, which constructs and updates the hyper-graphs of EEG sequences via a designed self-attentive model so as to capture the higher-order correlations within EEG sequences. In addition, the application of transfer learning and graph-based techniques has been investigated in [24] to improve emotion recognition accuracy, where the proposed 3D MobileNet employs a dynamic graph convolutional network to extract abundant features like the differential entropy and power spectral density, which yields the accuracy of 92.83% on the SEED dataset. Similarly, a novel transfer learning approach that integrates instance filtering and domain adaption has been introduced in [25], which obtains an average accuracy of 91.29% in the within-subjects emotion recognition tasks of the SEED dataset. In [26], a multi-head residual graph convolutional neural network (MRGCN) has been designed, which sufficiently exploit differential entropy (DE) feature of brain activity and achieve an average accuracy of 98.98% and 95.35% on the SEED and DEAP dataset by incorporates complex brain networks and graph convolution networks.

It is noticeable that in recent years, the spatial-temporal network-based DL models have aroused great interests in emotion recognition domain [27]–[29], which generally treat the EEG signals as spatial-temporal data streams, and both spatial distribution and temporal information of EEG signals are sufficiently exploited to promote accurate recognition. Typically, the combination of CNN and LSTM structure is a popular and practical manner to establish such models, see [9], [30]–[32] for some successful instances. In [30], the CNN module transforms the linear EEG sequence into a 2D-like frame sequence, facilitating the exploration of inter-channel correlations among physically neighboring EEG signals, while the LSTM module is used to learned the contextual information, which proved the feasibility of spatial-temporal network. On this basis, in [9], a 4D convolutional recurrent neural network (4D-CRNN) has been proposed, which simultaneously extract spatial, spectral, and temporal features from the raw EEG signals for accurate emotion recognition. In [31], an attention-based convolutional recurrent neural network (ACRNN) has been put forward according to extract more discriminative feature from EEG signals and adaptively select the useful information in channel and time. By applying a similar attention module, a 4D attention-based neural network (4D-ANN) has been designed in [32], which learns the spatial and temporal information via the CNN and Bi-LSTM, respectively, and importance of different brain regions and frequency bands can be adaptively evaluated.

On the basis of spatial-temporal structure, further embedding additional components such as various attention mechanisms has become a novel trend to develop powerful DL approaches for emotion recognition [14], [22], [23], [33]–[37].

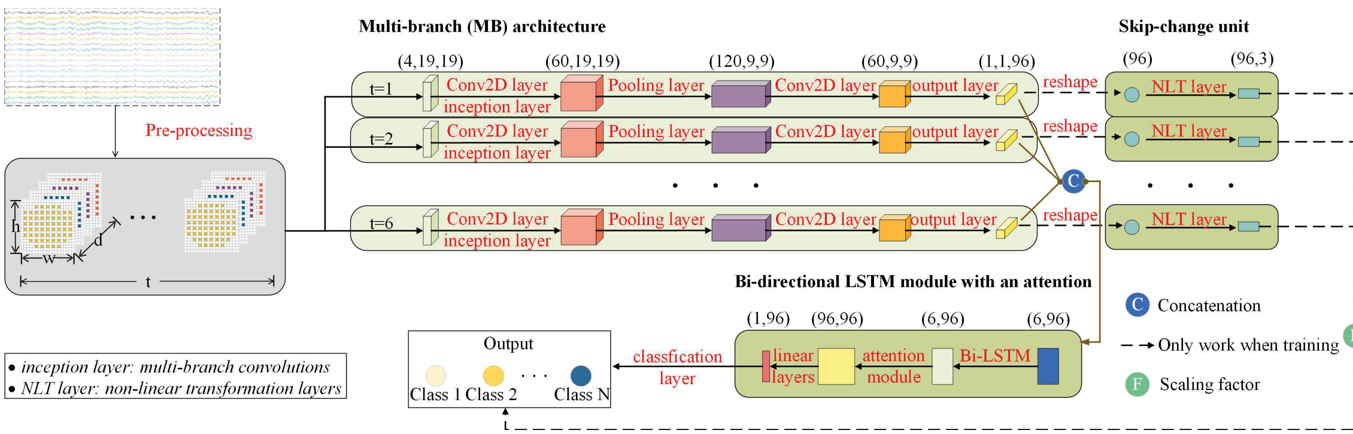


Fig. 1: Overall framework of the proposed S^{3}T-Net.

It should be pointed out that most of the spatial-temporal structure-based models have been dedicated to pursuing deep and comprehensive expansion, which may only bring very limited performance improvement and the significantly increased complexity could even hinder the model convergence. In this regard, another skip-change unit is equipped in the proposed S^{3}T-Net as a complement to the deployed Bi-LSTM structure for potential performance enhancement, and more details are presented in the following Section III.

III. Methods

In this section, the proposed S^{3}T-Net is elaborated with implementation details, whose main contributions lie in the designed multi-branch architecture and the Bi-LSTM module. First of all, the overall framework of the proposed S^{3}T-Net is illustrated in Fig. 1.

A. Overview of S^{3}T-Net

As can be seen from Fig. 1, in order to convert the raw high-dimensional data into a lower-dimensional and more discriminative feature space, the raw EEG data will be pre-processed to obtain a 4D feature input with rich information, next fed into the MB architecture for further extraction of the spatial-spectral representations, and in total six MB architectures are deployed (corresponding to different time steps). To further extract the temporal dependency among above representations, the outputs of six MB architectures are concatenated to enter the Bi-LSTM module. Moreover, above mentioned six outputs will be fed into an additional skip-change unit to assist parameter update of corresponding MB architecture, which only works in the model training stage. Finally, an output layer is placed to realize the final emotion classification. In the following subsections, details of the major components of the proposed S^{3}T-Net are presented.

B. EEG Signal Pre-processing Unit

During data pre-processing stage, spatial (2D), temporal (1D), and spectral (1D) features will be extracted from the raw EEG data. To be specific, each original EEG signal is divided into multiple non-overlapping segments with length of 3 seconds [21], which are assigned the same labels. Then, the butterworth filter is adopted to decompose above segments into four frequency bands, denoted as θ (4-8Hz), α (8-14Hz), β (14-31Hz), and γ (31-51Hz). Next, to effectively represent the chaotic characteristics of neural activity [38]–[40], the differential entropy (DE) features of each frequency band are extracted with a window in size of 0.5s, which can be calculated as:

$$h(Z) = \int_Z f(Z)log(f(z))dz \quad (1)$$

where Z refers to the EEG signals of any electrodes, and $f(z)$ is the corresponding probability density function. According to the relative projected positions of the electrodes, 2D sparse maps can be formed by the obtained DE features, which are then stacked on the frequency bands to obtain the joint spatial-spectral features. Finally, by further concatenating 2D maps acquired from different time windows, 4D features from the original EEG signals are extracted. In addition, an intuitive illustration of the electrode projection maps is presented in Fig. 2, and above entire data pre-processing procedure is illustrated in Fig. 3 for a clear view. As is shown in Fig. 2, the size of the electrode projection is set to 19*19, which considers that the symmetry of traditional convolution kernel and the sparse projection could promote the model to better focus on the information of each electrode.

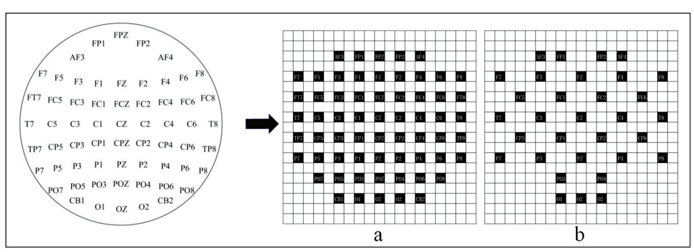


Fig. 2: 2D maps of electrode position projection.

C. Multi-branch (MB) Architecture

To fully exploit the 4D features of EEG signals, a multi-branch (MB) architecture is implemented in the proposed S^{3}T-Net. Compared with the traditional convolutional architecture,

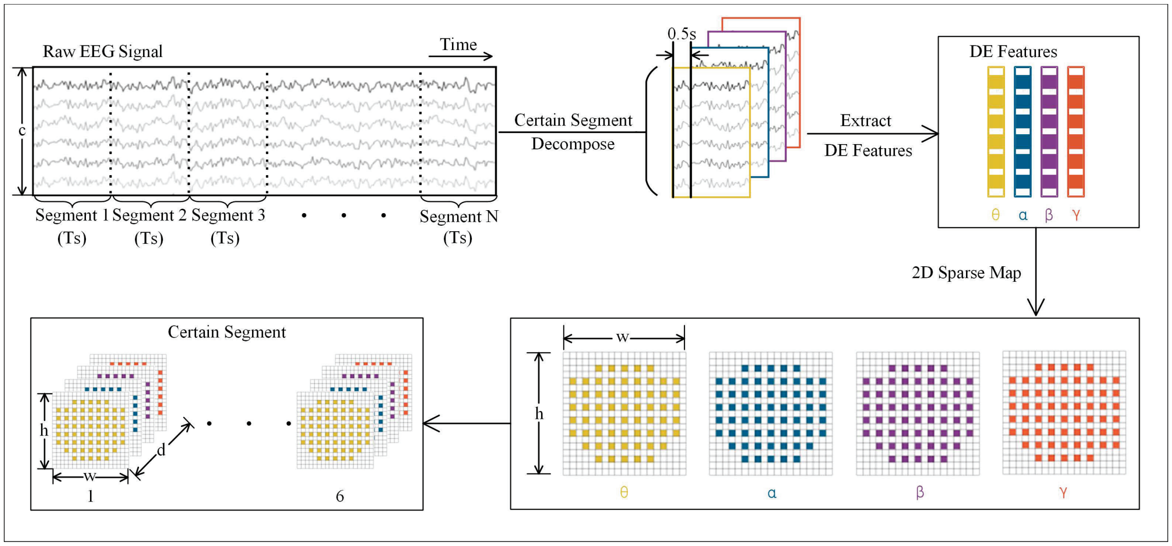


Fig. 3: Illustration of the data pre-processing.

this architecture extracts multi-scale spatial-spectral domain information from multiple network branches to obtain rich and strong representations, thus promoting accurate sentiment classification, while an asymmetric convolutional kernel is employed to minimize the number of network parameters, which is schematically shown in Fig. 4.

As is shown, the MB architecture mainly contains the Conv2D, pooling, inception, and output layers. Firstly, the input data are fed into a Conv2D layer to change the channel numbers, where the batch-normalization and ReLU activation are performed as well. Then, the normalized features will successively enter three inception layers, which are expected to capture multi-scale features by using parallel convolution kernels with different sizes and, by doing so, the model capability to tackle nonlinear representations can be greatly improved. According to Fig. 4, the inception layer includes three convolution branches and one pooling branch, where the application of asymmetric kernels benefits reducing the model parameters. Notice that the branch with more convolution operators corresponds to larger reception field, which pays more attention to the global information, and vice versa. At the end of each branch, a convolutional block attention module (CBAM) [41] is placed to highlight useful features. In addition, two residual connections are added to the three inception layers, which makes the model better adapt to network depths and improves the characterization ability of high-dimensional data.

Next, a pooling layer with max- and pyramid-pooling operations [42] is deployed to further refine the feature maps, where the former can eliminate the low-representative features and the latter can strengthen the correlation of context information. Moreover, a CBAM is also deployed in the pooling layer to enhance the attention on both spatial and spectral regions. Finally, in the output layer, the 1×1 convolution is employed to adjust the channel numbers, and the fully-connected (FC) layer is responsible to provide robust spatial-spectral feature representations.

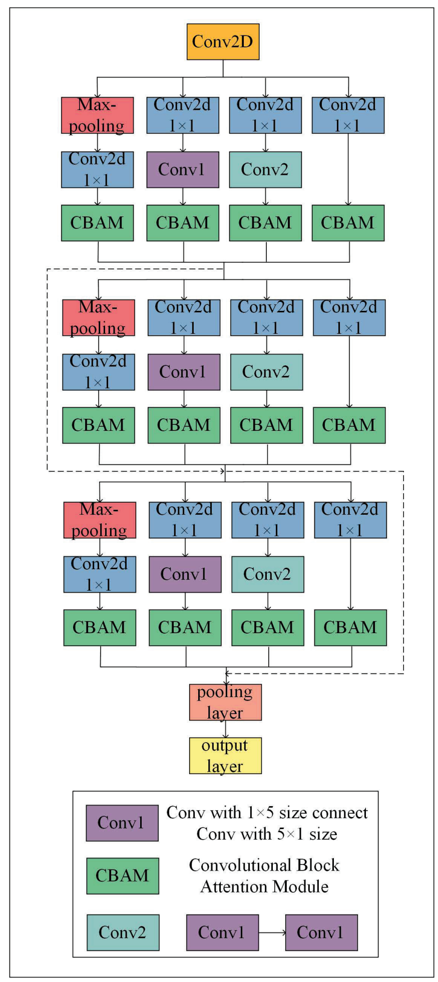


Fig. 4: Diagram of the designed multi-branch architecture.

D. Bi-directional LSTM with attention mechanism

In order to further acquire the temporal dependency among different spatial-spectral feature representations at each time step, the outputs of MB architecture are concatenated and fed into a bi-directional LSTM (Bi-LSTM) [43]–[45] with attention mechanism.

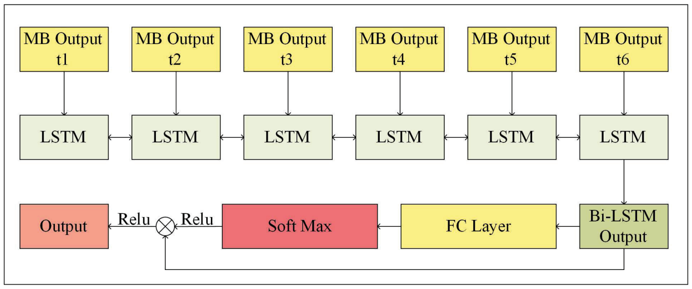


Fig. 5: Diagram of the applied Bi-LSTM module.

As is shown in Fig. 5, the input sequence of Bi-LSTM is processed by two LSTM layers, where the forward one can reflect the natural changing status of the EEG signals in time order, and the backward one can further facilitate the interaction of past and future information. In addition, an attention mechanism is employed at the end of Bi-LSTM to estimate the importance of each time step, which is composed of a fully-connected and a softmax layer. Given W and b as the weights and bias in the FC layer, for the Bi-LSTM output h_i, the corresponding importance coefficient is calculated as:

$$\omega_i = ReLu\Big(\frac{e^{Wh_i+b}}{\sum_j e^{Wh_j+b}}\Big),\ i=1,2,...,6 \tag{2}$$

where $ReLu(\cdot)$ is the activate function, and the final robust features used for emotion classification (in output layer) can be obtained as:

$$Feature_{final} = Relu(\omega_i \cdot h_i) \tag{3}$$

Remark 1: The input of Bi-LSTM is a concatenation of 6 spatial-spectral feature representations, and accordingly, the output of Bi-LSTM covers information at 6 time steps.

E. Skip-change Unit

When training DL models, it is commonly to encounter the problem of vanishing gradients, which generally occurs in the later stage of gradient back-propagation and can severely impede the model convergence. In the proposed S^{3}T-Net, such phenomenon will cause insufficient training of the MB architecture, which declines the model generalization ability.

To handle above issue, a skip-change unit is added to establish the direct associations between outputs of MB architectures and the finally predicted results (see Fig. 1), which only works in model training stage. By setting a non-linear transformation layer, the pseudo prediction tensors of different spectral-spatial feature representations at each time step can be obtained, which are used to calculate the pseudo loss function $\mathcal{L}_p$ according to the ground-truth as:

$$\mathcal{L}_p = \frac{1}{N}\sum_i\sum_{c=1}^{M} y_{ic}log(\tilde{y}_{ic}) \tag{4}$$

where N and M are the number of samples and categories, respectively; y_{ic} is the ground-truth label of sample i, and $\tilde{y}_{ic}$ is the pseudo prediction confidence of sample i belonging to category c. Similarly, another real loss $\mathcal{L}_r$ can be calculated by comparing the ground-truth labels and the finally prediction results of S^{3}T-Net as:

$$\mathcal{L}_r = \frac{1}{N}\sum_i\sum_{c=1}^{M} y_{ic}log(\hat{y}_{ic}) \tag{5}$$

where $\hat{y}_{ic}$ is the prediction confidence of sample i belonging to category c, and other symbols keep the same meanings as in Eq. (4). At last, the eventual loss function for back-propagation is defined as:

$$\mathcal{L} = \mathcal{L}_r + \theta\sum_{i=1}^{6}\mathcal{L}_{pi} \tag{6}$$

where θ is a scaling factor of the sum of six pseudo losses, which enables flexible parameter update and benefits alleviating the gradient vanishing problem.

IV. Results and Discussions

In this section, the proposed S^{3}T-Net is comprehensively evaluated on two representative public EEG-based emotion recognition datasets SEED [46] and DEAP [47], the former comes from Shanghai Jiao Tong University and comprises three categories of EEG signals, including positive, neutral, and negative, where the 62-channel EEG signals of 15 subjects (7 men and 8 women) are recorded by the ESI NeuroScan system while watching different types of film clips. The latter comes from Queen Mary University of London and contains two binary classification tasks, which estimates the levels of valence and arousal of 32 participants by 32 EEG channels, respectively.

Notice that in the officially provided available data, the EEG signals contaminated by the electromyography (EMG) and electrooculography (EOG) are already removed, and other irrelevant noises are filtered out as well. In addition, it should be pointed out that the recognition tasks of both SEED and DEAP datasets are quiet challenging due to the relatively insufficient training data.

A. Model Configurations and Experimental Environments

In Table I, the parameterization conditions of the proposed S^{3}T-Net are provided, including the major structural configurations and hyper-parameter settings in training stage.

Remark 2: Dropout strategy with a probability of 0.2 is applied at the input/output part of the Bi-LSTM module to avoid over-fitting problem caused by the limited training samples.

Moreover, to further verify the competitiveness and superiority of the proposed S^{3}T-Net, five state-of-the-art emotion recognition models are employed for comparison, including ACRNN [31], 4D-CRNN [9], 4D-ANN [32], MRGCN [26] and TDMNN [22] Results of above comparison models are cited from corresponding literature, and for a fair competition, the proposed S^{3}T-Net is tested under the same conditions. On both datasets, five-fold cross-validation is carried out. To be specific, all samples from each participant are randomly divided into five groups, each group will in turn serve as the

TABLE I: Parameterization of the proposed S^{3}T-Net

Variables and Model parameter	Values
Training epochs	200
Batch size	256
Optimizer	Adam
Initial learning rate	0.0005
Weight decay	0.0001
Dropout	0.2
Values of h, w and d in 4D feature	19, 19, 4
Convolution kernel of size (1,5)	padding = (0,2)
Convolution kernel of size (5,1)	padding = (2,0)
# of LSTM cells	64
Scaling factor of pseudo loss	0.3

testing set with the remaining four serving as the training one. Accordingly, the average recognition accuracy and the standard deviation are reported. In addition, all experiments in this study are run on the deep learning framework PyTorch with a single NVIDIA RTX A5000 GPU.

Remark 3: EEG signals in 1-4Hz frequency band are weakly correlated to the emotion, which are removed from the datasets in the experiments.

B. Comparisons with Advanced Models

Firstly, the comparison results with other five advanced models are presented in Table II, and it is found that the proposed S^{3}T-Net yields the best results on both tasks in DEAP dataset and achieves the sub-optimal results on SEED dataset, which sufficiently demonstrate the competitiveness and superiority of S^{3}T-Net. Moreover, as compared to the state-of-the-art TDMNN [22], performance improvements to different extents are obtained by the proposed S^{3}T-Net, especially on SEED dataset, where the average accuracy is improved by 0.43%, which mainly owes to the adoption of multi-scale convolutions enables simultaneous capture of both global and local features of EEG-based sparse maps. As a result, it can be concluded that the proposed S^{3}T-Net is reliable and competent in the EEG-based emotion recognition tasks.

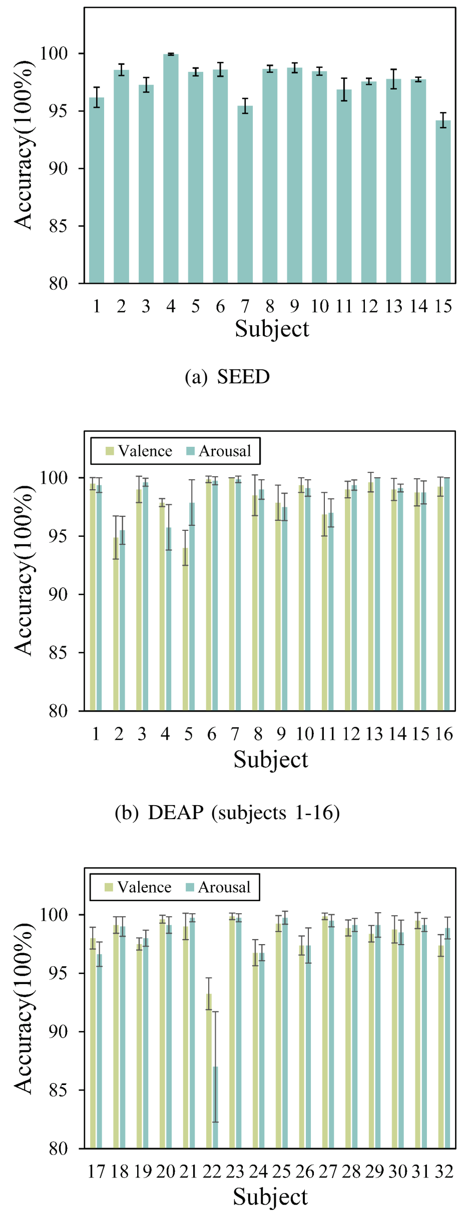
TABLE II: Performance comparison of different models on average accuracy

Model	DEAP-Valence ACC ± STD (%)	DEAP-Arousal ACC ± STD (%)	SEED ACC ± STD (%)
ACRNN [31]	93.72 ± 3.21	93.38 ± 3.73	-
4D-CRNN [9]	94.22 ± 2.61	94.58 ± 3.69	94.74 ± 2.32
4D-ANN [32]	96.90 ± 1.65	97.39 ± 1.75	96.25 ± 1.86
MRGCN [26]	94.97 ± 2.50	95.72 ± 3.80	**98.98 ± 1.50**
TDMNN [22]	98.08 ± 2.13	98.25 ± 2.85	97.20 ± 1.57
S^{3}T-Net(Ours)	**98.31 ± 1.92**	**98.28 ± 2.66**	97.63 ± 1.53

C. Generalization Validation of S^{3}T-Net

In this subsection, emotion recognition in subject-wise is performed to verify the generalization ability of the proposed S^{3}T-Net, and the experimental results are displayed in Fig. 6 and Table III.

Remark 4: In Fig. 6(a), model performance on dataset SEED with 15 subjects is presented, and that on DEAP dataset with 32 subjects in terms of valence and arousal level is displayed in Fig. 6(b) and Fig. 6(c), respectively. For each subject, 5-fold cross-validation is performed.



(a) SEED

(b) DEAP (subjects 1-16)

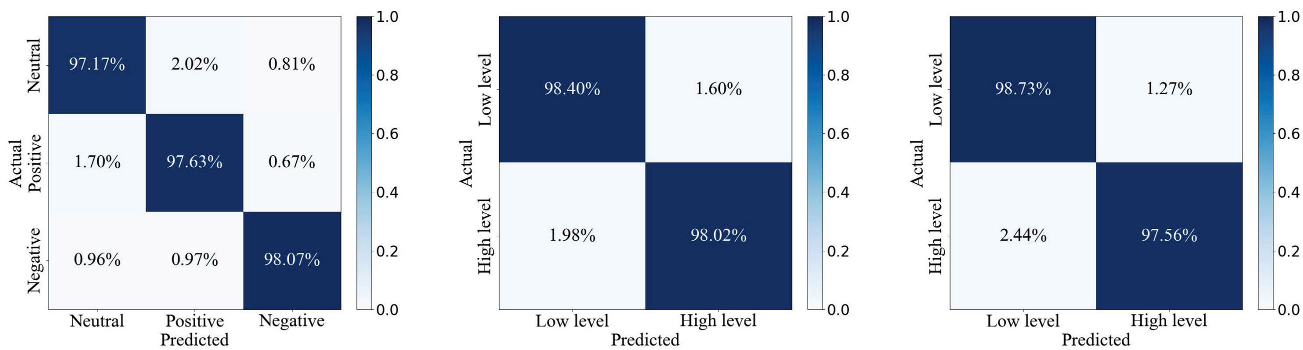
(c) DEAP (subjects 17-32)

Fig. 6: Visualization of subject-wise emotion recognition of the proposed S^{3}T-Net on SEED and DEAP datasets.

As can be observed, the accuracy of each subject and the average one for all subjects are statistically analyzed, where the high accuracy reflects the reliability of the model in dealing with diverse sentiment data. On SEED dataset, only the recognition accuracy of three subjects 1, 7, and 15 is lower than the average level of 97.63%; and on DEAP dataset, the

TABLE III: Subject-wise emotion recognition results on SEED and DEAP datasets

SEED		DEAP-Valence				DEAP-Arousal			
Subject	ACC ± STD (%)	Subject	ACC ± STD (%)	Subject	ACC ± STD (%)	Subject	ACC ± STD (%)	Subject	ACC ± STD (%)
1	96.18 ± 0.88	1	99.50 ± 0.52	17	98.00 ± 0.93	1	99.38 ± 0.63	17	96.63 ± 1.05
2	98.58 ± 0.50	2	94.88 ± 1.84	18	99.13 ± 0.71	2	95.50 ± 1.20	18	99.00 ± 0.84
3	97.28 ± 0.64	3	99.00 ± 1.14	19	97.50 ± 0.52	3	99.63 ± 0.34	19	98.00 ± 0.68
4	99.94 ± 0.08	4	97.88 ± 0.34	20	99.63 ± 0.34	4	95.75 ± 1.95	20	99.13 ± 0.71
5	98.40 ± 0.34	5	94.00 ± 1.51	21	99.00 ± 1.14	5	97.88 ± 1.96	21	99.75 ± 0.34
6	98.61 ± 0.60	6	99.88 ± 0.28	22	93.25 ± 1.35	6	99.75 ± 0.34	22	87.00 ± 4.73
7	95.44 ± 0.66	7	100.00 ± 0.00	23	99.88 ± 0.28	7	99.88 ± 0.28	23	99.75 ± 0.34
8	98.67 ± 0.30	8	98.50 ± 1.75	24	96.75 ± 1.12	8	99.00 ± 0.84	24	96.75 ± 0.68
9	98.76 ± 0.43	9	97.88 ± 1.51	25	99.25 ± 0.68	9	97.50 ± 1.17	25	99.75 ± 0.56
10	98.46 ± 0.34	10	99.38 ± 0.63	26	97.38 ± 0.81	10	99.13 ± 0.71	26	97.38 ± 1.49
11	96.86 ± 0.99	11	96.88 ± 1.88	27	99.88 ± 0.28	11	97.00 ± 1.20	27	99.50 ± 0.52
12	97.57 ± 0.27	12	99.00 ± 0.71	28	98.88 ± 0.68	12	99.38 ± 0.44	28	99.13 ± 0.56
13	97.78 ± 0.85	13	99.63 ± 0.84	29	98.38 ± 0.71	13	100.00 ± 0.00	29	99.13 ± 1.05
14	97.75 ± 0.19	14	99.00 ± 0.95	30	98.75 ± 1.17	14	99.13 ± 0.34	30	98.50 ± 1.05
15	94.20 ± 0.66	15	98.75 ± 1.17	31	99.50 ± 0.68	15	98.75 ± 0.99	31	99.13 ± 0.56
\	\	16	99.25 ± 0.81	32	97.38 ± 0.93	16	100.00 ± 0.00	32	98.88 ± 0.93
Average	97.63 ± 1.53	Average		98.31 ± 1.92		Average		98.28 ± 2.66	



(a) SEED (b) DEAP-Valence (c) DEAP-Arousal

Fig. 7: Confusion matrices obtained by S^{3}T-Net.

performance is satisfactory on both recognition tasks, where the average accuracy reaches 98.31% and 98.28% on the valence and arousal recognition task, respectively. Owing to the existence of individual difference in recording subjective emotion when creating the datasets, it can be deemed that above results are satisfactory and promising, which demonstrates the practicality of the proposed S^{3}T-Net.

TABLE IV: Generalization validation of the proposed S^{3}T-Net

Metrics	DEAP-Valence ACC ± STD (%)	DEAP-Arousal ACC ± STD (%)	SEED ACC ± STD (%)
Precision	0.9802	0.9827	0.9762
Recall	0.9802	0.9760	0.9763
Specificity	0.9840	0.9873	0.9882
F1 score	0.9802	0.9791	0.9762

In addition, the confusion matrices and related evaluation metrics obtained in the three test cases are presented in Fig. 7 and Table IV. In following Eqs. (7)-(10), the calculation of the adopted four metrics are provided [48]:

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

$$Specificity = \frac{TN}{FP+TN} \tag{9}$$

$$F1\ score = 2 \times \frac{Recall \times Precision}{Recall + Precision} \tag{10}$$

where the TP, TN, FP, and FN refer to the true-positive, true-negative, false-positive, and false-negative samples, respectively.

Precision and recall are two important model performance metrics, and the F1 score takes the harmonic average between them. For all of above three metrics, large value indicates that the proposed S^{3}T-Net has an overall excellent performance in the emotion recognition task, and is able to identify most of the real emotion labels effectively. Specificity is a measure of the model ability to recognize non-target emotions (negative category). This characteristic is particularly important for emotion recognition tasks in real-world applications to reduce false positive rates and improve user experience. According to the results, it is found that the proposed S^{3}T-Net can accurately identify different levels of both valance and arousal, implying the sufficient learning of rich EEG patterns generated by each subject. Meanwhile, on SEED dataset, S^{3}T-Net is also able to distinguish three emotion categories with fewer than 1.2% false positive rate, which further verifies that the proposed S^{3}T-Net is a reliable and competent emotion recognition model.

D. Ablation Studies on S^{3}T-Net

Secondly, results of comprehensive ablation studies are displayed in this subsection, which mainly aim to verify the contributions of different core components in S^{3}T-Net to the overall model performance.

1) Ablation on Multi-branch Architecture: In Table V, comparison results of three S^{3}T-Net variants regarding to MB architecture are presented, and for the three model variants, the residual structure in the inception layer (denoted as RES), the pyramid pooling module (denoted as PPM), and the convolutional block attention module (denoted as CBAM) are removed from the MB architecture, respectively. According to the results, it can be found that in average level, the most severe performance degradation is caused by removing the PPM, which decreases the accuracy by 0.24%, 0.27%, and 0.56% in the test case of DEAP-valence, DEAP-arousal, and SEED, respectively. By contrast, the CBAM seems to contribute least to the model performance, which may due to that the small size of the extracted EEG-2D projection maps and the large receptive fields of the designed MB convolutions can handle most target data collaboratively, thereby making the CBAM merely contribute a little in highlighting the important regions in features.

TABLE V: Ablation study on MB architecture, where "-" denotes the removal of subsequent component

Model	Params	DEAP-Valence ACC ± STD (%)	DEAP-Arousal ACC ± STD (%)	SEED ACC ± STD (%)
S^{3}T-Net	791991	**98.31 ± 1.92**	**98.28 ± 2.66**	**97.63 ± 1.53**
- RES	755811	98.05 ± 1.66	98.20 ± 2.23	**97.63 ± 1.45**
- PPM	781011	98.07 ± 1.82	98.01 ± 2.45	97.07 ± 1.64
- CBAM	773726	98.27 ± 1.74	**98.28 ± 2.44**	97.53 ± 1.48

2) Ablation on Bi-LSTM Module: In the proposed S^{3}T-Net, an attention mechanism is incorporated into the Bi-LSTM module, and to verify the effectiveness of such architectural designing, in this part, related ablation study is carried out and the results are displayed in Table VI.

TABLE VI: Ablation study on Bi-LSTM module, where "-" denotes the removal of subsequent component

Model	DEAP-Valence ACC ± STD (%)	DEAP-Arousal ACC ± STD (%)	SEED ACC ± STD (%)
S^{3}T-Net	**98.31 ± 1.92**	**98.28 ± 2.66**	**97.63 ± 1.53**
- Attention	98.17 ± 1.75	**98.28 ± 1.39**	97.57 ± 1.48

As can be seen, when further embedding the attention mechanism into the Bi-LSTM, there is performance improvement on both datasets, which suggests the necessity of evaluating the importance of different time steps to facilitate establishing robust temporal feature representations. In future work, it is also promising to introduce the contrastive learning methods to extract more generalized temporal features with greater differences.

3) Ablation on Skip-change Unit: In this part, advantages of employing the designed skip-change unit (denoted as Skip) are investigated, and according to the comparison results shown in Table VII, the adoption of skip-change unit leads to accuracy improvements of 3.01%, 2.67%, and 1.45% on three test cases, respectively, which implies that the skip-change unit plays an important role in addressing the vanishing gradient so as to benefit training a reliable model. It is also worth noting that the introduction of the skip-change unit hardly increases the parameters of model, which further proves that the method is effective and economical.

TABLE VII: Ablation study on skip-change unit, where "-" denotes the removal of subsequent component

Model	Params	DEAP-Valence ACC ± STD (%)	DEAP-Arousal ACC ± STD (%)	SEED ACC ± STD (%)
S^{3}T-Net	791991	**98.31 ± 1.92**	**98.28 ± 2.66**	**97.63 ± 1.53**
- Skip	791700	95.30 ± 2.88	95.61 ± 3.42	96.18 ± 2.08

Moreover, the gradient flow in back-propagation stage is visualized in Fig. 8 for a clear view. As is shown, without the skip-change unit (see the red arrows), the problem of gradient vanishing may easily occur in Bi-LSTM module due to the high sparsity and a certain similarity of the input data. Consequently, there will be insufficient momentum to well train the MB architecture. When the skip-change unit is introduced (see the blue arrows), a direct association between the output and MB architecture is established through the pseudo loss and, by doing so, the generated more gradient information can benefit sufficient model training to yield better model performance.

4) Ablation Studies on Combinational Components: To further validate the effectiveness of the combination of different core components, the ablation study results of other four model variants are presented in Table VIII, where "All Att" refer to CBAM and the attention mechanism in Bi-LSTM module, and other abbreviations keep the same meanings as previously mentioned. Based on the results, it is found that the original S^{3}T-Net obtains the best results in all test cases, which further demonstrates that it is the collaboration of different core components that endows the proposed S^{3}T-Net with the outstanding performance in EEG-based emotion recognition.

TABLE VIII: Ablation experiment on the combined MB and Bi-LSTM network structures, where "-" denotes the removal of subsequent components

Model	Params	DEAP-Valence ACC ± STD (%)	DEAP-Arousal ACC ± STD (%)	SEED ACC ± STD (%)
S^{3}T-Net	791991	**98.31 ± 1.92**	**98.28 ± 2.66**	**97.63 ± 1.53**
- All Att	764414	98.15 ± 1.84	98.11 ± 2.62	97.58 ± 1.44
- All Att & Skip	770969	95.72 ± 2.87	96.18 ± 3.71	95.96 ± 2.07
- All Att & RES	728234	98.16 ± 1.67	98.20 ± 1.41	97.53 ± 1.53
- All Att & PPM	760694	98.23 ± 1.84	98.00 ± 2.75	97.08 ± 1.72

E. Sensitivity Analysis for Scaling Factor of Pseudo Loss

In the proposed S^{3}T-Net, the total loss function contains six defined pseudo losses $\mathcal{L}_{p\cdot}$, as can be seen from Eq. (6).

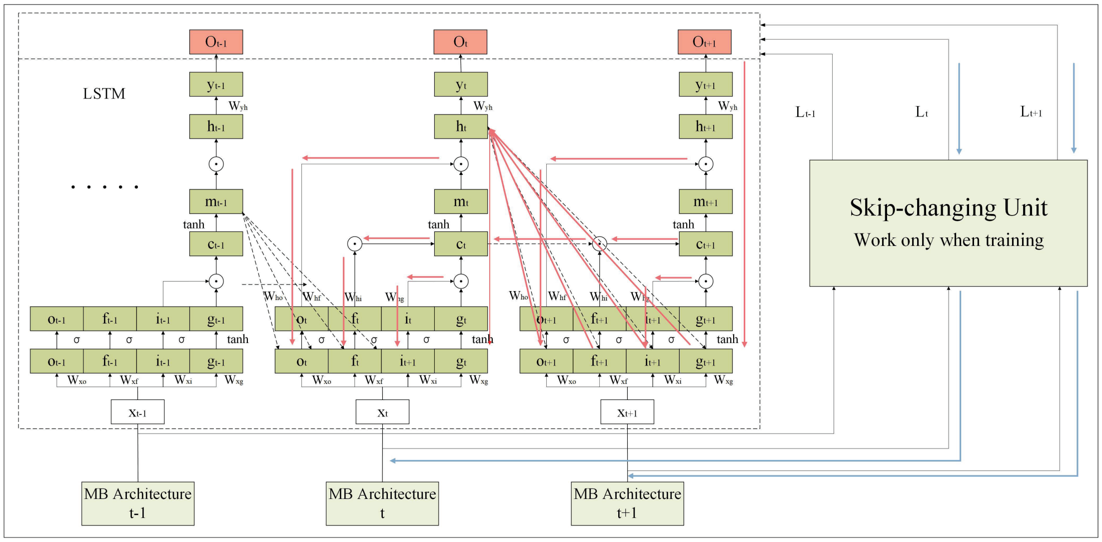


Fig. 8: Gradient flow in back-propagation stage with (in blue) / without (in red) the skip-change unit.

To investigate the influence of pseudo loss on the model generalization ability, sensitivity analysis regarding to the scaling factor θ, which determines the contribution ratio of $\mathcal{L}_p$, is carried out in this subsection, and the results are reported in Table IX.

TABLE IX: Sensitivity analysis for scaling factor of pseudo loss

θ	DEAP-Valence ACC ± STD (%)	DEAP-Arousal ACC ± STD (%)	SEED ACC ± STD (%)
0.3	**98.31 ± 1.92**	98.28 ± 2.66	**97.63 ± 1.53**
0.5	98.24 ± 1.77	98.29 ± 2.44	97.57 ± 1.41
0.7	98.29 ± 1.62	**98.38 ± 2.32**	97.55 ± 1.39

Based on the results, it is recommended to make pseudo loss occupy a relatively small ratio to the entire loss, and when $\theta = 0.3$, two out of three best results are obtained. Additionally, notice that the data in the test case of DEAP-arousal contain more repetitive patterns at different time steps, which may lead to significant gradient decay in Bi-LSTM module during training; and as a result, using a larger scaling factor to strengthen contributions of pseudo loss in this situation can effectively alleviate the gradient vanishing problem.

V. Conclusion

In this paper, a novel S^{3}T-Net for emotion recognition has been developed, which takes thorough explorations on the rich 4D features in the EEG signals. To be specific, a multi-branch architecture has been designed to obtain the multi-scale spatial-spectral representations, and a Bi-LSTM module with an attention mechanism has been placed to further capture the time dependency therein. Additionally, a skip-change unit has been set to alleviate the gradient vanishing problem.

Experiments have shown that the proposed S^{3}T-Net exhibits superior performance in the EEG-based emotion recognition tasks as compared to several other state-of-the-art models, which achieves the average accuracy of 97.63% and 98.30% on SEED and DEAP dataset, respectively. Moreover, extensive ablation studies have sufficiently demonstrated the effectiveness of the architecture designing of S^{3}T-Net. In future work, it is promising to 1) further embed the contrast learning framework to enhance the model performance in cross-subject representation tasks; 2) realize light-weight structure of the proposed S^{3}T-Net to enable the deployments on small edge devices [49], [50]; 3) enhance the model interpretability from the prior knowledge perspective by conducting EEG-related cross-disciplinary research; and 4) explore hybrid architectures and novel optimization algorithms to provide a more comprehensive insight in improving model performance on biomedical classification tasks [51], [52].

References

[1] M. Wei, M. Huang, and J. Ni, "Cross-subject EEG channel selection method for lower limb brain-computer interface," *International Journal of Network Dynamics and Intelligence*, p. 100008, 2023.

[2] N. Yu, R. Yang, and M. Huang, "Deep common spatial pattern based motor imagery classification with improved objective function," *International Journal of Network Dynamics and Intelligence*, pp. 73–84, 2022.

[3] C. Kolli, V. Krishnareddy, T. Reddy, M. Chandol, D. Dasari, and M. Reddy, "Deep learning-based privacy-preserving recommendations in federated learning," *International Journal of General Systems*, pp. 1–27, 2024.

[4] X. Li, M. Li, P. Yan, G. Li, Y. Jiang, H. Luo, and S. Yin, "Deep learning attention mechanism in medical image analysis: Basics and beyonds," *International Journal of Network Dynamics and Intelligence*, vol. 2, no. 1, pp. 93–116, 2023.

[5] G. Ma, Z. Wang, W. Liu, J. Fang, Y. Zhang, H. Ding, and Y. Yuan, "Estimating the state of health for lithium-ion batteries: A particle swarm optimization-assisted deep domain adaptation approach," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 7, pp. 1530–1543, 2023.

[6] N. Zeng, X. Li, P. Wu, H. Li, and X. Luo, "A novel tensor decomposition-based efficient detector for low-altitude aerial objects with knowledge distillation scheme," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 2, pp. 487–501, 2024.

[7] N. Zeng, P. Wu, Y. Zhang, H. Li, J. Mao, and Z. Wang, "DPMSN: A dual-pathway multiscale network for image forgery detection," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 5, pp. 7665–7674, 2024.

[8] J. Tan, Y. Li, Q. Xie, and X. Wang, "Modeling and causality analysis of human sensorimotor control system based on nvar method," *International Journal of Network Dynamics and Intelligence*, p. 100014, 2023.

[9] F. Shen, G. Dai, G. Lin, J. Zhang, W. Kong, and H. Zeng, "EEG-based emotion recognition using 4D convolutional recurrent neural network," *Cognitive Neurodynamics*, vol. 14, pp. 815–828, 2020.

[10] Y. Li, L. Wang, W. Zheng, Y. Zong, L. Qi, Z. Cui, T. Zhang, and T. Song, "A novel bi-hemispheric discrepancy model for EEG emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 13, no. 2, pp. 354–367, 2020.

[11] J. Li, Z. Zhang, and H. He, "Hierarchical convolutional neural networks for EEG-based emotion recognition," *Cognitive Computation*, vol. 10, pp. 368–380, 2018.

[12] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1290–1301, 2020.

[13] 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," *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2909–2917, 2020.

[14] J. Li, X. Wu, Y. Zhang, H. Yang, and X. Wu, "DRS-Net: a spatial–temporal affective computing model based on multichannel EEG data," *Biomedical Signal Processing and Control*, vol. 76, p. 103660, 2022.

[15] J. Li, S. Qiu, Y. Shen, C. Liu, and H. He, "Multisource transfer learning for cross-subject EEG emotion recognition," *IEEE Transactions on Cybernetics*, vol. 50, no. 7, pp. 3281–3293, 2019.

[16] Y. Li, J. Chen, F. Li, B. Fu, H. Wu, Y. Ji, Y. Zhou, Y. Niu, G. Shi, and W. Zheng, "GMSS: Graph-based multi-task self-supervised learning for EEG emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2512–2525, 2022.

[17] Y. Li, W. Cui, Huihuang, Y. Guo, K. Li, and T. Tan, "Epileptic seizure detection in eeg signals using sparse multiscale radial basis function networks and the fisher vector approach," *Knowledge-Based Systems*, vol. 164, pp. 96–106, 2019.

[18] N. Zeng, Z. Wang, Y. Li, M. Du, J. Cao, and X. Liu, "Time series modeling of nano-gold immunochromatographic assay via expectation maximization algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 12, pp. 3418–3424, 2013.

[19] M. V. T. Nguyen Huu Sau and N. T. Phuong, "Exponential stability for discrete-time impulsive positive singular system with time delays," *International Journal of Systems Science*, vol. 55, no. 8, pp. 1510–1527, 2024.

[20] H. Aldawsari, S. Al Ahmadi, and F. Muhammad, "Optimizing 1D-CNN-based emotion recognition process through channel and feature selection from EEG signals," *Diagnostics*, vol. 13, no. 16, p. 2624, 2023.

[21] A. Abgeena and S. Garg, "S-LSTM-ATT: a hybrid deep learning approach with optimized features for emotion recognition in electroencephalogram," *Health Information Science and Systems*, vol. 11, p. 40, 2023.

[22] X. Ju, M. Li, W. Tian, and D. Hu, "EEG-based emotion recognition using a temporal-difference minimizing neural network," *Cognitive Neurodynamics*, pp. 1–12, 2023.

[23] M. Li, M. Qiu, L. Zhu, and W. Kong, "Feature hypergraph representation learning on spatial-temporal correlations for EEG emotion recognition," *Cognitive Neurodynamics*, vol. 17, no. 5, pp. 1271–1281, 2023.

[24] R. Yuvaraj, A. Baranwal, A. Prince, M. Murugappan, and J. Mohammed, "Emotion recognition from spatio-temporal representation of EEG signals via 3D-CNN with ensemble learning techniques," *Brain Sciences*, vol. 13, no. 4, p. 685, 2023.

[25] L. Zhu, F. Yu, A. Huang, N. Ying, and J. Zhang, "Instance-representation transfer method based on joint distribution and deep adaptation for EEG emotion recognition," *Medical & Biological Engineering & Computing*, vol. 62, no. 2, pp. 479–493, 2024.

[26] X. Qiu, S. Wang, R. Wang, Y. Zhang, and L. Huang, "A multi-head residual connection GCN for EEG emotion recognition," *Computers in Biology and Medicine*, vol. 163, p. 107126, 2023.

[27] F. Yu, Y. Huang, C. Du, F. Li, Y. Gong, and G. Chen, "A CNN-LSTM model for road surface recognition of electric balance vehicles," *Systems Science & Control Engineering*, vol. 12, p. 2293912, 2024.

[28] S. Zhang, R. Jing, and X. Shi, "Crop pest recognition based on a modified capsule network," *Systems Science & Control Engineering*, vol. 10, no. 1, pp. 552–561, 2022.

[29] H. Qi and X. Meng, "Global dynamics of a stochastic reaction–diffusion predator–prey system with space-time white noise," *International Journal of Systems Science*, pp. 1–27, 2023.

[30] Y. Yang, Q. Wu, M. Qiu, Y. Wang, and X. Chen, "Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network," *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2018.

[31] 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*, vol. 14, no. 1, pp. 382–393, 2020.

[32] G. Xiao, M. Shi, M. Ye, B. Xu, Z. Chen, and Q. Ren, "4D attention-based neural network for EEG emotion recognition," *Cognitive Neurodynamics*, pp. 1–14, 2022.

[33] J. Li, W. Pan, H. Huang, J. Pan, and F. Wang, "STGATE: spatial-temporal graph attention network with a transformer encoder for EEG-based emotion recognition," *Frontiers in Human Neuroscience*, vol. 17, p. 1169949, 2023.

[34] N. Pusarla, A. Singh, and S. Tripathi, "Learning densenet features from EEG based spectrograms for subject independent emotion recognition," *Biomedical Signal Processing and Control*, vol. 74, p. 103485, 2022.

[35] X. Shi, Q. She, F. Fang, M. Meng, T. Tan, and Y. Zhang, "Enhancing cross-subject EEG emotion recognition through multi-source manifold metric transfer learning," *Computers in Biology and Medicine*, vol. 174, p. 108445, 2024.

[36] M. Jafari, A. Shoeibi, M. Khodatars, S. Bagherzadeh, A. Shalbaf, D. López-García, J. M. Górriz, and U. R. Acharya, "Emotion recognition in EEG signals using deep learning methods: A review," *Computers in biology and medicine*, vol. 165, p. 107450, 2023.

[37] Z. He, Y. Zhong, and J. Pan, "An adversarial discriminative temporal convolutional network for EEG-based cross-domain emotion recognition," *Computers in Biology and Medicine*, vol. 141, p. 105048, 2022.

[38] C. Stam, "Nonlinear dynamical analysis of EEG and MEG: review of an emerging field," *Clinical Neurophysiology*, vol. 116, no. 10, pp. 2266–2301, 2005.

[39] W. Zheng, J. Zhu, and B. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 417–429, 2017.

[40] R. Duan, J. Zhu, and B. Lu, "Differential entropy feature for EEG-based emotion classification," *International IEEE/EMBS Conference on Neural Engineering (NER)*, no. 6, pp. 81–84, 2013.

[41] S. Woo, J. Park, J. Lee, and I. Kweon, "CBAM: convolutional block attention module," *Proceedings of the European Conference on Computer Vision (ECCV)*, vol. 11211, pp. 3–19, 2018.

[42] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230–6239, 2017.

[43] L. Ni and V. Khim Sen Liew, "Carbon emission price forecasting in China using a novel secondary decomposition hybrid model of CEEMD-SE-VMD-LSTM," *Systems Science & Control Engineering*, vol. 12, p. 2291409, 2024.

[44] Y. Sun and K. Liao, "A hybrid model for metro passengers flow prediction," *Systems Science & Control Engineering*, vol. 11, p. 2191632, 2023.

[45] A. Dong, A. Starr, and Y. Zhao, "Neural network-based parametric system identification: a review," *International Journal of Systems Science*, vol. 54, no. 13, pp. 2676–2688, 2023.

[46] W. Zheng and B. 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.

[47] S. Koelstra, C. Muhl, M. Soleymani, J. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: a database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2011.

[48] H. Li, P. Wu, Z. Wang, J. Mao, F. E. Alsaadi, and N. Zeng, "A generalized framework of feature learning enhanced convolutional neural network for pathology-image-oriented cancer diagnosis," *Computers in Biology and Medicine*, vol. 151, p. 106265, 2022.

[49] L. Hu, Z. Wang, H. Li, P. Wu, J. Mao, and N. Zeng, "l-DARTS: Lightweight differentiable architecture search with robustness enhancement strategy," *Knowledge-Based Systems*, vol. 288, p. 111466, 2024.

[50] J. Xue and B. Shen, "A survey on sparrow search algorithms and their applications," *International Journal of Systems Science*, vol. 55, no. 4, pp. 814–832, 2024.

[51] S. Raziani, S. Ahmadian, S. M. J. Jalali, and A. Chalechale, "An efficient hybrid model based on modified whale optimization algorithm and multilayer perceptron neural network for medical classification problems," *Journal of Bionic Engineering*, vol. 19, pp. 1504 – 1521, 2022.

[52] Y. Özlük and E. Akman Aydin, "Fuzzy logic control of a head-movement based semi-autonomous human–machine interface," *Journal of Bionic Engineering*, vol. 20, pp. 645 – 655, 2023.