A Cross-Scale Transformer and Triple-View Attention Based Domain-Rectified Transfer Learning for EEG Classification in RSVP Tasks

Jie Luo, Weigang Cui, Song Xu, Lina Wang, Huiling Chen, *Associate Member, IEEE*, and Yang Li, *Senior Member, IEEE*

***Abstract*— Rapid serial visual presentation (RSVP)-based brain-computer interface (BCI) is a promising target detection technique by using electroencephalogram (EEG) signals. However, existing deep learning approaches seldom considered dependencies of multi-scale temporal features and discriminative multi-view spectral features simultaneously, which limits the representation learning ability of the model and undermine the EEG classification performance. In addition, recent transfer learning-based methods generally failed to obtain transferable cross-subject invariant representations and commonly ignore the individual-specific information, leading to the poor cross-subject transfer performance. In response to these limitations, we propose a cross-scale Transformer and triple-view attention based domain-rectified transfer learning (CST-TVA-DRTL) for the RSVP classification. Specially, we first develop a cross-scale Transformer (CST) to extract multi-scale temporal features and exploit the dependencies of different scales features. Then, a triple-view attention (TVA) is designed to capture spectral features from triple views of multi-channel time-frequency images. Finally, a domain-rectified transfer learning (DRTL) framework is proposed to simultaneously obtain transferable domain-invariant representations and untransferable domain-specific representations, then utilize domain-specific information to rectify domain-invariant representations to adapt to target data. Experimental results on two public RSVP datasets suggests that our CST-TVA-DRTL outperforms the state-of-the-art methods in the RSVP classification task. The source code of our model is publicly available in https://github.com/ljbuaa/CST_TVA_DRTL.**

***Index Terms*— Brain–computer interface, EEG, RSVP, transformer, transfer learning.**

Manuscript received 3 October 2023; revised 25 December 2023 and 17 January 2024; accepted 23 January 2024. Date of publication 29 January 2024; date of current version 5 February 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFC2416600; in part by the National Natural Science Foundation of China under Grant 62325301, Grant 62201023, and Grant U23A20335; in part by the Beijing Natural Science Foundation under Grant Z220017; and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LZ23F030001. *(Corresponding author: Yang Li.)*

Jie Luo is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China (e-mail: luojiecn@buaa.edu.cn).

Weigang Cui is with the School of Engineering Medicine, Beihang University, Beijing 100191, China (e-mail: cwg1994@buaa.edu.cn).

Song Xu and Lina Wang are with the National Key Laboratory of Science and Technology on Aerospace Intelligence Control, Beijing Aerospace Automatic Control Institute, Beijing 100070, China (e-mail: xusong618@163.com; violina@126.com).

Huiling Chen is with the College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou, Zhejiang 325035, China (e-mail: chenhuiling.jlu@gmail.com).

Yang Li is with the Department of Automation Science and Electrical Engineering and the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China (e-mail: liyang@buaa.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2024.3359191

I. Introduction

THE electroencephalogram (EEG)-based brain-computer interface (BCI) is a promising interactive technology that empowers humans to interact with computer directly through brain signals [1], [2]. Rapid serial visual presentation (RSVP) is a well-established BCI paradigms that has been widely used in speller [3] and image retrieval [4]. Nevertheless, the noisy single-trial EEG data and the class imbalance problem hinder EEG classification methods from achieving better performance on RSVP task [5], [6], [7].

Many researchers have proposed various temporal feature extraction methods to improve EEG classification performance in the RSVP task [8], [9]. For example, hierarchical discriminant component analysis (HDCA) introduced a group of spatial-temporal filters to extract discrimination temporal information from single-trial EEG signals [10]. With the successful application of deep learning technology in various fields, several deep learning-based frameworks have been proposed for EEG classification. The deep ConvNet (DCN) uses temporal convolution and spatial convolution to extract spatio-temporal features from EEG data, and realizes end-to-end EEG classification [11]. Similarly, the EEGNet captures EEG spatio-temporal features through the depth-wise and separable convolution layers with fewer parameters [5]. In order to extract multi-scale temporal features, Santamaría-Vázquez et al. developed an EEGInception method that incorporates the inception module into the convolutional network [12]. The inception module uses three convolution with different sizes kernels to extract multiple scales temporal features. This approach effectively improves the classification accuracy of the RSVP task. However, this method ignores the dependence of different scale temporal features, which may cause information redundancy of multi-scale temporal features, thus affecting the further improvement of model performance.

Additionally, recent studies have demonstrated that spectrogram of EEG signals can provide discriminative features for EEG classification. For instance, Kang et al. converted EEG data into spectrogram images by short-time Fourier transform, and then used an ensemble convolutional neural networks (CNN) to captures the spectral features from the time-frequency view [13]. To better discover important spectral features, Zhang et al. adopted the channel attention mechanism to enhance the spectral features on important channels after converting non-stationary EEG signals into multi-channel spectrogram images through continuous wavelet transform (CWT) [14]. These studies effectively improved the performance of EEG classification models by exploiting discriminative time-frequency features. However, the aforementioned works only capture spectral features from a single time-frequency view, ignoring the correlation of spectral features between different channels in multi-channel EEG spectrogram, which undermine the discriminability of spectral features.

In RSVP classification task, the risk of overfitting caused by insufficient training data and class imbalance limits the flexible application of RSVP-based BCIs [15], [16]. Recently, some studies have introduced transfer learning strategies to enable rapid application of RSVP-based BCIs. For example, Wei et al. proposed a multi-source transfer learning framework based on domain adversarial training, which reduces the amount of data required to train models on new subjects by learning the common features of other subjects' EEG data [17]. Similarly, He et al. developed a transfer learning method based on Euclidean space data alignment, which improves the learning performance for new subjects by aligning EEG trials from different subjects in Euclidean space [18]. The above studies show that cross-subject transfer performance of the deep learning methods can be effectively improved by learning the common EEG features among multiple subjects. However, due to substantial inter-individual variability in EEG signals [19], existing transfer learning methods tend to overlook individual-specific features and consequently compromise the transfer performance by incorporating untransferable individual information into common features.

In response to above issues, we propose a cross-scale Transformer and triple-view attention based domain-rectified transfer learning framework (CST-TVA-DRTL) for RSVP classification. First, a cross-scale Transformer (CST) temporal feature extractor is employed to extract multi-scale temporal features from EEG signals, which can characterize the dependencies of temporal features across scales and reduce redundant information. Second, a triple-view attention (TVA) spectral feature extractor is proposed, which can capture multi-channel spectral features from spectral-temporal view, spatio-temporal view and spatio-spectral view. Finally, we design a domain-rectified transfer learning (DRTL) strategy that can simultaneously encode transferable domain-invariant representations and untransferable domain-specific representations of EEG features, and then uses domain-specific representations to rectify the domain-invariant representations to adapt to the target domain. Two public datasets were used to evaluate the proposed CST-TVA-DRTL method, and the experimental results show that the proposed method is superior to the state-of-the-art methods in the RSVP classification task, which proves the effectiveness of our proposed CST-TVA-DRTL method.

The contributions of this study are fourfold:

(1) We propose a novel CST-TVA-DRTL method to detect ERP from single-trial EEG signals for RSVP-based BCIs, which effectively improves the RSVP classification performance by extracting multi-scale temporal features and multi-view spectral features from EEG data, and adopting domain-rectified transfer learning approach.

(2) A CST is designed to extract multi-scale temporal features from EEG data and reduce redundant information, which can significantly enhance the representation ability of extracted temporal features.

(3) We develop a TVA to capture spectral features of multi-channel EEG spectrogram from spectral-temporal, spatio-temporal and spatio-spectral views, which can provide more discriminative spectral representations.

(4) We propose a DRTL framework that can simultaneously obtain transferable domain-invariant representations and untransferable domain-specific representations, and use domain-specific representations to rectify the domain-invariant representations, which can enhance cross-subject transfer performance.

II. Methodology

A. Overview

The CST-TVA-DRTL framework mainly consists of two stages: pre-training and fine-tuning, as shown in Fig.1. In the pre-training stage, except the target subject data, other subject data are used for training whole framework. In the fine-tuning stage, the target domain data is used to fine-tune the parameters of the domain-specific feature encoder to rectify the domain-invariant representation.

For the framework structure, a spatial filtering algorithm xDAWN [20] is first used to filter raw EEG signals to enhance the P300 evoked potentials and the CWT is adopted to convert the EEG data into multi-channel spectrogram images. Then, the cross-scale Transformer temporal feature extractor is constructed to capture multi-scale temporal features of EEG signal and characterize the temporal dependencies of different scale temporal features. Meanwhile, the triple-view attention spectral feature extractor is adopted to extract spectral features from multiple views of multi-channel EEG spectrogram. Next, domain-specific and domain-invariant feature encoders are used to obtain domain-specific and domain-invariant representations, respectively. The domain rectification block uses domain-specific representations to rectify domain-invariant representations, and the rectified representations are used for the final RSVP classification.

B. CST Temporal Feature Extractor

By averaging multiple trials of EEG signals, significant differences can be observed between the waveform of the P300 signal evoked by target images and the EEG signals evoked by non-target images. In order to capture the time-scale difference

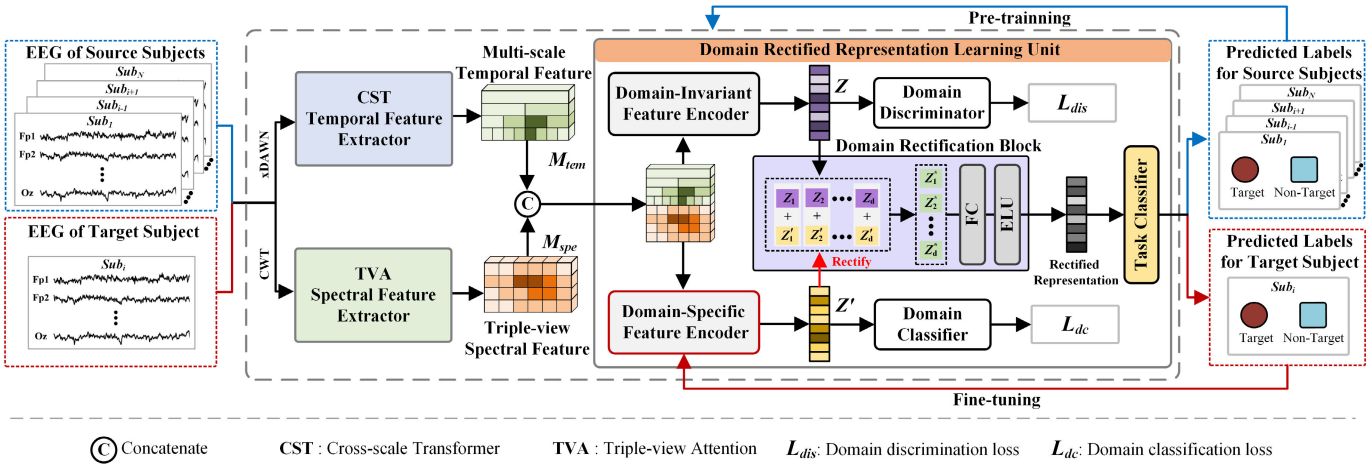

Fig. 1. The flowchart of our proposed CST-TVA-DRTL framework for RSVP classification.

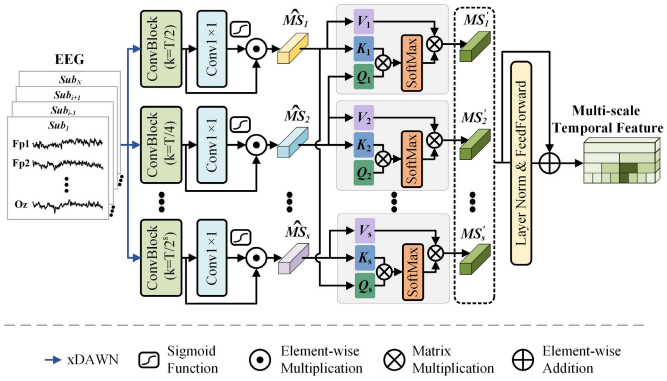

Fig. 2. Diagram of the CST temporal feature extractor.

of target and non-target evoked EEG, we have devised a CST temporal feature extractor, which leverages multi-scale convolution to extract temporal features across various scales, and subsequently utilizes a Transformer to model the temporal dependence of the features. Fig.2 demonstrates the flowchart of the CST. Specifically, multiple convolution blocks with different kernel sizes are first used to extracts different scale features of EEG signals as follows:

$$MS_s = Conv_k(X), \quad s = 1, 2, \ldots, S \tag{1}$$

where $X \in \mathbb{R}^{C\times T}$ is the EEG signal obtained through the xDAWN spatial filter, C represent the number of channels, T is the time length, and S is the number of scales. The $Conv_k(\cdot)$ denotes convolution block with kernel size $k = \frac{T}{2^s}$, each convolution block consist of a convolution layer, a batch normalization layer and an ELU activation layer. The MS_s is the s-th scale temporal features of EEG signal. In order to enhance key features at different scales by adaptive weighting, a convolution layer with 1×1 kernel size followed by a sigmoid function is adopted to calculate adaptive weights. Then, the weighted multi-scale features $\widehat{MS}_s$ is calculated as follows:

$$AW_s = \sigma\left(Conv_{1\times1}(MS_s)\right), \quad s = 1, 2, \ldots, S \tag{2}$$

$$\widehat{MS}_s = MS_s \odot AW_s \tag{3}$$

where $\sigma(\cdot)$ denotes sigmoid activation functions. The $Conv_{1\times1}(\cdot)$ is a 1×1 convolution layer. AW_s is the adaptive weights.

Owing to the interdependencies among multi-scale features, ordinary self-attention mechanisms may inadequately capture cross-scale dependencies, potentially leading to excessive redundancy in the fused multi-scale features [21]. To address this limitation, we propose a cross-scale multi-head self-attention block that effectively models the cross-scale dependence between multi-scale temporal features, thereby reducing redundancy in the fused multi-scale features. In the cross-scale multi-head self-attention block, two linear transformation layers are first used to transform the weighted multi-scale temporal features $\widehat{MS}_s \in \mathbb{R}^{d\times T}$ into the matrix $V_s \in \mathbb{R}^{d\times T}$ and $K_s \in \mathbb{R}^{d\times T}$, respectively. The query matrix $Q_s \in \mathbb{R}^{d\times T}$ is obtained by linear transformation of smaller-scale features $\widehat{MS}_{s+1}$. The transformations are defined as:

$$V_s = \widehat{MS}_s W^{V_s} \tag{4}$$

$$K_s = \widehat{MS}_s W^{K_s} \tag{5}$$

$$Q_s = \begin{cases} \widehat{MS}_{s+1} W^{Q_s}, & s < S \\ \widehat{MS}_1 W^{Q_1}, & s = S \end{cases} \tag{6}$$

where W^{V_s}, W^{K_s}, and W^{Q_s} are the learnable matrices of linear layers. Then, the cross-scale attention CA_s can be obtained by:

$$CA_s = Softmax\left(\frac{Q_s K_s^*}{\sqrt{d}}\right) V_s \tag{7}$$

where K_s^* is the transpose of K_s, $\sqrt{d}$ is the scaling factor and $Softmax(\cdot)$ is the softmax function.

The multi-scale features obtained through cross-scale multi-head attention [22] are fused through the concatenation function and the feed-forward network after linear transformation. The formulas for CST to obtain multi-scale temporal features are as follows:

$$M'_{\text{tem}} = Concat\left(MS'_1, \cdots, MS'_s\right), \quad s = 1, 2, \ldots, S \tag{8}$$

$$M_{\text{tem}} = \mathcal{F}^2_{fc}\left(GELU\left(\mathcal{F}^1_{fc}\left(LN\left(M'_{\text{tem}}\right)\right)\right)\right) + M'_{\text{tem}} \tag{9}$$

where $Concat(\cdot)$ is the concatenation function, $MS'_s = \mathcal{P}_s(CA_s)$, $\mathcal{P}_s$ are linear transform layers. $LN(\cdot)$ means layer

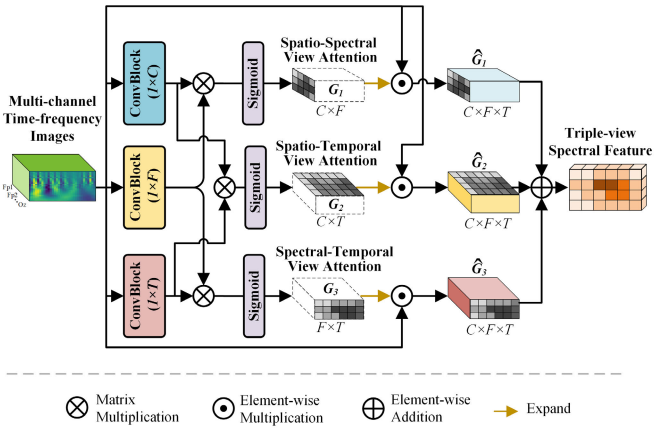

Fig. 3. Diagram of the TVA spectral feature extractor.

normalization, $GELU(\cdot)$ denotes Gaussian error linear units, $\mathcal{F}_{fc}^{1}$ and $\mathcal{F}_{fc}^{2}$ denote the FC layers.

C. TVA Spectral Feature Extractor

Many studies have shown that the spectral characteristics of EEG are helpful in classifying EEG signals [23], [24], [25]. However, converting EEG signals into multi-channel spectrogram images through CWT may lead to redundant information in spectrogram images. In order to extract features from high-dimensional time-frequency images more effectively, we construct a TVA spectral feature extractor to explore the spectral features.

Fig.3 presents the diagram of TVA. Concretely, spatial convolution block, spectral convolution block and temporal convolution block are first used to obtain triple-view features of multi-channel time-frequency images $U \in \mathbb{R}^{C\times F\times T}$, where F is the frequency dimension. The spatial convolution block consists of a convolution layer with a convolution kernel of $1\times C$ and a reshape layer. Similarly, the convolution kernels in the spectral convolution block and temporal convolution block are $1\times F$ and $1\times T$, respectively. The triple-view attention can be expressed by the following formula:

$$G_1 = \sigma(\Phi_{\text{spa}}(U)\Phi_{\text{spe}}(U)) \tag{10}$$

$$G_2 = \sigma(\Phi_{\text{spa}}(U)\Phi_{\text{tem}}(U)) \tag{11}$$

$$G_3 = \sigma(\Phi_{\text{spe}}(U)\Phi_{\text{tem}}(U)) \tag{12}$$

where $\Phi_{\text{spa}}(\cdot)$, $\Phi_{\text{spe}}(\cdot)$ and $\Phi_{\text{tem}}(\cdot)$ represent spatial convolution block, spectral convolution block and temporal convolution block respectively. With the triple-view attention, the output spectral features of the TVA unit are calculated by the following formula:

$$M_{spe} = Conv\left(\sum_{v=1}^{3} G_v \odot U\right) \tag{13}$$

where $\odot$ means element-wise multiplication.

D. Domain-Rectified Transfer Learning

Transfer learning is an effective strategy to improve the performance of deep learning models on EEG datasets with few labeled samples [26], [27], [28]. Existing transfer learning methods usually aim to learn domain-invariant features of EEG data from different subjects [29]. However, the learning of domain-invariant representations is difficult due to the large individual differences of EEG signals, which affects the performance of transfer learning methods. To address this issue, we design a domain-rectified transfer learning (DRTL) framework, which adapts to the target domain by rectifying the domain-invariant representation through target domain-specific representation.

The diagram of domain-rectified representation learning is shown in Fig. 1. In particular, domain-specific feature encoder $\boldsymbol{E}_{\varphi}$ and domain-invariant feature encoder $\boldsymbol{E}_{\phi}$ are used to obtain domain-specific representations $z' \in \mathbb{R}^{L}$ and common domain-invariant representations $z \in \mathbb{R}^{L}$ of different subjects' EEG data, respectively. The L denotes the dimension of representation vectors. Domain-specific feature encoder and domain-invariant feature encoder have the same network structure, both consisting of a convolutional block and a fully-connected layer. To make domain-invariant representations adaptive to different domains, domain-specific representations are used to rectify domain-invariant representations. The calculation process of domain-rectified is as follows:

$$z' = \boldsymbol{E}_{\varphi}\left(Concat\left(M_{tmp}, M_{spe}\right)\right) \tag{14}$$

$$z = \boldsymbol{E}_{\phi}\left(Concat\left(M_{tmp}, M_{spe}\right)\right) \tag{15}$$

$$z^{*} = ELU\left(\mathrm{F_z}\left(z' + z\right)\right) \tag{16}$$

where z^{*} denotes the rectified representations, $ELU(\cdot)$ is exponential linear unit activation function and $\mathrm{F_z}$ denotes the FC layer. Then, the rectified representation is input into the task classifier $\boldsymbol{G}_{\psi}$ for RSVP classification. $\widehat{y} = \boldsymbol{G}_{\psi}(z^{*})$ is the predicted class label and the classification loss is defined as follows:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{H}\sum_{h=0}^{H}\left(y_h log\left(\widehat{y}_h\right)\right) \tag{17}$$

where y_h and $\widehat{y}_h$ are the actual and predicted label for the h-th sample respectively, H is the total number of samples. In the pre-training stage, in order to make the domain-specific feature encoder learn more domain-specific information, a domain classifier $\boldsymbol{D}_{dc}$ is used to classify the domain feature representation. $\widehat{d} = \boldsymbol{D}_{dc}(z')$ is the predicted domain label. The domain-specific feature encoder is constrained by minimizing the domain classification loss. The domain classification loss is defined as follows:

$$\mathcal{L}_{\text{dc}} = -\frac{1}{H}\sum_{j=0}^{J}\sum_{h=0}^{H}\left(d_n^j log(\widehat{d}_n^j)\right) \tag{18}$$

where d_h^j and $\widehat{d}_h^j$ are the actual and predicted domain label for the h-th sample respectively, and J is the total number of domains. Meanwhile, similar to DANN, a domain discriminator is used to identify domain labels for domain-invariant representation learning. To constrain the distance of features between different domains, we confuse the domain discriminator by maximizing the domain discrimination loss [30]. $\bar{d} = \boldsymbol{D}_{dis}(z)$ is the predicted domain label.The domain discrimination loss is defined as follows:

$$\mathcal{L}_{\text{dis}} = \frac{1}{H}\sum_{j=0}^{J}\sum_{h=0}^{H}\left(d_n^j log(\bar{d}_n^j)\right) \tag{19}$$

Finally, a joint loss function L_{pt} based on task classification loss, domain classification loss and domain discrimination loss is used to constrain the parameter optimization of the model in the pre-training stage.

$$\mathcal{L}_{pt} = \mathcal{L}_{cls} + \mathcal{L}_{dc} + \mathcal{L}_{dis} \tag{20}$$

In the fine-tuning stage, only the model parameters of the task classifier of the domain-specific feature encoder need to be fine-tuned under the constraints of the task classification loss function L_{cls}. The optimization procedure of the CST-TVA-DRTL framework is summarized in **Algorithm 1**.

Algorithm 1 The Parameters Update in the Proposed CST-TVA-DRTL Framework

Input: Raw EEG signals x, the corresponding class labels y, the maximum epoch K, T and the CST-TVA-DRTL *Net* $(\theta, \varphi, \phi, \psi)$.
Output: The RSVP classification result O.

Initializing temporal and spectral feature extractor parameters θ and parameters of $\boldsymbol{E}_{\varphi}$, $\boldsymbol{E}_{\phi}$ and $\boldsymbol{G}_{\psi}$.
Initializing source domain data x_{src}, y_{src} and target domain data x_{tg}, y_{tg}
Initializing $K = 1 \times 10^3$, $T = 2 \times 10^2$, $\kappa = 0$, $\tau = 0$.
while $\kappa < K$
    Generate conditional probability: $\widehat{y} = Net(\theta^{\kappa}, \varphi^{\kappa}, \phi^{\kappa}, \psi^{\kappa}, x_{src}^{\kappa})$;
    Calculate loss: $\mathcal{L}_{pt} = \mathcal{L}_{cls} + \mathcal{L}_{dc} + \mathcal{L}_{dis}$
    Update the parameters: $\theta^{\kappa+1}, \varphi^{\kappa+1}, \phi^{\kappa+1}, \psi^{\kappa+1} \leftarrow \hat{u} \arg \min_{\theta^{\kappa}, \varphi^{\kappa}, \phi^{\kappa}, \psi^{\kappa}} \mathcal{L}_{pt}$;
    $\kappa \leftarrow \kappa + 1$;
end while
while $\tau < T$
    Generate conditional probability: $\widehat{y} = Net(\theta^{\kappa}, \varphi^{\kappa}, \phi^{\tau}, \psi^{\tau}, x_{tg}^{\tau})$;
    Calculate loss: $\mathcal{L}_{cls}$
    Update the parameters: $\phi^{\tau+1}, \psi^{\tau+1} \leftarrow \hat{u} \arg \min_{\phi^{\tau}, \psi^{\tau}} \mathcal{L}_{cls}$;
    $\tau \leftarrow \tau + 1$;
end while
Get the classification result $O = Net(\theta^{\kappa}, \varphi^{\kappa}, \phi^{\tau}, \psi^{\tau}, x)$

III. Experiments and Results

A. RSVP Datasets

Two publicly available RSVP target detection EEG datasets are used to evaluate the performance of the proposed CST-TVA-DRTL: Tsinghua RSVP dataset [31] and PhysioNet RSVP dataset [32]. The RSVP paradigms are shown in Fig.4.

1) Tsinghua RSVP dataset: This dataset comprises RSVP EEG data collected from 64 subjects, with a gender distribution of 32 females. Each subject observed 160 sequences of stimulus images. Within each sequence, a total of 100 distinct images were presented for 0.1s each. In these images, street view images with people are targets, while street view images with no people are non-targets. The number of non-target images is approximately 64 times the number of target images. EEG recordings were obtained using the Synamps2 system equipped with 64 channels, sampling at a rate of 1,000 Hz. The electrodes were placed according to the international 10–20 system.

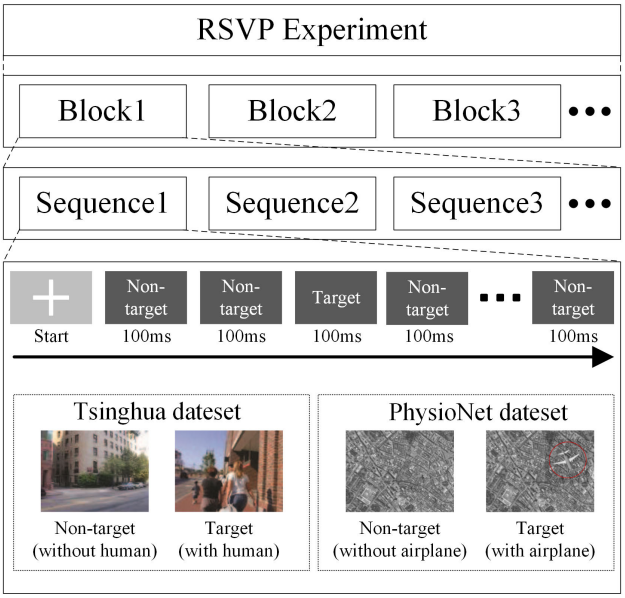

Fig. 4. The RSVP paradigm for target detection.

2) PhysioNet RSVP dataset: This dataset encompasses RSVP EEG data acquired from 10 subjects (4 females) while viewing 8 sequences of images at presentation rates of 10 Hz. The stimulus images can be divided into target images with airplane and non-target images without airplane. In these stimulus images, the number of target images is only one tenth of the total number of images. The EEG recordings were captured using BioSemi ActiveTwo system equipped with 64 channels at a sample rate of 2048 Hz, but only 8 channels: P7, P8, PO7, PO3, PO4, PO8, O1, O2 are available. Electrode placement followed the international 10–20 system.

B. Data Preprocessing

Before the classification experiments, we used the Python EEG toolkit MNE to preprocess the raw EEG signal data. We follow the data preprocessing method in [31]. First, the electrooculography data is removed and the EEG data were processed by a band-pass filter with a bandwidth of [2 30] Hz. Then, EEG data epochs were extracted according to event triggers and the EEG data was intercepted within the time interval of [−200 1000] ms, where the [−200 0] ms was used for baseline correction. In order to reduce the amount of data, the EEG epoch data was down-sampled to 100Hz and then used for model training.

C. Evaluation Metrics and Comparison Models

Since the number of non-target samples is far more than the number of target samples. RSVP classification has a class imbalance problem. In order to effectively evaluate the performance of the proposed CST-TVA-DRTL in the RSVP classification task, the balanced accuracy (BA), true positive rate (TPR), true negative rate (TNR) and area under the receiver operating characteristic curve (AUC) are used as evaluation metrics. BA is obtained by averaging TPR and TNR, which is used to measure the average classification accuracy for target and non-target samples. To verify the classification

performance of the proposed CST-TVA-DRTL method, five widely used EEG classification methods, including EEGNet, Deep ConvNet (DCN), EEGInception, EEGConformer and STSTNet were used for fair comparative experiments:

1) EEGNet [5]: This is a lightweight convolutional neural network that extracts the spatiotemporal features of EEG through convolution, depth-wise convolution and separable convolution. EEGNet has been proven to be effective in various EEG classification tasks but its simple network structure limits its representation learning capabilities.

2) DCN [11]: This is a deep convolutional neural network that mainly captures high-level features from EEG signals through four convolution blocks. This model can serve as a versatile tool for decoding EEG signals across various tasks. However, due to its traditional convolutional network structure, DCN is difficult to efficiently capture the multi-scale feature of EEG signals.

3) EEGInception [12]: To extract multi-scale time domain features, EEGInception introduces the Inception mechanism into the deep convolutional neural network. Each Inception module can extract multi-scale temporal features from EEG through convolution layers with varying convolution kernel sizes. This method demonstrates its efficacy in detecting ERP. Although this model can effectively extract multi-scale information, it cannot effectively capture the long-term dependencies of time domain features.

4) EEGConformer [33]: First, the low-level local features were extracted by using the temporal and spatial convolution. Then, the Transformer was used to capture the global correlation within the local temporal features for EEG classification. However, this model ignores the important time-frequency feature of EEG signals.

5) STSTNet [2]: This is a multi-view features based EEG decoding method, which can simultaneously extract the spatio-temporal and spectral-temporal features from EEG signals and spectrum images, and then fuses the multi-view features through the spatio-temporal-spectral Transformer. Although this model can comprehensively extract spatio-temporal and spectral-temporal features, it does not consider the multi-scale time domain information of EEG signals.

We also compare our method with three other transfer learning approaches:

1) **Domain-Adversarial Neural Networks (DANN)** [29]: This is a domain adaptation transfer learning method, which achieves the alignment of different domain features through a domain classifier and gradient reversal layer. It has been used for cross-subject transfer learning on EEG data [34]. Due to the large inter-individual differences in EEG, aligning features with the source domain may cause the loss of discriminative information.

2) **Adaptive Transfer Learning based on DCN (ATLDCN)** [35]: The ATLDCN handles the substantial inter-subject variability of EEG data through five adaptation scheme. However, the selection of adaptive transfer learning schemes relies on time-consuming optimization processes.

3) **Source-free Subject Adaptation (SFSA)** [36]: The SFSA transfer learning method generates source domain data through a classifier-based source domain data generator, and then aligns the target subject features with the generated source subject features. Although this method can utilize common information from different subjects, it ignores subject-specific information that contributes to the classification task.

The above comparison methods and proposed CST-TVA-DRTL are all implemented using the Python 3.9.7 and the PyTorch 1.13.1, and comparative experiments are conducted on the same hardware platform. The 5-fold cross-validation strategy is used to conduct comparative experiments. When training the model, all model parameters are optimized by the Adam optimizer. The initial learning rate and batch size are set to 10^{-4} and 64, respectively. Due to the data of RSVP paradigm has the characteristics of extremely unbalanced class, the oversampling strategy is adopt to balance data categories of the training set.

D. Overall Performance

The feature extraction capabilities and transfer performance of the proposed method and the above baseline methods are comprehensively compared in subject-dependent experiments and cross-subject experiments, respectively. In subject-dependent experiments, models are trained and tested on the same subject data. Following the setting of 5-fold cross-validation experiments, each subject data is evenly divided into five parts. In each fold experiment, one part of the data is used as the test set, one part is used as the validation set, and the other three parts are used as the training set. For a fair comparison, five deep learning-based EEG classification models such as DCN, EEGNet, EEGInception, EEGConformer and STSTNet are compared with CST-TVA method that without domain-rectified transfer learning. The results of subject-dependent experiments on Tsinghua and PhysioNet RSVP datasets are presented in Table I, where BA, TPR, TNR and AUC are the mean values of all subjects' results, and std is the standard deviation. The results in Table I show that CST-TVA achieves the best BA on Tsinghua and PhysioNet RSVP datasets. For Tsinghua dataset, CST-TVA reaches 92.56%, which is 1.96%, 1.65%, 0.68%, 1.09% and 0.55% higher than EEGNet, DCN, EEGInception, EEGConformer and STSTNet, respectively. Although our method is slightly lower than DCN on the TPR, it is higher than other methods on the TNR and AUC, reaching 0.9551 and 0.9415 respectively. We perform the paired-sample t-test between the proposed CST-TVA method and other methods. The adjusted p-values with Bonferroni correction in the significance test are provided in Table I. The results show that p-values of BA are less than 0.001 for all comparison methods. The significant improvements in BA suggests that the CST-TVA has stronger EEG feature extraction capabilities. For PhysioNet dataset, the CST-TVA achieved a BA of 72.02%, outperforming EEGNet ($p < 0.01$), DCN ($p < 0.01$), EEGInception ($p < 0.01$), EEGConformer ($p < 0.05$) and STSTNet ($p < 0.05$) by 2.66%, 1.01%, 1.06%, 0.89% and 0.66%, respectively. Although EEGConformer achieves the highest result on TPR, its TNR is lower than ours CST-TVA, which is caused by the class imbalance on the training set. In terms of TNR and AUC, CST-TVA achieves 0.7277 and 0.7448 respectively, which performs better than five baseline methods. Subject-dependent experiments on the

TABLE I
THE OVERALL COMPARISON OF SUBJECT-DEPENDENT CLASSIFICATION PERFORMANCE ON TSINGHUA AND PHYSIONET DATASETS

Dataset	Subject-Dependent Methods	BA±std (%)	TPR±std	TNR±std	AUC±std
Tsinghua	EEGNet [5]	90.60±4.20⁂	0.9084±0.0461†	0.9035±0.0406⁂	0.9117±0.0312⁂
	DCN [11]	90.91±4.14⁂	**0.9096**±0.0446†	0.9085±0.0405⁂	0.9238±0.0308⁂
	EEGInception [12]	91.88±3.92⁂	0.9052±0.0483†	0.9323±0.0320⁂	0.9270±0.0289⁂
	EEGConformer [33]	91.47±3.91⁂	0.8867±0.0619‡	0.9427±0.0292⁂	0.9255±0.0299⁂
	STSTNet [2]	92.01±3.47⁂	0.8923±0.0481*	0.9481±0.0256⁂	0.9322±0.0262⁂
	Our CST-TVA	**92.56**±3.58	0.8962±0.0474	**0.9551**±0.0195	**0.9415**±0.0227
PhysioNet	EEGNet [5]	69.36±5.91‡	0.7311±0.1007†	0.6561±0.0531‡	0.7079±0.0937*
	DCN [11]	71.01±6.79‡	0.7258±0.0583†	0.6944±0.1011‡	0.7132±0.0951*
	EEGInception [12]	70.96±6.55‡	0.7027±0.1017‡	0.7165±0.0615‡	0.7125±0.1080*
	EEGConformer [33]	71.13±6.50*	**0.7451**±0.0912†	0.6775±0.0375‡	0.7123±0.1001*
	STSTNet [2]	71.36±6.35*	0.7342±0.0795†	0.6931±0.0625‡	0.7174±0.0918†
	Our CST-TVA	**72.02**±6.14	0.7166 ±0.0655	**0.7238**±0.0561	**0.7281**±0.0962

where the best results indicate highlighted with bold font, the *, ‡ and ⁂ mean p-value < 0.05, 0.01 and 0.001 respectively and † indicate an insignificant advantage or disadvantage.

TABLE II
THE OVERALL COMPARISON OF CROSS-SUBJECT CLASSIFICATION PERFORMANCE ON TSINGHUA AND PHYSIONET DATASETS

Dataset	Cross-Subject Methods	BA±std (%)	TPR±std	TNR±std	AUC±std
Tsinghua	DANN [29]	92.54±3.43⁂	**0.9073**±0.0464†	0.9435±0.0258⁂	0.9438±0.0266⁂
	ATLDCN [35]	92.63±3.62⁂	0.9066±0.0498†	0.9459±0.0263⁂	0.9479±0.0329⁂
	SFSA [36]	92.71±3.48⁂	0.9012±0.0544†	0.9530±0.0189⁂	0.9496±0.0245⁂
	Our CST-TVA-DRTL	**93.07**±3.40	0.9025±0.0532	**0.9589**±0.0177	**0.9581**±0.0213
PhysioNet	DANN [29]	72.51±7.89‡	0.7219±0.0808‡	**0.7283**±0.0737†	0.7317±0.1079*
	ATLDCN [35]	72.71±5.44‡	0.7449±0.0767†	0.7092±0.0862‡	0.7407±0.0954†
	SFSA [36]	72.94±8.27‡	0.7326±0.0734‡	0.7262±0.0769†	0.7422±0.0902†
	Our CST-TVA-DRTL	**73.95**±5.96	**0.7513**±0.0601	0.7277±0.0465	**0.7448**±0.0691

where the best results indicate highlighted with bold font, the *, ‡ and ⁂ mean p-value < 0.05, 0.01 and 0.001 respectively and † indicate an insignificant advantage or disadvantage.

Tsinghua and PhysioNet RSVP datasets indicates that our proposed method can effectively extract discriminative features to achieve better balanced accuracy.

In cross-subject experiments, one subject from the dataset is selected as the target subject in turn for fine-tuning and testing the model, and the rest subjects are used as the source domain for pre-training the model. In the pre-training stage, only the source subjects' data are used to train the model. In the fine-tuning stage, the same 5-fold cross-validation strategy as in the subject-dependent experiment is adopted. The data from the target subject is divided into five parts, three parts for fine-tuning the model, one part for validation, and one part for testing. The results of cross-subject experiments are presented in Table II. From Table II, we can observe that compared with other transfer learning methods such DANN, ATLDCN and SFSA, our CST-TVA-DRTL achieves better results on both datasets. For Tsinghua dataset, the BA of our method reaches 93.07%, which is 0.53%, 0.44% and 0.36% higher than three baseline transfer learning methods, respectively. In terms of TNR and AUC, our method achieves the best results, reaching 0.9589 and 0.9581 respectively. The adjusted p-values of BA are less than 0.001 for all comparison methods, indicating that the proposed CST-TVA-DRTL not only demonstrates excellent performance in specific subjects but also exhibits statistically significant superiority overall. For PhysioNet dataset, our CST-TVA-DRTL also outperforms other methods. The BA of our method reaches 73.95%, which is 1.44%, 1.24% and 1.01% higher than other methods, respectively. The results of the paired-sample t-test show that the improvement is significant compared to DANN ($p < 0.01$), ATLDCN ($p < 0.01$) and SFSA ($p < 0.01$). Cross-subject experimental results on the Tsinghua and PhysioNet RSVP datasets indicate that our CST-TVA-DRTL achieves a substantially higher BA than DANN, ATLDCN and SFSA, which confirms that our proposed transfer learning framework can leverage data the data from other subjects to boost the method's performance.

Fig.5 illustrates the mean and standard deviation of BA for each subject in the 5-fold cross-validation experiment. The mean and standard deviation for all subjects are also shown in the figure. As shown, after adopting domain-rectified transfer learning, the classification accuracy of the CST-TVA-DRTL on most subjects is improved. In particular, the BA of the third subject in the Tsinghua dataset was improved from 92.48% to 94.59%. It is evident that the incorporation of domain-rectified transfer learning can effectively use cross-subject information to enhance the performance of the CST-TVA-DRTL. In addition, the standard deviation of the BAs obtained by the CST-TVA-DRTL on different subjects is smaller, indicating that domain-rectified transfer learning can also make the CST-TVA-DRTL more stable.

IV. DISCUSSION

A. Comparison With the State-of-the-Art Methods

We compare the proposed method with recently reported state-of-the-art (SOTA) methods to demonstrate its effectiveness. For fair comparison, these SOTAs for subject-dependent

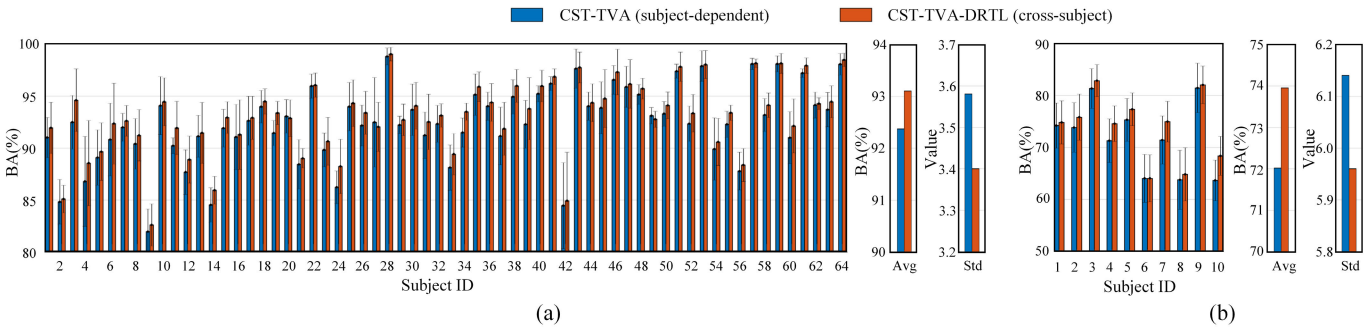

Fig. 5. Comparison between CST-TVA and our CST-TVA-DRTL on (a) Tsinghua dataset and (b) PhysioNet dataset.

TABLE III
RESULTS OBTAINED BY THE STATE-OF-THE-ART SUBJECT-DEPENDENT METHODS ON TSINGHUA AND PHYSIONET DATASETS

Dataset	Subject-Dependent Methods	Year	BA (%)	AUC
Tsinghua	iEEGNet [8]	2022	-	0.9279
	XGB-DIM [9]	2023	84.40	0.9051
	Our CST-TVA	2023	**92.56**	**0.9415**
PhysioNet	PPNN [7]	2022	67.23	-
	DRL [16]	2022	68.80	-
	Our CST-TVA	2023	**72.02**	**0.7281**

where the best results indicate highlighted with bold font.

TABLE IV
RESULTS OBTAINED BY THE STATE-OF-THE-ART CROSS-SUBJECT METHODS ON TSINGHUA AND PHYSIONET DATASETS

Dataset	Cross-Subject Methods	Year	BA (%)	AUC
Tsinghua	MACRO [19]	2021	-	0.9309
	MCGRAM [15]	2022	-	0.9352
	Our CST-TVA-DRTL	2023	**93.07**	**0.9581**
PhysioNet	xDAWN-SVM# [18]	2020	61.37	0.6256
	Our CST-TVA-DRTL	2023	**73.95**	**0.7448**

where methods marked with # indicates its result obtained by ourselves.

experiments and cross-subject experiments are separately compared with CST-TVA and CST-TVA-DRTL methods in Table III and Table IV respectively. Since the SOTAs mainly report BA and AUC results in their papers, we list the results of BA and AUC in the tables for comparison.

As shown in Table III, the proposed CST-TVA exhibits significant improvements compared to the SOTAs on the Tsinghua and PhysioNet RSVP datasets. For Tsinghua dataset, XGB-DIM [9] yields the worst results due to its reliance on machine learning methods, which struggle to effectively extract the key features from RSVP EEG signals. iEEGNet [8] is an enhanced version of the EEGNet model, achieves an AUC of 0.9274, second only to the CST-TVA. For the PhysioNet dataset, the CST-TVA method outperforms the PPNN [7] and DRL [16], both based on deep neural networks, by 4.76% and 3.22% in terms of BA, respectively. This indicates that the CST-TVA possesses stronger representation learning capabilities. The main reason is that our proposed CST-TVA can simultaneously learn multi-scale temporal features and multi-view spectral features.

Table IV presents a comparison between the proposed cross-subject method CST-TVA-DRTL and other SOTA cross-subject approaches. As can be seen from the table, the CST-TVA-DRTL achieves the best results on both datasets. For the Tsinghua dataset, CST-TVA-DRTL achieves an AUC of 0.9581, while the MACRO [19] and MCGRAM [15] only achieve AUCs of 0.9309 and 0.9352, respectively. This may be attributed to the fact that MACRO and MCGRAM do not consider subject-specific information, thereby weakening their adaptability to the target domain data. For the PhysioNet dataset, the proposed CST-TVA-DRTL method significantly outperforms the xDAWN-SVM [18] approach, highlighting the weaker feature transfer capabilities of machine learning methods. In summary, the proposed CST-TVA-DRTL exhibits superior performance compared to existing SOTA methods, as it not only leverages the temporal and spectral features of EEG signals but also utilizes subject-specific information to enhance cross-subject transfer learning capabilities.

B. Ablation Studies

The proposed CST-TVA-DRTL method primarily consists of three key components: CST, TVA, and DRTL. These components are designed to empower the model with the capabilities of extracting temporal features, extracting spectral features, and performing transfer learning, respectively. To investigate the impact of these components on the model's classification performance, a series of ablation experiments were conducted. The results of the ablation experiments are presented in Table V.

1) Efficacy of CST: As shown in Table V, the BA of CST on the Tsinghua and PhysioNet datasets reached 91.94% and 71.37%, respectively, which increased by 1.03%, 0.36% compared with the baseline DCN method. The reason for this is that DCN model only considers the single scale temporal characteristics of EEG signal, and it is difficult to effectively capture the difference in the time-domain waveform of the target and non-target EEG. The CST can simultaneously extract different scales temporal features and model their temporal dependencies, which improves its ability to extract temporal features, and leading to better RSVP classification results. To investigate the impact of adaptive weighting on the performance of CST, we compared the CST with a multi-scale feature (MSF) extraction method without adaptive weighting. As shown in Table V, the BA of CST exhibits a significant ($p < 0.05$) improvement compared to MSF on both datasets. This indicates that the adaptive weighting in CST effectively enhances the performance of the model.

TABLE V
THE RESULTS OF ABLATION EXPERIMENTS ON TSINGHUA AND PHYSIONET RSVP DATASETS

Dataset	Methods	BA±std (%)	TPR±std	TNR±std	AUC±std
Tsinghua	DCN (Baseline)	90.91±4.14	**0.9096**±0.0446	0.9085±0.0405	0.9238±0.0308
	MSF	91.75±3.95⁂	0.9002±0.0479†	0.9347±0.0332⁂	0.9263±0.0287⁂
	SVF	91.30±3.82†	0.8658±0.0549†	**0.9602**±0.0243⁂	0.9231±0.0309 †
	CST	91.94±3.57⁂	0.8896±0.0496†	0.9492±0.0285⁂	0.9291±0.0273⁂
	TVA	91.87±3.62⁂	0.8974±0.0485⁂	0.9400±0.0298†	0.9294±0.0266⁂
	CST+TVA	92.56±3.55⁂	0.8962±0.0474†	0.9551±0.0195⁂	0.9415±0.0227⁂
	CST+TVA+DRTL(Ours)	**93.07**±3.40⁂	0.9025±0.0532‡	0.9589±0.0177⁂	**0.9581**±0.0213⁂
PhysioNet	DCN (Baseline)	71.01±6.79	0.7258±0.0583	0.6944±0.1011	0.7132±0.0951
	MSF	71.12±5.99*	0.7116±0.0629†	0.7108±0.0754*	0.7141±0.0933†
	SVF	70.18±7.33†	0.7225±0.0483†	0.6810±0.0927†	0.7039±0.1082†
	CST	71.37±5.58*	0.7135±0.0531†	0.7140±0.0669*	0.7197±0.0928*
	TVA	70.66±5.37‡	0.7181±0.0501†	0.6951±0.0738‡	0.7054±0.1057*
	CST+TVA	72.02±6.14‡	0.7166 ±0.0655†	0.7238±0.0561‡	0.7281±0.0962‡
	CST+TVA+DRTL(Ours)	**73.95**±5.96‡	**0.7513**±0.0601‡	**0.7277**±0.0465*	**0.7448**±0.0691‡

where the best results indicate highlighted with bold font, the *, ‡ and ⁂ mean p-value < 0.05, 0.01 and 0.001 respectively and † indicate an insignificant advantage or disadvantage.

TABLE VI
MODEL PARAMETERS AND COMPUTATION TIMES OF DIFFERENT METHODS

Methods	BA (%)	Parameters (1×10^3)	Training time (s)	Inference time (s)
EEGNet [5]	90.60	3.03	12.18	0.11
DCN [11]	90.91	127.47	26.60	0.23
EEGInception [12]	91.88	20.51	23.54	0.24
EEGConformer [33]	91.47	272.50	48.59	0.31
STSTNet [2]	92.01	283.07	22.81	0.34
Our CST-TVA	92.56	290.65	32.67	0.45
DANN [29]	92.54	61.98	111.65	0.21
ATLDCN [35]	92.63	127.47	74.13	0.26
SFSA [36]	92.71	177.50	57.19	0.31
Our CST-TVA-DRTL	**93.07**	963.82	74.80	0.46

2) Efficacy of TVA: The TVA is designed to extract multi-view spectral features from EEG time-frequency images. To validate the effectiveness of TVA, we compared it with a conventional single-view spectral feature (SVF) extraction method. As shown in Table V, the TVA method with triple-view attention mechanism demonstrates a significant ($p < 0.01$) improvement in BA compared to SVF on both datasets. This suggests that the triple-view attention mechanism is more effective in capturing spectral features. However, if only time-frequency features are used, the results obtained by TVA are worse than those of CST, which suggests that temporal features are indispensable for the RSVP classification task. After incorporating TVA into the CST method, the CST+TVA method surpasses the CST method by 0.62% and 0.65% in BAs on the Tsinghua and PhysioNet datasets, respectively. The superior performance of CST+TVA is due to its ability to extract both temporal and spectral features of EEG signals, while the CST method only focuses on temporal features. The incorporation of TVA into the CST+TVA method enables it to capture key spectral features of multi-channel time-frequency images from different perspectives, thus enhancing its ability of representation learning.

3) Efficacy of DRTL: After introducing DRTL into CST+TVA method, the BA of proposed CST+TVA+DRTL on both datasets increased by 0.51% and 1.93% respectively. The proposed DRTL framework is beneficial because it can use more subjects EEG data to improve the feature extraction ability of the model through the transfer learning strategy, and use domain-specific representations to rectify domain-invariant representations to adapt to the target subject data. The domain-specific and domain-invariant representations learned by the proposed CST-TVA-DRTL on Tsinghua dataset and PhysioNet dataset are visualized in Fig.6. We can see that the domain-specific representations obtained by the proposed model from source subjects and target subject have clear boundaries, while domain-invariant representations of source subjects and target subject are indistinguishable. This shows that our model can simultaneously extract individualized information and common invariant information of different subjects' data, thus effectively improving the transfer learning performance of the model.

C. Saliency Map Analysis of EEG Channels

We investigate the correlation between target and non-target visual stimuli and different channels of EEG signals by using the saliency map method [37]. The saliency map is a commonly used method to visualize the classification inference process of deep learning models in the field of computer vision [38], which can reveal the importance of each part of the input data to the classification results through a single gradient back-propagation. In this study, to investigate the importance of different EEG channels on target and not-target classification, the EEG data were input into the well-trained DCN model, and then the Gradient-weighted Class Activation Mapping (Grad-CAM) method was used to obtain the importance score of the EEG data [39]. The channel importance score was obtained by accumulating the importance score of all data points in each channel. Fig.7 visualizes the averaged and normalized channel importance score of two classes of EEG data from the Tsinghua and PhysioNet datasets. As shown in Fig.7 (a), target visual stimuli evoke brain activity in both prefrontal and occipital cortexes, while non-target

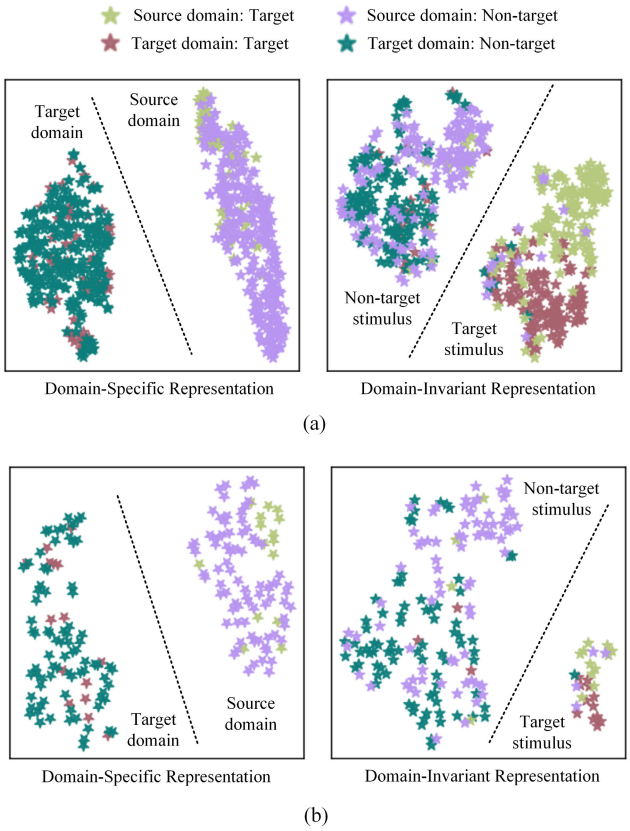

Fig. 6. The t-SNE visualization of domain-specific and domain-invariant representations learned by the proposed CST-TVA-DRTL on (a) Tsinghua dataset and (b) PhysioNet dataset.

visual stimuli only elicit brain activity in the occipital cortex. Due to the EEG signals provided in the PhysioNet dataset have only 8 channels: P7, P8, PO7, PO3, PO4, PO8, O1, O2, the occipital cortex activities evoked by target and non-target visual stimuli are similar as shown in Fig.7 (b), which means that it is challenging to accurately identify target versus non-target visual stimuli based solely on occipital cortex activities. Due to the lack of prefrontal cortex signals in PhysioNet dataset, the average balanced accuracy of the proposed method can only reach 73.95%, while it can reach 93.07% on Tsinghua dataset. Therefore, the prefrontal cortex signals are crucial to the RSVP classification task.

In order to visually explore the differences between target and non-target EEG signals, we conducted a visualization of the multi-channel EEG signals in both spatial and temporal dimensions in Fig.8. From the scalp topography in Fig.8 (a), it can be observed that for non-target visual stimuli, there is periodic brain activity in the occipital visual area, whereas for target visual stimuli, significant activity appears in the prefrontal cortex at around 300ms. The same results can be clearly observed from the time-domain waveforms as well. The target EEG waveforms at channel O2 exhibited clear P300 and N400 components, while the non-target EEG waveforms were near-sinusoidal signals. These findings are consistent with previous studies [31], demonstrating a strong correlation between prefrontal cortex activity and target visual stimuli. Since the PhysioNet dataset only provides EEG signals from the visual area, it is difficult to observe significant differences between

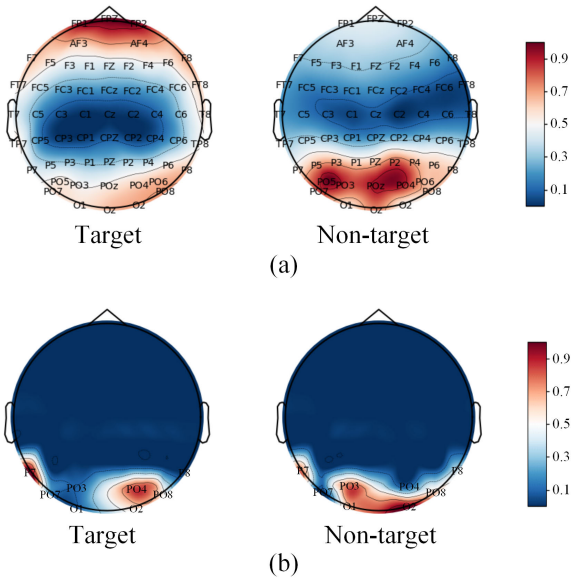

Fig. 7. Saliency maps on (a) Tsinghua dataset and (b) PhysioNet dataset.

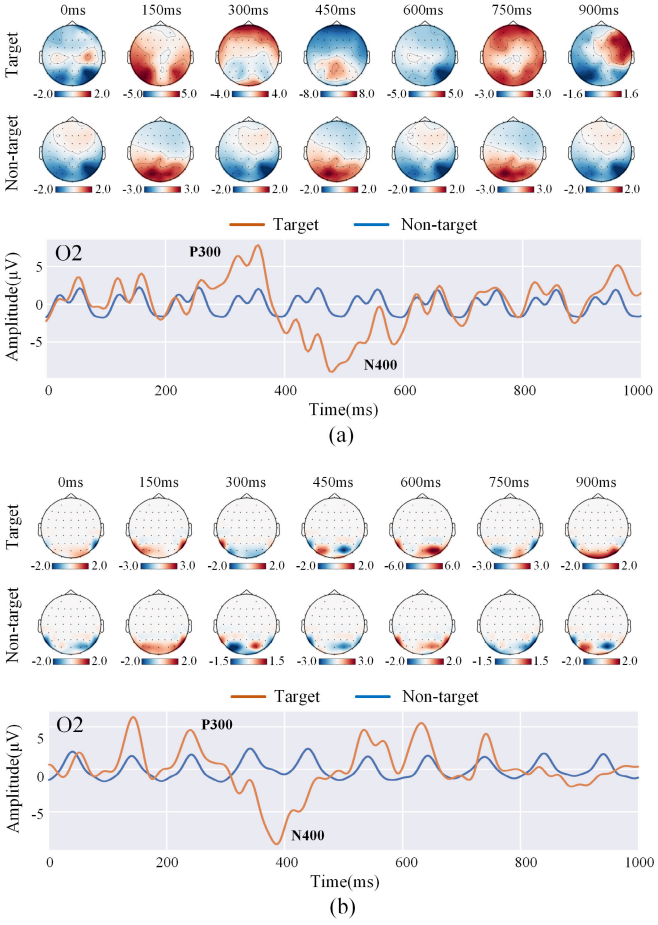

Fig. 8. Visualization of target and non-target EEG from spatial and temporal dimensions. (a) Tsinghua dataset, (b) PhysioNet dataset.

target and non-target EEG signals from a spatial dimension. However, from the temporal dimension, clear components such as the P300 and N400 can be observed in the target EEG, while the non-target EEG exhibits periodic patterns.

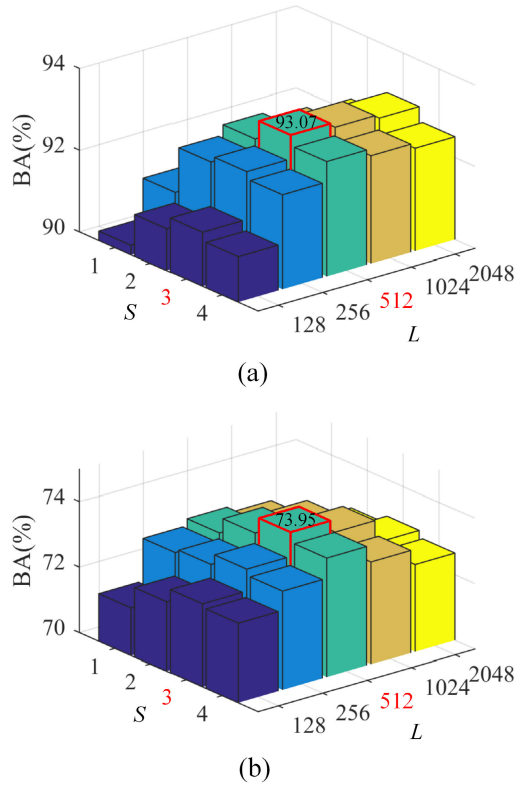

Fig. 9. Performance comparison with various S and L on (a) Tsinghua dataset and (b) PhysioNet dataset.

D. Analysis of Hyperparameter Settings

As the hyperparameters of deep learning models have a significant impact on performance, it is crucial to set appropriate values for optimal model [40], [41], [42], [43], [44]. In this study, we employed a simple yet effective grid search strategy to optimize hyperparameters such as the number of scales S for multi-scale temporal features and the dimension of domain-specific representation L. Performance comparison results of the proposed model under different parameter settings are presented in Fig.9. As shown, increasing the number of scales from 1 to 3 leads to improved model performance, indicating that multi-scale temporal features provide more discriminative information. However, when the number of scales exceeds 3, model performance slightly decreases due to redundant information contained within additional scale temporal features. Similarly, it can be observed that setting L at 128 results in lower balanced accuracy because smaller representation dimensions contain less information resulting in loss of details; whereas increasing L up to 512 leads to optimal model performance. However, exceeding an L value greater than 1024 may cause redundancy and negatively affect efficiency.

E. Computational Complexity

The computational complexity of our proposed CST-TVA-DRTL is evaluated based on the number of parameters and the training and inference time of the model. For comparison purposes, Table VI presents the average BA, model parameter count, training time, and inference time of CST-TVA-DRTL and its competitors on the Tsinghua dataset. It is important to note that the time consumption of all models was measured on 1×10^3 a hardware platform with an Intel Core i7-7700 3.60GHz CPU and NVIDIA GTX 1080 GPU. The software environment includes Python 3.9.7, PyTorch 1.13.1, and CUDA 11.0. From Table VI, it can be observed that the EEGNet has the lowest computational complexity. However, due to its simple network architecture, it struggles to effectively capture the most discriminative features in EEG signals, resulting in the lowest average BA. In contrast, although the proposed CST-TVA-DRTL involves more parameters due to the introduction of Transformer and feature encoder, its average BA is significantly improved. Furthermore, compared with the existing methods, the training time and inference time of the proposed CST-TVA-DRTL do not exhibit a significant increase. This is primarily attributed to the fact that the majority of the parameters in the CST-TVA-DRTL originate from the fully connected layers of the feature encoders, and the fully connected layers are computationally efficient.

F. Limitations and Future Directions

The proposed CST-TVA-DRTL framework achieves better RSVP classification results than existing methods, but it still has some limitations that need to be addressed in future work. First, all channels of EEG data were used for RSVP classification, but only a subset of channels were significantly correlated with the RSVP task. The performance of the CST-TVA-DRTL can be degraded by noisy signals on uncorrelated channels. Thus, an adaptive channel enhancement strategy will be further studied to reduce the interference of noisy channels. Second, although our proposed method effectively improves the transfer performance, it requires data from multiple subjects, which may limit its applicability to datasets with a small number of subjects. Therefore, we will investigate further improving RSVP classification performance by generating minority class data through generative adversarial networks in future work.

V. Conclusion

This paper presents a novel CST-TVA-DRTL framework for RSVP classification. Specially, this framework leverages a CST temporal feature extractor to obtain multi-scale temporal features from EEG signals and characterize the temporal feature dependencies across different scales. In addition, a TVA spectral feature extractor is adopted to capture discriminative spectral features of multi-channel EEG spectrogram from three different views. Moreover, a DRTL framework is designed to improve the cross-subject transfer learning performance by simultaneously exploiting the common invariant information and subject-specific information of multiple subject data. Experimental results on the Tsinghua and PhysioNet RSVP datasets confirm that our proposed CST-TVA-DRTL outperforms the state-of-the-art methods, demonstrating its viability as a solution for RSVP classification.

References

[1] A. Subasi, A. Saikia, K. Bagedo, A. Singh, and A. Hazarika, "EEG-based driver fatigue detection using FAWT and multiboosting approaches," *IEEE Trans. Ind. Informat.*, vol. 18, no. 10, pp. 6602–6609, Oct. 2022.

[2] J. Luo et al., "A dual-branch spatio-temporal-spectral transformer feature fusion network for EEG-based visual recognition," *IEEE Trans. Ind. Informat.*, vol. 20, no. 2, pp. 1721–1731, Feb. 2024.

[3] J. Pan, X. Chen, N. Ban, J. He, J. Chen, and H. Huang, "Advances in P300 brain–computer interface spellers: Toward paradigm design and performance evaluation," *Frontiers Hum. Neurosci.*, vol. 16, Dec. 2022, Art. no. 1077717.

[4] W. Wei, S. Qiu, Y. Zhang, J. Mao, and H. He, "ERP prototypical matching net: A meta-learning method for zero-calibration RSVP-based image retrieval," *J. Neural Eng.*, vol. 19, no. 2, Apr. 2022, Art. no. 026028.

[5] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 15, no. 5, Jul. 2018, Art. no. 056013.

[6] M. Xu, Y. Chen, Y. Wang, D. Wang, Z. Liu, and L. Zhang, "BWGAN-GP: An EEG data generation method for class imbalance problem in RSVP tasks," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 251–263, 2022.

[7] F. Li et al., "Phase preservation neural network for electroencephalography classification in rapid serial visual presentation task," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 6, pp. 1931–1942, Jun. 2022.

[8] H. Zhang, Z. Wang, Y. Yu, H. Yin, C. Chen, and H. Wang, "An improved EEGNet for single-trial EEG classification in rapid serial visual presentation task," *Brain Sci. Adv.*, vol. 8, no. 2, pp. 111–126, Jun. 2022.

[9] B. Li, S. Zhang, Y. Hu, Y. Lin, and X. Gao, "Assembling global and local spatial–temporal filters to extract discriminant information of EEG in RSVP task," *J. Neural Eng.*, vol. 20, no. 1, Feb. 2023, Art. no. 016052.

[10] L. Parra et al., "Spatiotemporal linear decoding of brain state," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 107–115, Jan. 2008.

[11] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Aug. 2017.

[12] E. Santamaría-Vázquez, V. Martínez-Cagigal, F. Vaquerizo-Villar, and R. Hornero, "EEG-inception: A novel deep convolutional neural network for assistive ERP-based brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2773–2782, Dec. 2020.

[13] J.-S. Kang, S. Kavuri, and M. Lee, "ICA-evolution based data augmentation with ensemble deep neural networks using time and frequency kernels for emotion recognition from EEG-data," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 616–627, Apr. 2022.

[14] H. Zhang, X. Zhao, Z. X. Wu, B. Sun, and T. Li, "Motor imagery recognition with automatic EEG channel selection and deep learning," *J. Neural Eng.*, vol. 18, Feb. 2020, Art. no. 016004.

[15] Z. Li, C. Yan, Z. Lan, D. Tang, and X. Xiang, "MCGRAM: Linking multi-scale CNN with a graph-based recurrent attention model for subject-independent ERP detection," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 12, pp. 5199–5203, Dec. 2022.

[16] F. Li et al., "Decoupling representation learning for imbalanced electroencephalography classification in rapid serial visual presentation task," *J. Neural Eng.*, vol. 19, no. 3, May 2022, Art. no. 036011.

[17] W. Wei, S. Qiu, X. Ma, D. Li, C. Zhang, and H. He, "A transfer learning framework for RSVP-based brain computer interface," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 2963–2968.

[18] H. He and D. Wu, "Transfer learning for brain–computer interfaces: A Euclidean space data alignment approach," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 2, pp. 399–410, Feb. 2020.

[19] Z. Lan, C. Yan, Z. Li, D. Tang, and X. Xiang, "MACRO: Multi-attention convolutional recurrent model for subject-independent ERP detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 1505–1509, 2021.

[20] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "XDAWN algorithm to enhance evoked potentials: Application to brain–computer interface," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 8, pp. 2035–2043, Aug. 2009.

[21] Q. Guo, L. X. Fang, R. Wang, and C. M. Zhang, "Multivariate time series forecasting using multiscale recurrent networks with scale attention and cross-scale guidance," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 2023, doi: 10.1109/TNNLS.2023.3326140.

[22] Y. Li et al., "Global transformer and dual local attention network via deep-shallow hierarchical feature fusion for retinal vessel segmentation," *IEEE Trans. Cybern.*, vol. 53, no. 9, pp. 5826–5839, Sep. 2023.

[23] Y. Li, Y. Liu, Y.-Z. Guo, X.-F. Liao, B. Hu, and T. Yu, "Spatio-temporal-spectral hierarchical graph convolutional network with semisupervised active learning for patient-specific seizure prediction," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 12189–12204, Nov. 2022.

[24] L. Guo, T. Yu, S. Zhao, X. Li, X. Liao, and Y. Li, "CLEP: Contrastive learning for epileptic seizure prediction using a spatio-temporal-spectral network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 3915–3926, 2023.

[25] W. G. Cui, M. Y. Sun, Q. X. Dong, Y. Z. Guo, X. F. Liao, and Y. Li, "A multiview sparse dynamic graph convolution-based region-attention feature fusion network for major depressive disorder detection," *IEEE Trans. Computat. Social Syst.*, early access, Jul. 2023, doi: 10.1109/TCSS.2023.3291950.

[26] X. Y. Chen, S. N. Wang, B. Fu, M. S. Long, and J. M. Wang, "Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2019, pp. 1908–1918.

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

[28] Z. Kou, K. C. You, M. S. Long, and J. M. Wang, "Stochastic normalization," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2020, pp. 16304–16314.

[29] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. 32nd Int. Conf. Mach. Learn.*, Jun. 2015, pp. 1180–1189.

[30] Y. T. Li, M. Murias, G. Dawson, and D. E. Carlson, "Extracting relationships by multi-domain matching," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2018, pp. 6799–6810.

[31] S. Zhang, Y. Wang, L. Zhang, and X. Gao, "A benchmark dataset for RSVP-based brain–computer interfaces," *Frontiers Neurosci.*, vol. 14, Oct. 2020, Art. no. 568000.

[32] A. Matran-Fernandez and R. Poli, "Towards the automated localisation of targets in rapid image-sifting by collaborative brain–computer interfaces," *PLoS ONE*, vol. 12, no. 5, May 2017, Art. no. e0178498.

[33] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG conformer: Convolutional transformer for EEG decoding and visualization," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 710–719, 2023.

[34] Y. Li, W. Zheng, Y. Zong, Z. Cui, T. Zhang, and X. Zhou, "A bi-hemisphere domain adversarial neural network model for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 494–504, Apr. 2021.

[35] K. Zhang, N. Robinson, S.-W. Lee, and C. Guan, "Adaptive transfer learning for EEG motor imagery classification with deep convolutional neural network," *Neural Netw.*, vol. 136, pp. 1–10, Apr. 2021.

[36] P. Lee, S. Jeon, S. Hwang, M. Shin, and H. Byun, "Source-free subject adaptation for EEG-based visual recognition," in *Proc. 11th Int. Winter Conf. Brain–Comput. Interface (BCI)*, Feb. 2023, pp. 1–6.

[37] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. Int. Conf. Learn. Represent.*, Apr. 2014, pp. 1–8.

[38] M. Y. Sun et al., "Attention-rectified and texture-enhanced cross-attention transformer feature fusion network for facial expression recognition," *IEEE Trans. Ind. Informat.*, vol. 19, no. 12, pp. 11823–11832, Dec. 2023.

[39] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[40] J. Luo et al., "Visual image decoding of brain activities using a dual attention hierarchical latent generative network with multiscale feature fusion," *IEEE Trans. Cognit. Develop. Syst.*, vol. 15, no. 2, pp. 761–773, Jun. 2023.

[41] S. Li, H. Chen, M. Wang, A. A. Heidari, and S. Mirjalili, "Slime mould algorithm: A new method for stochastic optimization," *Future Gener. Comput. Syst.*, vol. 111, pp. 300–323, Oct. 2020.

[42] Y. Li, Y. Zhang, W. Cui, B. Lei, X. Kuang, and T. Zhang, "Dual encoder-based dynamic-channel graph convolutional network with edge enhancement for retinal vessel segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 8, pp. 1975–1989, Aug. 2022.

[43] X. Zhou, W. Gui, A. A. Heidari, Z. Cai, G. Liang, and H. Chen, "Random following ant colony optimization: Continuous and binary variants for global optimization and feature selection," *Appl. Soft Comput.*, vol. 144, Sep. 2023, Art. no. 110513.

[44] Y. Li, Y. Liu, W.-G. Cui, Y.-Z. Guo, H. Huang, and Z.-Y. Hu, "Epileptic seizure detection in EEG signals using a unified temporal-spectral squeeze-and-excitation network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 4, pp. 782–794, Apr. 2020.