OPEN ACCESS
OPUS AUGSBURG

Universität Augsburg
Universitätsbibliothek

Multi-view domain-adaptive representation learning for EEG-based emotion recognition

Chao Li, Ning Bian, Ziping Zhao, Haishuai Wang, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Li, Chao, Ning Bian, Ziping Zhao, Haishuai Wang, and Björn W. Schuller. 2023. "Multi-view domain-adaptive representation learning for EEG-based emotion recognition." Information Fusion 104 (November): 102156. https://doi.org/10.1016/j.inffus.2023.102156.

Nutzungsbedingungen / Terms of use: **CC BY-NC-ND 4.0**

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY-NC-ND 4.0: Creative Commons: Namensnennung - Nicht kommerziell - Keine Bearbeitung
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de

Multi-view domain-adaptive representation learning for EEG-based emotion recognition

Chao Li [a], Ning Bian [a], Ziping Zhao [a,*], Haishuai Wang [b], Björn W. Schuller [c]

[a] College of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China
[b] College of Computer Science, Zhejiang University, Hangzhou, China
[c] University of Augsburg, Germany and GLAM, Imperial College London, UK

ARTICLE INFO

Keywords:
Affective computing
Cross-attention
Domain adaptation
EEG
Emotion recognition
Multi-view learning

ABSTRACT

Current research suggests that there exist certain limitations in EEG emotion recognition, including redundant and meaningless time-frames and channels, as well as inter- and intra-individual differences in EEG signals from different subjects. To deal with these limitations, a Cross-attention-based Dilated Causal Convolutional Neural Network with Domain Discriminator (CADD-DCCNN) for multi-view EEG-based emotion recognition is proposed to minimize individual differences and automatically learn more discriminative emotion-related features. First, differential entropy (DE) features are obtained from the raw EEG signals using short-time Fourier transform (STFT). Second, each channel of the DE features is regarded as a view, and the attention mechanisms are utilized at different views to aggregate the discriminative affective information at the level of the time-frame of EEG. Then, a dilated causal convolutional neural network is employed to distill nonlinear relationships among different time frames. Next, a feature-level fusion is used to fuse features from multiple channels, aiming to explore the potential complementary information among different views and enhance the representational ability of the feature. Finally, to minimize individual differences, a domain discriminator is employed to generate domain-invariant features, which projects data from both the different domains into the same data representation space. We evaluated our proposed method on two public datasets, SEED and DEAP. The experimental results illustrate that our CADD-DCCNN method outperforms the SOTA methods.

1. Introduction

Emotion is a state that combines a human's feelings, thoughts, and behaviors, and which influences their rational decision-making, perception and cognition; accordingly, emotion has a significant impact on interpersonal communication [1]. Therefore, emotions contribute greatly in a wide range of studies. Emotion recognition has also been widely put into practice in many fields like depression diagnosis [2] and human–computer interaction [3]. Typically, multiple modalities are used for emotion recognition, including facial expressions [4], body gestures [5], voice [6], electrocardiography (ECG) [7], electroencephalography (EEG) [8], and electromyography (EMG) [9]. A great number of studies [10,11] have found that multi-view EEG signals, which have a strong relationship with emotion, are more difficult to disguise compared to other modalities. Therefore, they can be utilized as an effective approach for detecting emotions [12,13].

In EEG-based emotion recognition, significant progress has been made by numerous researchers in recent years. However, there are still two problems in urgent need of a solution. First, multi-view EEG signals can vary significantly among individuals due to differences in brain structure and patterns of brain activity across different subjects. Training a common model that can adopt across different datasets or subjects is a challenging task. Therefore, the primary problem in this context is that of how to deal with individual differences between different subjects or even different sessions of the same subject. Second, due to the varying contributions of the different time frames and channels of multi-view EEG signals to emotion recognition, it is vital to develop better ways of identifying and utilizing these signals in emotion recognition classifiers. Therefore, it will be necessary to obtain more distinguishing spatio-temporal features related to emotions to boost the effectiveness of emotion recognition.

To address the previously mentioned concerns in multi-view EEG-based emotion recognition, we propose a Cross-attention-based Dilated Causal Convolutional Neural Network with Domain Discriminator for multi-view EEG-based emotion recognition called CADD-DCCNN. We

* Corresponding author.
E-mail addresses: superlee@tjnu.edu.cn (C. Li), 2111090041@stu.tjnu.edu.cn (N. Bian), zhaoziping@tjnu.edu.cn (Z. Zhao), haishuai.wang@zju.edu.cn (H. Wang), schuller@tum.de (B.W. Schuller).

use multiple source domains that correspond to one target domain for domain adaptation to learn the commonalities between different domains. Moreover, a multi-view cross-attention mechanism is applied to learn the connection between EEG signals and emotional stimuli from multiple channel views, thereby enhancing the efficiency and stability of emotion recognition. The CADD-DCCNN method we propose takes a multi-view EEG signal as input, which is represented by a sequence of data from different electrodes, and produces an emotion label corresponding to this input. We focus on resolving several key issues in emotion recognition by adopting the following two strategies: (1) we aim to explore the complementary information between multiple channels through multi-view learning and combine attention mechanisms to identify the most informative spatio-temporal features for emotion recognition by selecting the most relevant channels and time frames; (2) we use domain adaptation techniques to eliminate individual differences among different subjects and build domain-invariant features.

Channel and time frame selection. Collected multi-view EEG signals have varying relevance to emotion, and different time frames may also activate emotions to varying degrees. In this paper, our purpose is to identify the most informative features for emotion recognition by assigning different weights to different channels and time frames. We use neuro-physiological research and a data-driven approach to explore subtle relations both within and between channels and time frames. We treat each channel as a view and analyze EEG signals from multiple views. To achieve this goal, a multi-view cross-attention mechanism is used in our proposed CADD-DCCNN. Firstly, attention mechanisms are applied within each view to explore the activation levels of different time frames. Then, through multi-view learning, we explore the importance of different views to search for the optimal channels and time frames relevant to emotions. We employ multi-view learning to explore complementarity and correlation between multi-view EEG signals, which helps enhance the performance and generalizability of emotion recognition.

Domain-invariant features. Because of the shift in data distribution among different individuals, many previous studies in multi-view EEG-based emotion recognition build the model based on an individual's brain responses. Although user-dependent models are popular, some recent studies [14–16] suggest building specially designed user-independent models. Accordingly, to deal with the problem of distribution shift in data, we integrate a domain discriminator to limit the distributions of the features obtained from the training (source) and testing (target) data to be similar.

In summary, our proposed CADD-DCCNN method utilizes multi-view cross-attention mechanism to extract emotion-related features that are distinctive by capturing the nonlinear connections between multiple channels. Furthermore, it utilizes a global domain discriminator to ensure a domain-invariant data representation. Our CADD-DCCNN method is evaluated on two widely used EEG emotional datasets (SEED [17] and DEAP [18]) to demonstrate its effectiveness. Additionally, ablation studies are executed to exhibit the effectiveness of our multi-view learning module, cross-attention mechanism, the dilated causal convolutional neural network, and domain discriminator module. In particular, the primary contributions of this paper are:

(1) We employ a multi-view cross-attention mechanism, where each channel of the DE features is considered as a view. Attention mechanisms are then applied to each view to extract discriminative information for each emotion. We explore the importance of different time frames on each channel and uncover the complementarity between different views to enhance the performance of our method. Our experimental results indicated the effectiveness of multi-view cross-attention mechanism in selecting emotion-related channels and time frames.
(2) A dilated causal convolutional neural network is utilized to capture the causal interactions among the features.
(3) A domain discriminator is integrated to generate a common feature space that constrains similar feature distributions between different (source and target) domains; this allows us to not only reduce discrepancy in data distribution in cross-subject scenarios but also improve model generalization in cross-session experiments.
(4) We propose a framework for multi-view EEG-based emotion recognition, CADD-DCCNN, which addresses the challenges of channel and time frame selection and building domain-invariant features. We assess our proposed approach using two publicly available datasets. Our approach obtains the SOTA performance on both datasets. Specifically, on SEED, it carries out an average accuracy of 92.44%. On DEAP, it achieves mean accuracies of 69.45% and 70.50% for valence and arousal, respectively.

The remainder of this paper is organized as follows. Section 2 describes the related works. The proposed CADD-DCCNN-based emotion recognition method is presented in Section 3. Experiments conducted on the two emotion datasets and the analysis of experimental results are illustrated in Section 4. Finally, in Section 5, we conclude this paper.

2. Related work

2.1. EEG features for emotion recognition

Traditional handcrafted features and machine learning classifiers are commonly utilized in emotion detection. However, traditional machine learning approaches are significantly constrained by feature design and selection, which often demand a substantial amount of prior knowledge. To overcome these limitations, deep learning technology was developed to learn data representations [19,20]. Inspired by the success of deep learning in speech recognition and natural language processing [21–23], some scholars have utilized deep learning to extract features from EEG signals in emotion recognition.

Time-domain features, such as amplitude, variance, and mean, are usually implemented to extract the time-domain statistics of EEG signals, which are the most intuitive and accessible features in EEG signal analysis. Frequency-domain features show how the EEG waveform changes with frequency. By using an algorithm, the time domain signals are transformed into frequency domain signals, which is the main idea behind frequency-domain features; the results reflect the way the features of the signal change with frequency, allowing the distribution of the various rhythms in EEGs to be more intuitively observed. In order to extract frequency-domain features, EEG signals are usually decomposed into several frequency bands (δ band (1–3 Hz), θ band (4–7 Hz), α band (8–13 Hz), β band (14–30 Hz), γ band (31–50 Hz)) [24,25]. Then methods such as differential entropy (DE) [25], wavelet transform (WT) [26], and power spectral density (PSD) are utilized to extract frequency-domain features from every frequency band. Time–frequency features consider the above two features of the signal, describing the ways in which it changes with both time and frequency. Methods such as STFT [27], and Hilbert–Huang transform (HHT), among others, are commonly employed to extract time–frequency features.

2.2. Domain adaptation

Creating subject-specific models for every subject is a feasible but impractical solution for addressing individual differences, as it would require a significant amount of effort to collect a labeled dataset for each subject. Another way to tackle this issue is to employ domain adaptation (DA) methods, which intend to minimize the distribution differences between different domains, thereby aiding the learning of transferable features for emotion recognition.

Wang et al. [28] categorized deep domain adaptation into discrepancy-based, adversarial-based, and reconstruction-based methods. Adversarial-based methods aim to minimize the distance between

different domains by training a domain discriminator to classify both domain types. Bao et al. [29] developed a two-level domain adaptation neural network (TDANN) that reduces the distribution discrepancy of deep features between different domains through employment of maximum mean discrepancy (MMD) and domain adversarial neural network (DANN). Chen et al. [30] proposed a multi-source EEG-based emotion recognition network (MEERNet), which adopts multiple source domains to adapt to an individual target domain separately for domain adaptation, in order to extract domain-invariant and domain-specific features. Liu et al. [31] proposed an extended domain adaptation method based on subject clustering (DASC), which mitigates the effects of "negative transfer" by incorporating subject clustering. However, although these methods can effectively reduce individual differences and improve generalization performance, the feature extraction methods they choose may ignore important information in EEG signals. TDANN selected a deep CNN method, which incorporates multiple convolutional layers and two max pooling layers, but which may lead to a lot of valuable information being lost and the relationships between the whole and the parts being ignored. MEERNet and DASC chose to use a multi-layer perceptron (MLP) method, which might cause the omission of the spatial information of EEG signals. In summary, when extracting features, these methods may neglect the relationships within and between channels as well as the temporal dimension. This also raises a second problem that needs to be solved, namely that of how to identify EEG samples that contain a higher amount of emotional information.

2.3. Multi-view learning for EEG emotion recognition

Data is often portrayed using various perspectives, incorporating multiple modalities or features [32]. In the field of medical research, Alzheimer's disease (AD) encompasses different types of data modalities, including Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET), along with diverse features [33]. Multimedia data, including text, video, and audio, is commonly sourced from various origins or described by multiple features [34,35], such as temporal and frequency features [36]. Empirical evidence from various studies [37,38] consistently indicates that multi-view learning combines different views to explore the complementarity between them, thereby improving model accuracy. Multi-view learning has gained extensive popularity and adoption across diverse tasks, owing to its notable effectiveness [39,40].

Moreover, several recent works in the field of multi-view learning have successfully incorporated deep learning techniques [41–43]. But few researchers have applied multi-view learning to emotion recognition in EEG signals. EasyDA [44] utilizes an approximate empirical kernel map generated from samples in the source and target domains to map each view into a domain generalization feature space, followed by a parameterless weighted combination for multi-view. However, this method overlooks the interrelationships between multiple channels in EEG signals and the degree of association between different channels and different emotions. Since different channels play different roles under different emotions, we consider each channel (i.e., electrode) as a view and explore the complementarity between multiple channels. We conduct emotion recognition research on EEG signals using multi-view learning based on multiple channels.

2.4. Attention mechanism

Some researchers have utilized attention mechanisms to extract emotion-related spatio-temporal features in EEG-based emotion recognition. Jia et al. [45] utilized an attention mechanism to transform channels into a 2D map when calculating weights, and used pooling to calculate an attention matrix. Li et al. [46] proposed the transferable attention neural network (TANN), which incorporates a global attention layer to merge the features from the entire brain regions and emphasize significant regions for emotion classification. However, these methods tend to neglect the interactions between EEG signal channels and temporal relevance. To address this limitation, the Transformer model, based on self-attention mechanism, assigns different weights to different time frames in the temporal dimension. This allows to fully consider the temporal dynamics in emotions. Consequently, many researchers have integrated the Transformer model into EEG-based emotion recognition tasks. Wang et al. [47] utilized weight-shared transformer encoders to adaptively capture the importance of different time frames within each channel. They also combined a hierarchical spatial encoder to capture the correlations between channels. Si et al. [48] proposed a hierarchical hybrid model called MACTN, which extracts local emotional features, global emotional features, and emotion-relevant channels through CNN, Transformer, and channel attention, respectively. Transformer models mainly focus on global temporal feature extraction, often overlooking local temporal information. On the other hand, CNN and Transformer have different abilities to extract information at various scales. Therefore, some scholars have combined CNN and Transformer to extract both local and global temporal information. However, it is worth noting that convolution and pooling operations can disrupt temporal and channel correlations.

To address this issue, we note that Hao et al. [49] opted to employ an attention mechanism, which is popularly employed in computer vision, natural language processing, and classifying multivariate time series (MTS) tasks due to its ability to locate key regions in images, key parts in sentences, and key variables in MTS. We hypothesize that this model also has the ability to locate global key information. We consider each channel as a view, and each view contains multiple time frames. Based on this, we first utilize an attention mechanism in the temporal dimension to locate key time frames. Then, through multi-view learning, we extract key information and complementary information between multiple views, forming a cross-attention mechanism based on multi-view learning. This allows the model to automatically identify key emotional-related information in multi-view EEG signals.

3. Proposed methodology

3.1. System overview

In our model, as shown in Fig. 1, features (Fig. 1(a)) containing temporal and spatial information are generated from the raw signals for each subject. These features are then input into a domain adversarial neural network in which the deep representation of the features is extracted by means of a multi-view cross-attention mechanism (Fig. 1(b)) and a dilated causal convolutional neural network (Fig. 1(d)). Finally, the fused deep representation, which is fused through a max-pooling layer, is fed to two parts, including a domain discriminator (Fig. 1(f)) and a label classifier (Fig. 1(e)). The domain discriminator is utilized to determine the domain from which the input originates (training data or testing data), in order to narrow down the distinguish shift. The label classifier assigns the deep representation to a class label within a classification space.

3.2. Input EEG signal representation

Since DE features have exhibited remarkable performance in multi-view EEG-based emotion recognition [50], we implement the proposed CADD-DCCNN method with DE features obtained from multi-view EEG signals as input. Using (1), we formulate DE:

$$\begin{aligned} f(S) &= -\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(s-\mu)^2}{2\sigma^2}} log \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(s-\mu)^2}{2\sigma^2}} ds \\ &= \frac{1}{2} log 2\pi e \sigma^2 \end{aligned} \tag{1}$$

where S is a slice of the EEG signal that obeys the Gaussian distribution $N(\mu, \sigma^2)$. Particularly, in the datasets referred to in Section 4.1, every

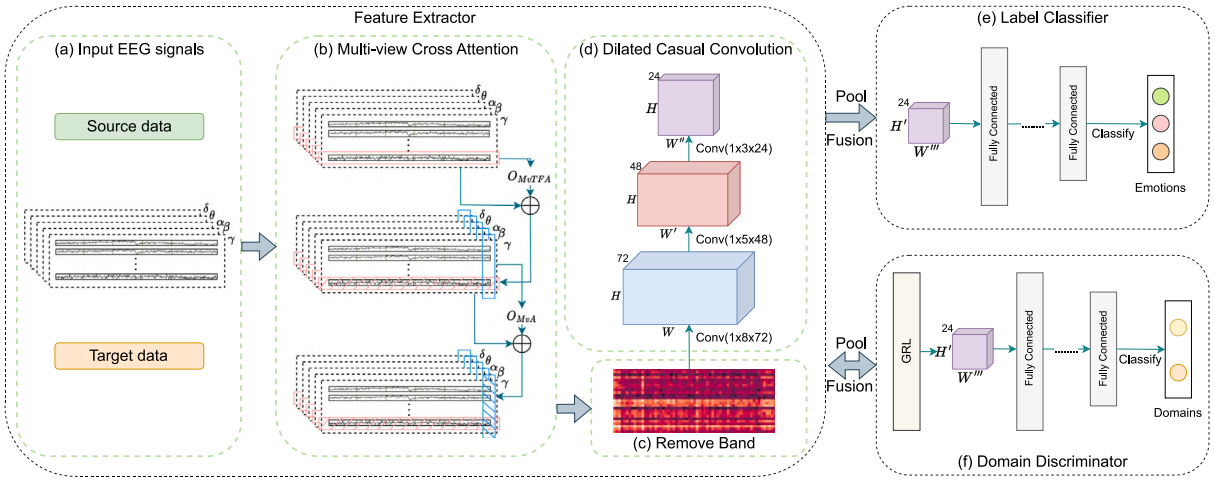


Fig. 1. The network structure of the CADD-DCCNN framework for EEG emotion recognition.

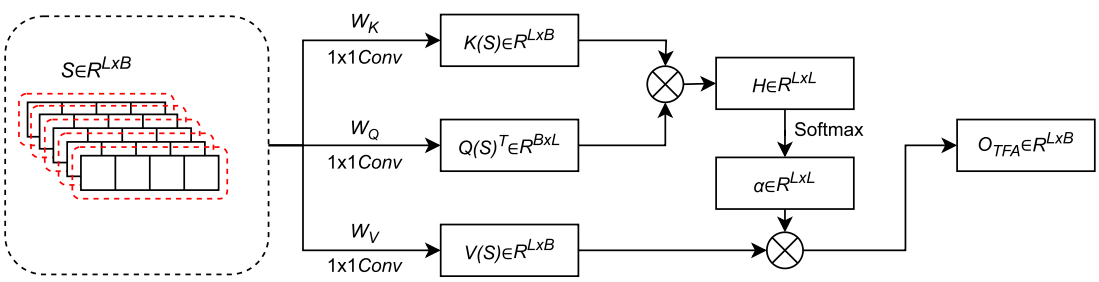


Fig. 2. Time-frame Attention mechanism for calculating the attention of one channel.

subject has multiple captured trials, and every trial consisting of a series of multi-view EEG signals of specific duration recorded while experiencing a specific emotion. For each trial, DE features are extracted from five distinct frequency bands of each channel using STFT with a non-overlapping Hanning window of 1-s. The size of DE features for 1-s windows is (62, 5) and (32, 5) for the SEED and DEAP datasets, respectively. We concatenate the DEs of all windows into a feature vector that represents one trial.

The proposed CADD-DCCNN method uses an input matrix denoted as t_m, which corresponds to the mth subject. Specifically, $t_m = [t_{m,1}, t_{m,2}, \ldots, t_{m,n}]^\top \in R^{n\times d_s}$, where d_s represents the size of a DE feature vector, and the vector $t_{m,i}$, with d_s dimensions, represents a feature associated with the ith trial of the mth subject.

3.3. Attention-based feature extractor

The feature extractor consists of a multi-view cross-attention mechanism (MvCA), a band removal mechanism, and a dilated causal convolutional neural network (DCCNN).

3.3.1. Multi-view Cross-Attention mechanism (MvCA)

The input of the feature extractor is a multivariate (time frame and channel) EEG signal sequence. Since human emotions are progressive and diverse, and the activated degrees of different emotions also differ across brain regions. Therefore, we consider each channel as a view. Additionally, since different emotions exhibit varying activation patterns across different time frames, we employee an attention mechanism within each view to calculate attention weights on the time dimension. Finally, we employ a similar attention mechanism across multiple views to dynamically learn the weights for each view, and then combines them with the original signal to build a new vector representation for the EEG trial using multi-view learning.

The MvCA module in our model comprises two modules: multi-view time-frame attention (MvTFA) and multi-view attention (MvA). In this paper, we consider a channel as a view. MvTFA extracts the long- and short-term dependencies of past values in each view of signals. MvA evaluates connections between multiple views. The MvCA module initially applies the MvTFA module. For one view, we employ $S = (s_1^1, s_1^2, \ldots, s_1^L) \ldots (s_C^1, s_C^2, \ldots, s_C^L)$ to embody the original data feature sequence, where L represents the count of time frames and C denotes the quantity of views. The specific computation process is demonstrated in Fig. 2.

As the first stage in the MvCA module, S is converted into three distinct feature domains (Q, K, and V) using (2):

$$Q(S) = S \cdot W_Q, K(S) = S \cdot W_K, V(S) = S \cdot W_V \tag{2}$$

where W_Q, W_K, and W_V are weight matrices with dimension $B\times B$. The outputs $Q(S)$, $K(S)$, and $V(S)$ share the dimension $L\times B$ and represent the query space, key space, and value space, respectively. The MvCA seizes the connections between a potential query and the key–value pairs in the data. The conversion of S into each feature space can be accomplished by utilizing a 1×1 convolutional operation on S.

During the second stage, the time-frame attention for a single view, indicated as α in Fig. 2, is computed using the features in the query and key spaces by two equals depicted in (3) and (4).

$$H = Q(S) \cdot K(S)^\top \tag{3}$$

$$\alpha_{q,k} = exp(H_{q,k}) / \sum_{j=1}^{q} exp(H_{q,k}) (1 \le k \le q \le L). \tag{4}$$

Here, $H_{q,k}$ is a hidden state within the H matrix (where $H \in R^{L\times L}$) that records the attention of features from previous time step k to current time step q. To ensure that $H_{q,k}$ is only applicable when $k \le q$, H is updated to H' through setting $H_{m,n}$ to zero when $m < n$. That is to

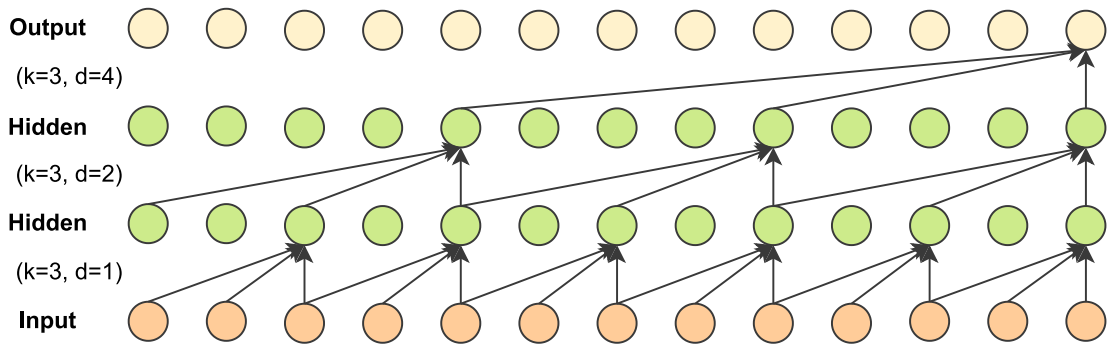


Fig. 3. A dilated causal convolution with dilation factors d = 1,2,4 and filter size k = 3.

say, the upper right corner of the H' matrix is assigned zero, resulting a lower triangular matrix. Hence, H' can directly seize the attention of both long- and short-term past values. Finally, (4) is used to normalize the attention, which produces the attention matrix α.

During the third action, α is employed on the features in $V(S)$ to compute the attention O_{MvTFA} using (5). Next, O_{MvTFA} is again merged with S to obtain the hidden states Y using (6).

$$O_{MvTFA} = \alpha \cdot V(S) \tag{5}$$

$$Y = O_{MvTFA} + S. \tag{6}$$

To our knowledge, to compute attention, current attention mechanisms either disregard the temporal order of features [45], or only consider features that already indicate the temporal order. The MvTFA mechanism in our proposed model differs from these existing attention mechanisms because it can capture the time dependency of values directly.

Apart from developing the MvTFA module to learn the long- and short-term dependencies of features in time sequence from each view, we also construct the MvA module to assess the relationships between views. The input of the MvA is a feature sequence at time step t from C views, denoted as $Y^t = (y_1^t, y_2^t, \dots, y_C^t)$. Using a similar process to that used to compute α in the MvTFA module, we can compute a normalized view attention. We then apply the view attention to Y_t and denote the output as O_{MvA}. Finally, O_{MvA} is merged with the Y once more to obtain the hidden states Z using (7).

$$Z = O_{MvA} + Y. \tag{7}$$

The ultimate features in Z, which combine both the O_{MvTFA} and O_{MvA}, are referred to as multi-view cross-attention features.

3.3.2. Band removal

The second part is a band removal mechanism. We perform a $1 \times 1 \times 1$ convolution on the features from MvCA to compress the dimension of frequency band direction and reduce the amount of subsequent computation.

3.3.3. Dilated Causal Convolutional Neural Network (DCCNN)

The third part of the feature extractor is a DCCNN. The DCCNN is formed by the combination of dilated convolution and causal convolution. Causal convolution is a convolutional model used for solving time series problems, where each node only considers preceding nodes, ensuring that information cannot flow into the future and capturing the causal relationships between time frames. Dilated convolution, on the other hand, expands the receptive field, allowing the current node to look 'very far' into the past. When the dilation factor is 1, dilated convolution is equivalent to a standard Convolutional Neural Network (CNN). Fig. 3 shows an illustration of a DCCNN. The kernal size k is 3 and the dilation factors d of three dilated layers are 1, 2, and 4, respectively. A 1D dilated convolution is calculated as shown in (8).

$$p(s) = \sum_{l=0}^{k-1} f(l) h(s - d \cdot l). \tag{8}$$

Here, $h(*)$ represents the input, $p(*)$ represents the output of the dilated convolution, $f(l)$ represents the filter of length k, and $s - d \cdot l$ denotes the historical direction of feature s.

In simple terms, the dilated convolution, without increasing the number of parameters, incorporates local information of different sizes by setting different dilation factors at different layers. Therefore, by using DCCNN, we can further extract the temporal features and causal relationships of EEG signals while reducing the number of parameters for subsequent calculations, thereby improving computational speed.

3.4. Domain Discriminator (DD)

Motivated by the generative adversarial network (GAN), we develop an adversarial training process for the feature extractor and the DD. Throughout the training process, the DD seeks to determine whether the features belong to the source or target domain. Meanwhile, the feature extractor is trained to transform the inputs from different domains into a shared latent space. By reducing the classification capability of the DD, the feature extractor is trained to produce features that are domain-independent. By adopting this approach, our proposed method can reduce the problem of feature distribution shift.

Particularly, we first use average pooling to fuse the deep representation from multi-view features and transform the input $P(S)$ to a vector d_m. Then, we introduce a gradient reversal layer (GRL), which can maximize DD loss, prior to applying ReLU activation to d_m which is calculated by (9). The GRL is not effective during forward propagation, but changes the direction of gradient transfer during backpropagation, so that the updating direction is reversed.

$$d_m^r = ReLU(W^r \cdot d_m + b^r) \tag{9}$$

$$d_m^s = softmax(W^s \cdot d_m^r + b^s). \tag{10}$$

Finally, we derive the probability of the input originating from the source or target domain by transforming d_m^r into a 2D space and implementing a softmax function in (10), where W^r, b^r, W^s, and b^s are weight matrices and bias vectors that can be learned during training.

3.5. Label classifier

The label classifier is connected to the emotion recognition task and is trained to learn the distribution of emotions to output emotion labels from deep representations of EEG signals. We first use average pooling to fuse the multi-view features to learn more discriminative representations. Then, to decode these deep representations, we use

several fully connected (FC) layers to construct a classifier. The label classifier predicts emotions by mapping the deep representations from the common space to the emotion space. Specifically, the classifier comprises three FC layers and a softmax function, which transforms the network predictions into emotions using (11).

$$d^s_{m,f} = softmax(W^s \cdot d_m + b^s). \tag{11}$$

Here, W^s and b^s are the learnable weight matrix and bias vector.

4. Experiments and results

4.1. Datasets

4.1.1. SEED

The SEED dataset is a public affective EEG dataset for emotion recognition. It contains EEG data acquired from 15 subjects, recorded via 62 EEG electrodes while they watched 15 film videos, each lasting about four minutes. The videos elicited three types of emotions (positive, neutral, and negative). Each subject participated in the experiments three times on different days, watching the same 15 movie videos within each experiment; these settings allow for subject-dependent (SD) and subject-independent (SI) experiments to validate the robustness and transferability of emotion recognition models. In SEED, a bandpass filter ranging from 1.0 to 75.0 Hz in frequency was employed. Then the DE features were extracted. Since the 15 trials undertaken by each subject were of different durations, we performed a zero-filling operation: specifically, we selected the longest duration of the trial as the final duration, then performed zero-filling operations after other trials to unify the trial length.

4.1.2. DEAP

The DEAP dataset is also a public affective EEG dataset for emotion recognition, containing the collected data of 32 subjects watching 40 one-minute music videos. Participants rate their levels from 1 to 9 after watching each video on four dimensions: Arousal, Valence, Liking, and Dominance. In this paper, we removed the eight peripheral channels and used only EEG signals for emotion recognition. We split the valence and the arousal dimension into high/low, separately, resulting in two binary classification task. In DEAP, a bandpass filter ranging from 4.0 to 45.0 Hz in frequency was used, as reported in [51]. Initially, we disassemble the EEG signals into the same five frequency bands. Next, DE features are extracted from each channel for every frequency band.

4.2. Experimental settings

The feature extractor of our model includes a multi-view cross-attention, a band removal, and a DCCNN. The DCCNN includes three convolutional layers with kernels of (1, 8), (1, 5), and (1, 3), respectively. The dilation factors for the second and third layers are (1, 8) and (1, 5). As DCCNN only works in the time dimension, the convolutional kernel and dilated factor are set to 1 in the channel dimension. The output tensor of the three convolutional layers has 72, 48, and 24 channels, respectively. The label classifier consists of three fully connected layers, with outputs of size 1024, 150, and 3 (for the SEED dataset) or 2 (for the DEAP dataset). During the training process, we utilized a batch size of 5 for 200 epochs and employed the Adam optimizer with a learning rate of 1e−4.

4.3. Compared methods

Within this part, we exhibit experimental results on two commonly adopted EEG datasets (as outlined above, SEED and DEAP) and compare our proposed CADD-DCCNN with several other methods, which are listed below:

- Two traditional shallow machine learning techniques: SVM [52] and kNN [53];
- Three deep neural network models: DGCNN [50], SparseD [54], and MATCN [48];
- Seven cutting-edge domain adaptation models for emotion recognition in EEG: ATDD-LSTM [51], MEERNet [30], MS-MDA [55], AD-TCNs [56], HVF$_2$N-DBR [57], MMDA-VAE [58], MSDA-SFE [59], TMLP+SRDANN [60], and TSFIN [61].

While SVM and kNN can only process EEG signal channels one at a time, other deep networks are capable of handling multi-channel EEG signals. These methods are all characteristic approaches in prior research studies on emotion recognition from EEG signals. To ensure a persuasive comparison with the proposed approach, we directly quoted the outcomes from the relevant literature. Additionally, we have conducted ablation studies to investigate the roles of each part in CADD-DCCNN and their influence on the whole method.

4.4. Experiment on two publicly available datasets

4.4.1. Experiment on SEED dataset

SD Experiment. Compared to the DEAP dataset, each subject in SEED participated in three trials at different time, resulting in data consisting of three sessions. We utilized this characteristic to design a SD cross-session trial that forecasts the emotions from the same subject at different time. The results illustrate the robustness of different models over time, making this experiment setting more useful for practical applications. However, few research works have conducted cross-session experiments until now. In the cross-session scenario, we used leave-one-session-out cross-validation. In particular, two out of three sessions of each subject were utilized as training data, whereas the unused session was employed to test. The classification accuracy for each subject was computed as the mean accuracy across three trials. Finally, the mean classification accuracy (ACC) and standard deviation (STD) across 15 subjects were calculated as the final result of cross-session experiment.

SI Experiment. We used leave-one-subject-out cross-validation (LOSO-CV), where 14 subjects were utilized as training data and the unused subject was employed to test. We computed the ACC and STD across 15 subjects as the final results for cross-subject experiment.

4.4.2. Experiment on DEAP dataset

To perform a binary classification task, we categorized the emotions in the DEAP dataset as high/low arousal and valence, using the same partition scheme and threshold as described in [18].

SD Experiment. We applied leave-one-clip-out cross-validation, where 39 out of 40 trials of one subject were utilized as training data and the unused trial was employed to test. An ACC and STD were computed as the final result of SD experiment over 40 trials for each of the 32 subjects. In the DEAP dataset, each trial of each subject contains only one data point; however, in this experiment, using one trial for testing would result in an insufficient amount of data. Therefore, we applied sliding windows with a size of 9 s and no overlap to split the data into multiple segments, which divided the data into 7 data points for one trial. Thus, the training data was made up of 273 ($39 * 7$) data points, and the testing data contained 7 ($1 * 7$) data points for one subject.

SI Experiment. Similar to our approach on SEED, we applied LOSO-CV for the current experiment. This means that 31 out of 32 subjects were utilized as training data, and the unused subject was employed to test. We compute an ACC and STD across 32 subjects as the final result of SI experiment.

Table 1
Mean accuracies (%) and STD for SD achieved by different methods on SEED dataset.

Model	ACC/STD
kNN [53]	72.00/12.60
DGCNN [50]	73.06/10.36
SVM [52]	81.19/14.79
MEERNet [30]	86.20/05.80
CADD-DCCNN	**87.41/01.80**

Table 2
Mean accuracies (%) and STD for SI achieved by different methods on SEED dataset.

Model	ACC/STD
SVM [52]	56.73/16.29
kNN [53]	73.93/09.95
DGCNN [50]	79.95/09.02
TMLP+SRDANN [60]	81.04/06.28
MMDA-VAE [58]	85.07/11.81
MEERNet [30]	87.10/02.00
HVF_2N-DBR [57]	89.33/10.13
MS-MDA [55]	89.63/06.79
MSDA-SFE [59]	91.65/02.91
CADD-DCCNN	**92.44/06.16**

Table 3
Mean accuracies (%) and STD for SD achieved by different methods on DEAP dataset.

Model	ACC/STD	
	Valence	Arousal
kNN [53]	50.11/21.92	57.76/23.14
SVM [52]	53.76/19.56	55.67/20.91
DGCNN [50]	86.06/02.61	85.61/02.44
ATDD-LSTM [51]	90.91/12.95	90.87/11.32
SparseD [54]	**95.72/09.52**	91.75/05.23
CADD-DCCNN	90.97/13.96	**92.42/12.72**

Table 4
Mean accuracies (%) and STD for SI achieved by different methods on DEAP dataset.

Model	ACC/STD	
	Valence	Arousal
kNN [53]	54.38/09.27	54.38/11.93
SVM [52]	55.62/09.44	52.66/14.47
TMLP+SRDANN [60]	57.70/07.23	61.88/05.55
DGCNN [50]	59.29/06.83	61.10/12.28
SparseD [54]	60.65/06.24	65.39/09.41
AD-TCNs [56]	64.33/07.06	63.25/04.62
MATCN [48]	66.10/06.10	67.80/08.10
TSFIN [61]	67.03/–	68.13/–
HVF_2N-DBR [57]	68.91/–	69.22/–
MSDA-SFE [59]	69.26/–	70.10/–
CADD-DCCNN	**69.45/05.60**	**70.50/09.39**

4.4.3. Results and analysis

The results of our experiments are summarized in Tables 1 to 4. Bold indicates the highest average accuracy. Based on the experimental results, we made three key observations:

(1) Our proposed CADD-DCCNN method outperformed all comparable methods on both the SEED and DEAP datasets. Specifically, compared to non-domain adaptation methods such as SVM, kNN and DGCNN, the average precision improvement of CADD-DCCNN was approximately 20.97%, 16.96% and 13.42% on the SEED dataset. On the DEAP dataset, the average precision increased by around 25.52%, 27.97%, 8.54%, 2.03%, and 3.35% for valence and 27.30%, 25.39%, 8.11%, 2.89%, and 2.7% for arousal when compared to SVM, kNN, DGCNN, SparseD, and MATCN, respectively. These results validate the effectiveness of our learned transferable data representation for EEG-based emotion recognition and demonstrate the practicality of the domain adaptation method in cross-subject EEG-based emotion recognition.

(2) Our proposed CADD-DCCNN method outperformed existing domain adaptation methods. Compared to models trained using domain adaptation learning strategies, such as ATDD-LSTM, MEERNet, MS-MDA, AD-TCNs, HVF_2N-DBR, MMDA-VAE, MSDA-SFE, TMLP+SRDANN, and TSFIN, the CADD-DCCNN achieved an average precision improvement of 5.14%, along with improvements of 4.00% for valence and 3.98% for arousal, in SI scenarios on the SEED and DEAP datasets, respectively. These results indicate that our proposed multi-view cross-attention mechanism can effectively explore the complementary and consistent information among channels. Furthermore, the use of attention mechanisms within each view enables effective learning of more discriminative temporal information.

(3) On the DEAP dataset, we observed that the model performance was much lower in SI scenarios than in SD scenarios, which demonstrates that individual differences have a negative impact on multi-view EEG-based emotion recognition. In cross-session scenarios on the SEED dataset, we found that differences between different testing times for the same subject are even harder to eliminate than individual differences between different subjects, resulting in lower precision in SD scenarios than in SI scenarios. However, despite these challenges, our method still outperformed the compared methods. Therefore, overall, our proposed CADD-DCCNN method has an advantage in minimizing both intra-individual and inter-individual differences.

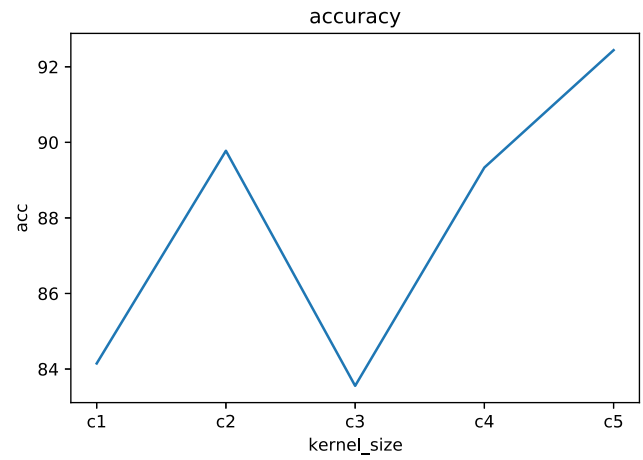


Fig. 4. Line chart of average accuracy of CADD-DCCNN under different combinations of kernel size in DCCNN. c1 represents the kernel sizes in the temporal dimension as (3, 3, 3). c2 represents the kernel sizes in the temporal dimension as (5, 5, 5). c3 represents the kernel sizes in the temporal dimension as (8, 8, 8). c4 represents the kernel sizes in the temporal dimension as (3, 5, 8). c5 represents the kernel sizes in the temporal dimension as (8, 5, 3). Among them, c5 is the kernel sizes used in our model.

Table 5
Mean accuracies (%) and STD for SI achieved in an ablation study on SEED dataset.

Model	ACC/STD
DANN	88.15/07.99
DCCNN-DANN	89.48/07.34
TFA-DANN	88.44/06.89
MvCA-DANN	90.37/07.11
CADD-DCCNN	**92.44/06.16**

4.5. Ablation study and visualization

4.5.1. Ablation study

We conducted an ablation study on the SEED and DEAP datasets to evaluate the efficiency of each part in our CADD-DCCNN approach.

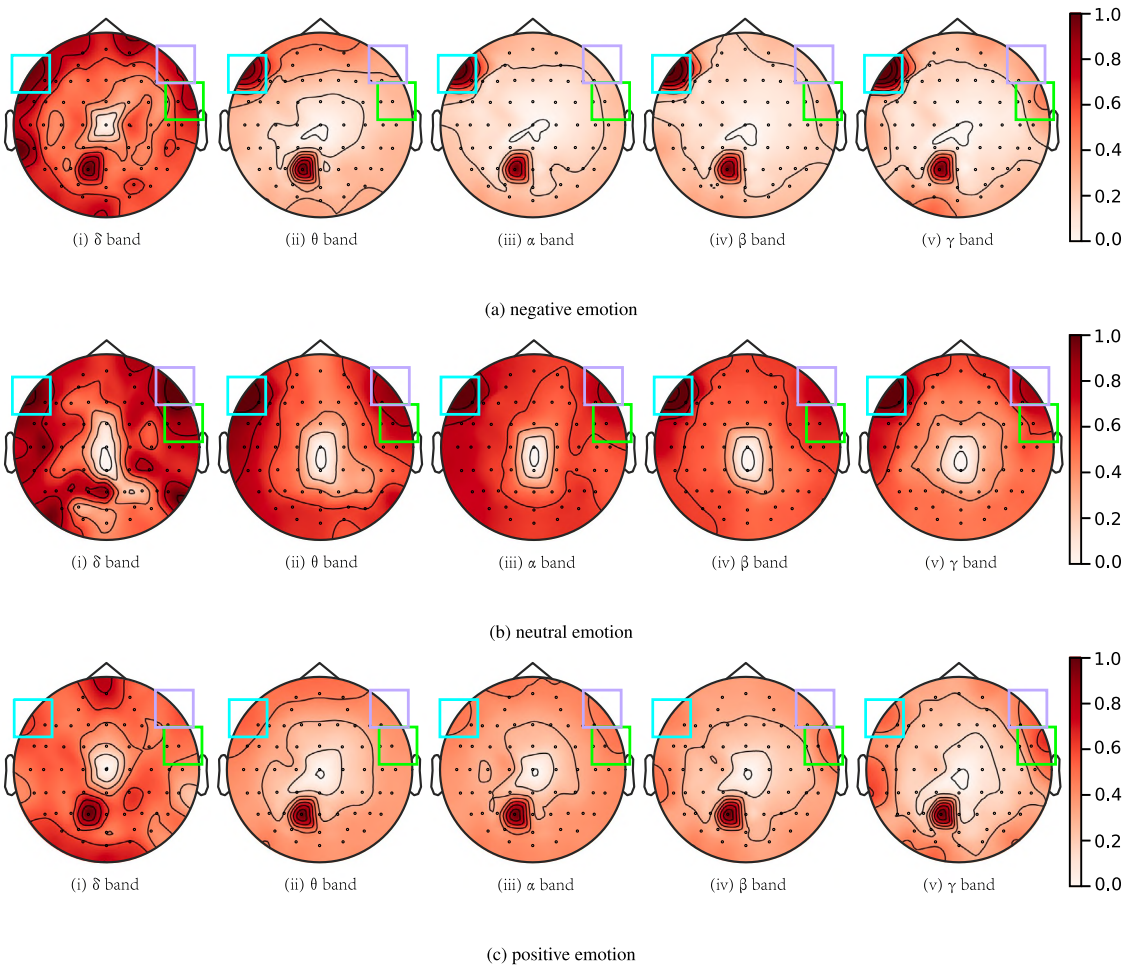


Fig. 5. Brain topography maps generated from raw EEG signals under different emotional states.

Table 6
Mean accuracies (%) and STD for SI achieved in an ablation study on DEAP dataset.

Model	ACC/STD	
	Valence	Arousal
DANN	66.95/07.48	68.67/10.36
DCCNN-DANN	67.19/07.29	69.14/09.79
TFA-DANN	67.50/05.86	68.79/10.20
MvCA-DANN	68.91/06.38	69.45/09.43
CADD-DCCNN	**69.45/05.60**	**70.50/09.39**

CADD-DCCNN was constructed based on DANN, with three supplementary modules: a time-frame attention, a multi-view cross-attention mechanism and a DCCNN. In particular, we compared our full CADD-DCCNN method with four alternate methods:

- DANN: only DANN model without two supplementary modules;
- DCCNN-DANN: the DANN model with only the DCCNN;
- TFA-DANN: the DANN model with only the time-frame attention mechanism;
- MvCA-DANN: the DANN model with only the multi-view cross-attention mechanism.

The outcomes of our ablation study are presented in Tables 5 and 6. Our model achieves, by far, the best recognition performance. Furthermore, DANN's recognition accuracy is 2.87% lower than CADD-DCCNN's, showing the effectiveness of DCCNN and multi-view cross-attention mechanism. The accuracy of DCCNN-DANN is 0.68% higher than DANN, demonstrating the positive role of DCCNN in extracting contextual relationships in multi-channel EEG signals during the improvement of this algorithm. The accuracy of TFA-DANN is 0.32% higher than DANN, validating that the same channel exhibits different emotional responses across different time frames. By employing an attention mechanism to assign different weights to different time frames, more influential emotion-related features can be extracted. The accuracy of MvCA-DANN is 1.86% higher than TFA-DANN, indicating that the information between multiple channels is complementary. Learning the complementary information among multiple channels can effectively improve the performance of the model, further validating the effectiveness of multi-view learning. Drawing on these findings, we believe that the proposed CADD-DCCNN method is beneficial for improving EEG-based emotion recognition performance.

4.5.2. Impact of kernel size in DCCNN

Fig. 4 shows the relationship between different convolutional kernels and the performance of CADD-DCCNN. DCCNN increases the receptive field by introducing dilations in the convolutional kernel. Therefore, using different kernel sizes can capture features at different scales. Larger kernels can capture global contextual information, while smaller kernels can extract local detailed information. As shown in the figure, using the same kernel size in all three layers cannot simultaneously capture global and local information. By using different kernel sizes and combining features at different levels, the model can better capture multi-scale sequential patterns, enhancing its modeling capability. Additionally, placing larger kernels at earlier layers

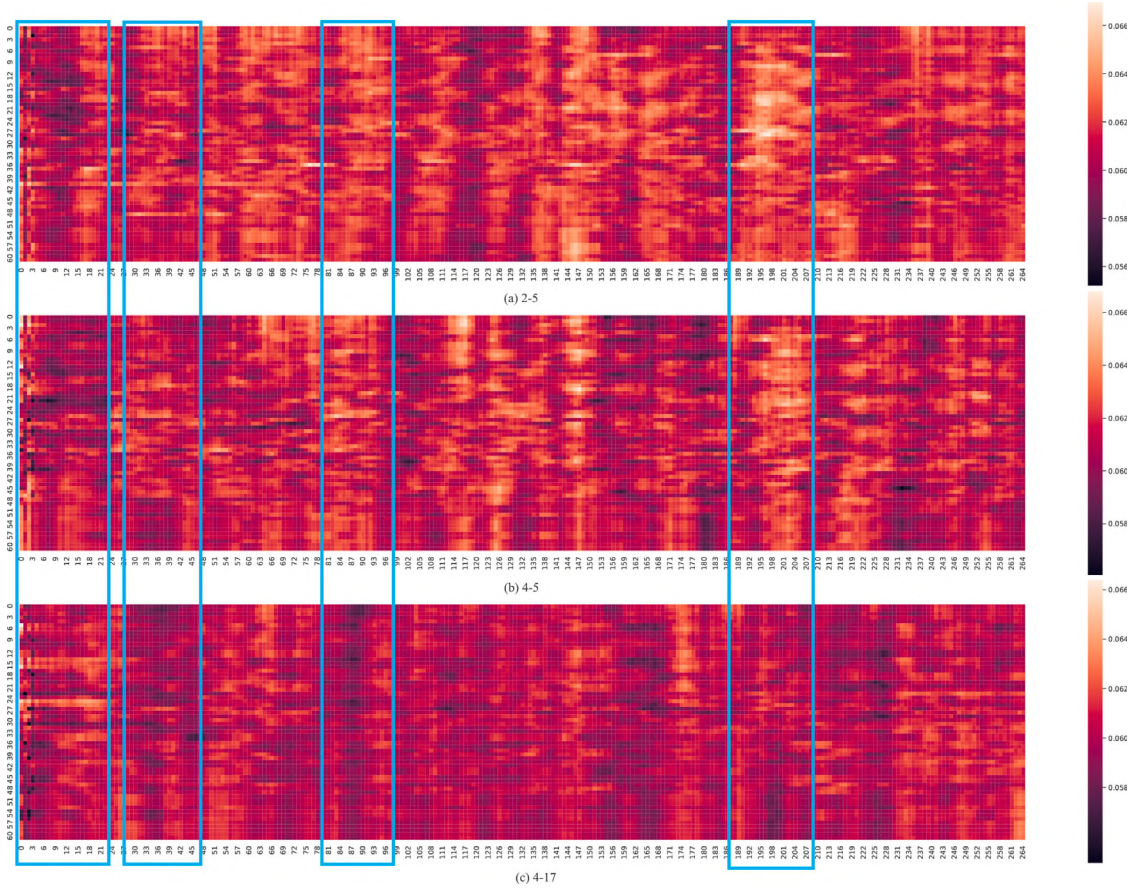


Fig. 6. (a) The 5th trial from the 1st session of the 2nd subject, (b) The 5th trial from the 1st session of the 4th subject, (c) The 2nd trial from the 2nd session of the 4th subject.

allows the model to capture global contextual information first when processing input sequences, helping to better understand the overall sequence structure. Furthermore, gradually decreasing kernel sizes can reduce the total number of model parameters. Therefore, we chose the combination of c5.

4.5.3. Validation of model components

To further comprehend and demonstrate the impacts of multi-view learning, multi-view cross-attention (MvCA), DCCNN, and the domain discriminator in EEG-based emotion recognition, we utilized two visualization techniques and a set of comparative experiments. Specifically, we employed EEG topographic maps to visualize the spatial distribution of electrical activity in different brain regions under various conditions, which enabled us to observe how different patterns of brain activity were associated with different emotions in the presence of multi-view cross-attention. We conducted comparative experiments by combining DCCNN and standard CNN with the baseline model DANN to validate the effectiveness of DCCNN. Additionally, we used 2-dimensional t-SNE plots to analyze the shift in feature distribution between training and testing data, as well as to evaluate the performance of the domain discriminator.

As shown in Fig. 5, to verify the usability of multi-view learning, we display the activation degrees of all channels at the same moment under different emotions. In negative and neutral emotions, the activation levels are relatively higher in the blue regions, while in positive emotions, the activation levels are relatively lower. In negative emotions, the activation levels are relatively lower in the green regions, while in positive and neutral emotions, the opposite is observed. In neutral emotions, the activation levels are relatively higher in the purple regions, while in positive and negative emotions, the opposite is observed. This indicates that the activation degrees of different channels vary significantly under different emotions. Therefore, we can consider each channel as a view and explore the complementarity between multiple views to learn emotion-related features more comprehensively, further enhancing the model's generalization ability.

To analyze the effectiveness of MvCA, we plotted heatmaps using the multi-view time-frame attention (MvTFA) and EEG topographic maps after applying MvCA. We selected key channels based on the EEG topographic map and conducted experiments after removing those key channels. First, we plotted heatmaps of the same experiment for different subjects and different trials for the same subject after applying MvTFA, as shown in Fig. 6.(a) and (b) represent the same trial for different subjects, while (b) and (c) represent different trials for the same subject. The higher the activation level, the lighter the color. From the figure, it can be observed that during the time intervals of 30 s to 45 s, 81 s to 96 s, and 189 s to 207 s, (a) and (b) exhibit relatively higher activation levels, while (c) shows relatively lower activation levels. However, during the time interval of 0 s to 21 s, the activation levels are reversed, indicating that the activation levels in the temporal dimension are similar for the same trial, while they differ for different trials. Therefore, we can extract key time frames using MvTFA.

Next, we plotted topographic maps of EEG signals after applying MvCA (as shown in Fig. 7). Fig. 7 depicts a time frame of EEG data from the 7th subject in the SEED dataset. (a), (b), and (c), respectively, show the activation of brain regions in the five frequency bands under negative, neutral, and positive emotions. Darker colors indicate higher activation levels and greater attention weights. From Fig. 7, we can observe that during negative emotion, the pre-frontal and frontal lobes exhibit high activation levels, and the right brain region on the β and

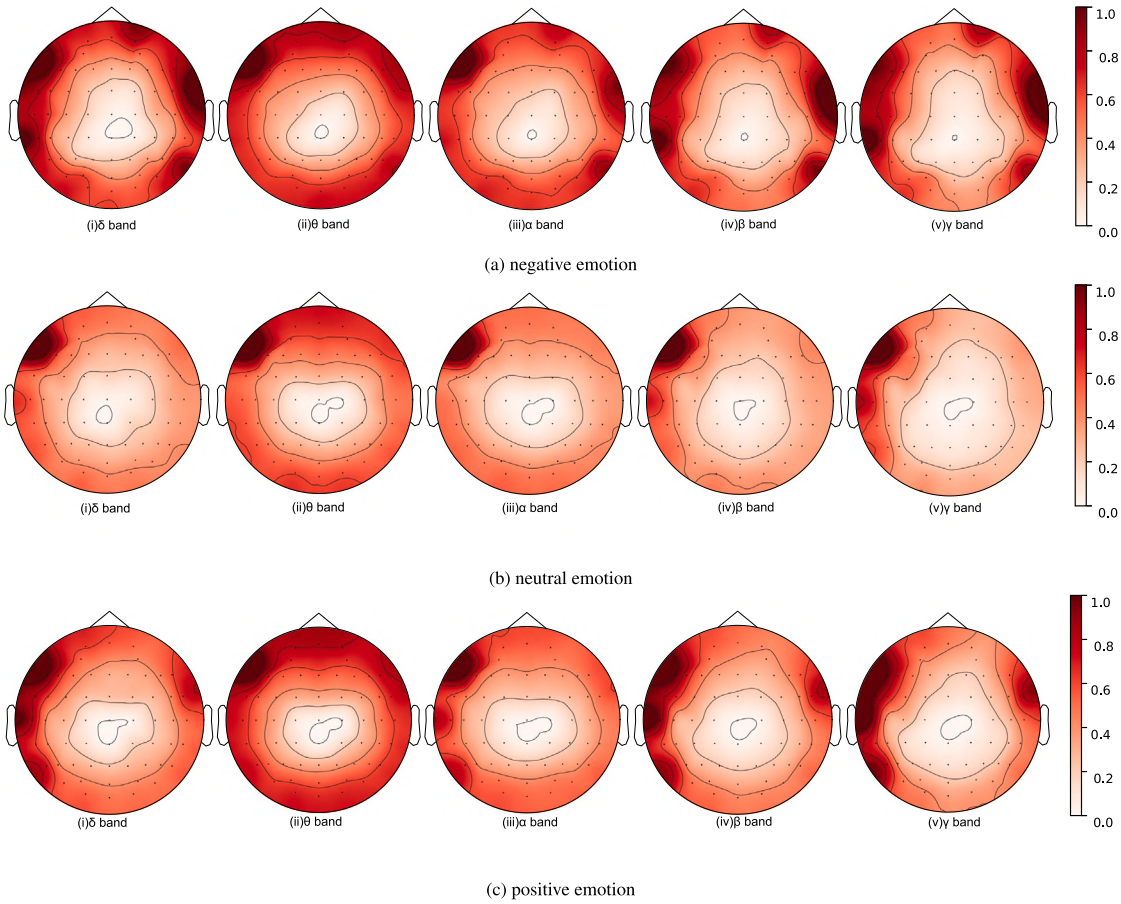


Fig. 7. EEG topographic maps under various emotions on SEED.

Table 7
Channels of EEG signals, mean accuracies (%), and STD for SI achieved on SEED dataset.

Parameters	All-Channels	No-Key-Channels
Channels	62	57
ACC/STD	**92.44/06.16**	78.52/09.67

Table 8
Parameters of DCCNN-DANN and CNN-DANN, mean accuracies (%), and STD for SI achieved on SEED dataset.

Parameters	DCCNN-DANN	CNN-DANN
Convolution layers	3	3
Kernel size	(1, 8), (1, 5), (1, 3)	(1, 8), (1, 5), (1, 3)
Dilation factors	(1, 8), (1, 5)	None
Output tensors	72, 48, 24	72, 48, 24
ACC/STD	**89.48/07.34**	88.30/07.00

γ bands also shows high activity. During neutral emotion, the prefrontal cortex is more active. During positive emotion, the left brain region exhibits high activity. These findings are consistent with those of neuroscience studies [62,63].

Based on Fig. 7, we selected the top 5 channels with higher activation levels, namely FP2, F7, F5, T7, and P7. We conducted experiments after removing these 5 channels, and the number of channels and results are shown in Table 7. After removing the key channels, the accuracy decreased by 13.92%, demonstrating that MvCA can indeed extract key channels, thus validating the effectiveness of MvCA.

We conducted a series of comparative experiments to assess the effectiveness of DCCNN. In our study, we integrated DCCNN and standard CNN into the baseline DANN model for EEG-based emotion recognition. The parameters and experimental results are shown in Table 8. The accuracy of CNN-DANN is 88.30%, which is 0.15% higher than the DANN model but 1.18% lower than DCCNN-DANN. These findings suggest that the CNN model is proficient at extracting emotion-related features. However, when employing multi-layer convolution operations in CNN, the feature map size tends to decrease and the receptive field size also decreases. In contrast, DCCNN utilizes dilated convolution, which not only preserves the feature map size but also expands the receptive field through dilation factors. Consequently, DCCNN effectively captures temporal features at various scales, enabling the extraction of both local and global temporal information.

Finally, to show the efficiency of the domain discriminator, we display the training and testing feature distributions in the same 2D space before and after passing through the domain discriminator, respectively (see Fig. 8). Each colored shape stands for a subject, i. e., a domain. Fig. 8 displays the feature distribution of each subject in SEED and DEAP before and after passing through the domain discriminator. As depicted in (a) and (c), the allocation of EEG data among various subjects (represented by distinct colors) is similar, with the majority of trials clustering together and only a small number of outliers occurring in certain subjects. However, after passing through the domain discriminator, the feature distribution of each subject becomes more uniform.

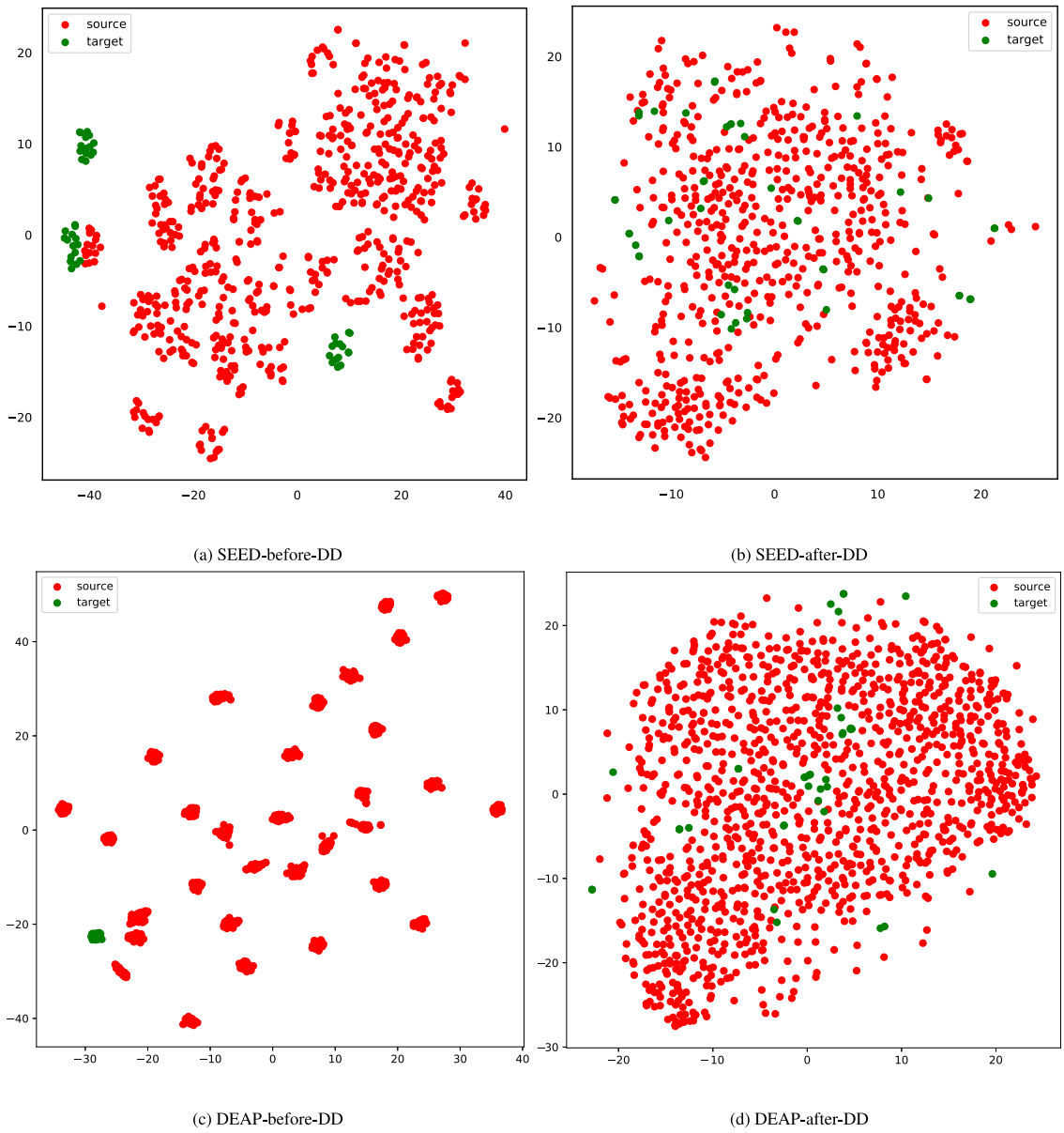


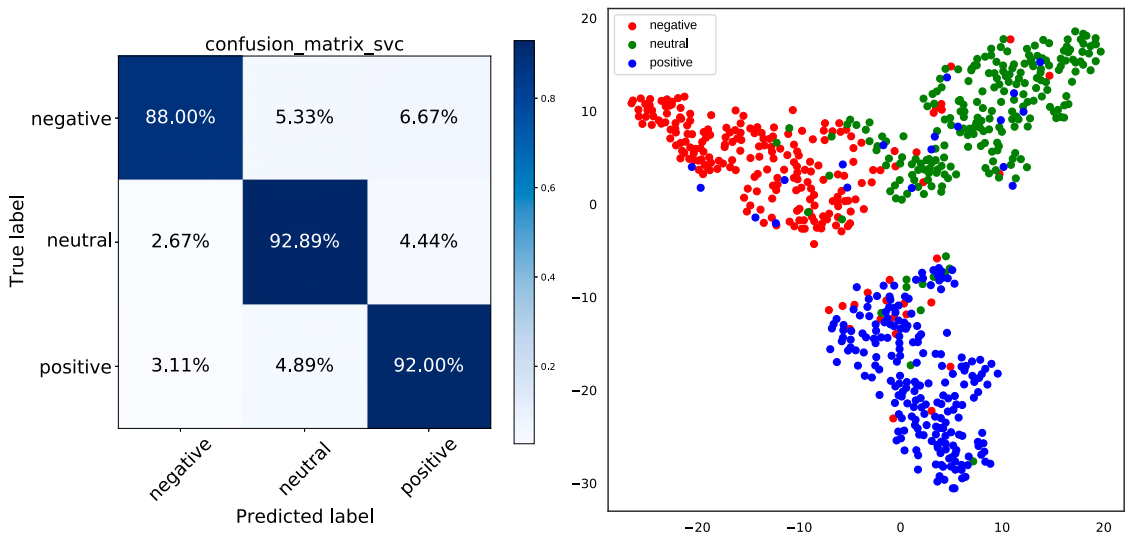
Fig. 8. T-SNE-based visualization of feature embedding on every subject of SEED and DEAP. (a) displays the feature space on the SEED of using CADD-DCCNN before the domain discriminator (DD). (b) displays the feature space on the SEED of using CADD-DCCNN after the domain discriminator.

Therefore, the inclusion of the domain discriminator into our model allows us to ensure data representation invariance while minimizing significant feature distribution shifts between different subjects.

Fig. 9 presents the outcomes of our experiments on SEED. Specifically, the figure includes a confusion matrix (Fig. 9(a)) displaying percentages with row normalization, where the horizontal axis represents the true label and the vertical axis stands for the predicted label. The element (i, j) represents the percentage of samples in class i that were classified as class j, with the matrix block on the diagonal line indicating the probability of correct prediction. Additionally, the figure includes a feature map (Fig. 9(b)) in which red points represent true labels for negative emotion, green points represent true labels for neutral emotion, and blue points correspond to true labels for positive emotion. From the results presented in Fig. 9, it is clear that our model achieves high average accuracies for recognition of all emotions in general. However, the recognition accuracy for negative emotions is relatively poor when compared to that for neutral and positive emotions. In the experiments, negative emotions were more prone to be misclassified as positive emotions. One possible reason is that the EEG signals activated during negative and positive emotions bear similarities. Subjects may exhibit strong responses to both types of emotions. Additionally, the dataset may have a relatively smaller number of samples for negative emotions, leading to misclassification of negative emotions as positive emotions.

5. Discussion

In this paper, we proposed the CADD-DCCNN method for emotion recognition from EEG signals, achieving SOTA outcomes on the popular SEED and DEAP datasets. The success of our method can be largely attributed to its utilization of multi-view learning combined with attention mechanisms to effectively select emotion-related channels and



(a) The confusion matrix

(b) The feature maps

Fig. 9. The confusion matrix and feature maps of the SI EEG emotion recognition results using our CADD-DCCNN on the SEED dataset.

time frames. Additionally, the model leverages the ability of dilated causal convolutional neural networks to extract temporal information from multi-view features. Moreover, feature-level fusion is performed on the features from multiple views, enabling the exploration of complementary information between different views. Furthermore, a domain discriminator is incorporated to ensure uniform feature distribution coverage and invariant data representation, thus minimizing the problem of data distribution shift in cross-subject scenarios and improving model generalization. Finally, we performed an ablation study to evaluate the individual contributions of each part, and the experimental outcomes confirmed their validity. Future efforts will need to verify the effectiveness in real-world 'in-the-wild' settings.

CRediT authorship contribution statement

Chao Li: Conceptualization, Data curation, Writing – review & editing, Supervision, Funding acquisition. **Ning Bian:** Methodology, Writing – original draft, Software, Visualization, Funding acquisition. **Ziping Zhao:** Writing – review & editing, Supervision, Validation, Funding acquisition. **Haishuai Wang:** Validation, Supervision, Funding acquisition. **Björn W. Schuller:** Supervision, Writing – review & editing, Revised paper, Funding acquisition.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

Data availability

Data will be made available on request.

Acknowledgments

This work was substantially supported by the National Natural Science Foundation of China (Grant Nos: 62071330, 61702370, 61902282), the National Science Fund for Distinguished Young Scholars (Grant No: 61425017), the Key Program of the National Natural Science Foundation of China (Grant No: 61831022), the Key Program of the Natural Science Foundation of Tianjin (Grant No: 18JCZDJC36300), the Technology Plan of Tianjin (Grant No: 18ZXRHSY00100), and the Tianjin Research Innovation Project for Postgraduate Students (Grant No: 2022SKYZ267).

References

[1] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion (2022).

[2] U.R. Acharya, V.K. Sudarshan, H. Adeli, J. Santhosh, J.E. Koh, S.D. Puthankatti, A. Adeli, A novel depression diagnosis index using nonlinear features in EEG signals, Eur. Neurol. 74 (1–2) (2015) 79–83.

[3] N. Kumar, J. Kumar, Measurement of cognitive load in HCI systems using EEG power spectrum: an experimental study, Procedia Comput. Sci. 84 (2016) 70–78.

[4] G. Recio, A. Schacht, W. Sommer, Recognizing dynamic facial expressions of emotion: Specificity and intensity effects in event-related brain potentials, Biol. Psychol. 96 (2014) 111–125.

[5] H. Gunes, M. Piccardi, Bi-modal emotion recognition from expressive face and body gestures, J. Netw. Comput. Appl. 30 (4) (2007) 1334–1345.

[6] Z. Zhao, Q. Li, Z. Zhang, N. Cummins, H. Wang, J. Tao, B.W. Schuller, Combining a parallel 2D CNN with a self-attention dilated residual network for CTC-based discrete speech emotion recognition, Neural Netw. 141 (2021) 52–60.

[7] F. Agrafioti, D. Hatzinakos, A.K. Anderson, ECG pattern analysis for emotion detection, IEEE Trans. Affect. Comput. 3 (1) (2011) 102–115.

[8] C. Li, Z. Bao, L. Li, Z. Zhao, Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition, Inf. Process. Manage. 57 (3) (2020) 102185.

[9] B. Cheng, G. Liu, Emotion recognition from surface EMG signal using wavelet transform and neural network, in: 2008 2nd International Conference on Bioinformatics and Biomedical Engineering, IEEE, 2008, pp. 1363–1366.

[10] M. Hamada, B. Zaidan, A. Zaidan, A systematic review for human EEG brain signals based emotion classification, feature extraction, brain condition, group comparison, J. Med. Syst. 42 (2018) 1–25.

[11] M. Li, H. Xu, X. Liu, S. Lu, Emotion recognition from multichannel EEG signals using K-nearest neighbor classification, Technol. Health Care 26 (S1) (2018) 509–519.

[12] S. Liu, Y. Zhao, Y. An, J. Zhao, S.-H. Wang, J. Yan, GLFANet: A global to local feature aggregation network for EEG emotion recognition, Biomed. Signal Process. Control. 85 (2023) 104799.

[13] S. Liu, Z. Wang, Y. An, J. Zhao, Y. Zhao, Y. dong Zhang, EEG emotion recognition based on the attention mechanism and pre-trained convolution capsule network, Knowl.-Based Syst. 265 (2023) 110372.

[14] S. Tripathi, S. Acharya, R.D. Sharma, S. Mittal, S. Bhattacharya, Using deep and convolutional neural networks for accurate emotion classification on deap dataset., in: Twenty-Ninth IAAI Conference, 2017.

[15] Y. Li, W. Zheng, Z. Cui, T. Zhang, Y. Zong, A novel neural network model based on cerebral hemispheric asymmetry for EEG emotion recognition., in: IJCAI, 2018, pp. 1561–1567.

[16] P. Pandey, K. Seeja, Subject-independent emotion detection from EEG signals using deep neural network, in: International Conference on Innovative Computing and Communications, Springer, 2019, pp. 41–46.

[17] W.-L. Zheng, B.-L. Lu, Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks, IEEE Trans. Auton. Ment. Dev. 7 (3) (2015) 162–175.

[18] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis; using physiological signals, IEEE Trans. Affect. Comput. 3 (1) (2011) 18–31.

[19] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[20] K. Ezzameli, H. Mahersia, Emotion recognition from unimodal to multimodal analysis: A review, Inf. Fusion (2023) 101847.

[21] M.M. Lopez, J. Kalita, Deep learning applied to NLP, 2017, arXiv preprint arXiv:1703.03091.

[22] S. Dara, P. Tumma, Feature extraction by using deep learning: A survey, in: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, 2018, pp. 1795–1801.

[23] K. Noda, Y. Yamaguchi, K. Nakadai, H.G. Okuno, T. Ogata, Audio-visual speech recognition using deep learning, Appl. Intell. 42 (2015) 722–737.

[24] D. Nie, X.-W. Wang, L.-C. Shi, B.-L. Lu, EEG-based emotion recognition during watching movies, in: 2011 5th International IEEE/EMBS Conference on Neural Engineering, IEEE, 2011, pp. 667–670.

[25] L.-C. Shi, Y.-Y. Jiao, B.-L. Lu, Differential entropy feature for EEG-based vigilance estimation, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2013, pp. 6627–6630.

[26] S.M. Alarcao, M.J. Fonseca, Emotions recognition using EEG signals: A survey, IEEE Trans. Affect. Comput. 10 (3) (2017) 374–393.

[27] Y.-J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, Y. Shi, Real-time movie-induced discrete emotion recognition from EEG signals, IEEE Trans. Affect. Comput. 9 (4) (2017) 550–562.

[28] M. Wang, W. Deng, Deep visual domain adaptation: A survey, Neurocomputing 312 (2018) 135–153.

[29] G. Bao, N. Zhuang, L. Tong, B. Yan, J. Shu, L. Wang, Y. Zeng, Z. Shen, Two-level domain adaptation neural network for EEG-based emotion recognition, Front. Hum. Neurosci. 14 (2021) 605246.

[30] H. Chen, Z. Li, M. Jin, J. Li, Meernet: multi-source EEG-based emotion recognition network for generalization across subjects and sessions, in: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2021, pp. 6094–6097.

[31] J. Liu, X. Shen, S. Song, D. Zhang, Domain adaptation for cross-subject emotion recognition by subject clustering, in: 2021 10th International IEEE/EMBS Conference on Neural Engineering (NER), IEEE, 2021, pp. 904–908.

[32] C. Zhang, Y. Cui, Z. Han, J.T. Zhou, H. Fu, Q. Hu, Deep partial multi-view learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (5) (2020) 2402–2415.

[33] Y. Liu, L. Fan, C. Zhang, T. Zhou, Z. Xiao, L. Geng, D. Shen, Incomplete multi-modal representation learning for Alzheimer's disease diagnosis, Med. Image Anal. 69 (2021) 101953.

[34] C. Zhang, H. Fu, J. Wang, W. Li, X. Cao, Q. Hu, Tensorized multi-view subspace representation learning, Int. J. Comput. Vis. 128 (8–9) (2020) 2344–2361.

[35] Z. Li, C. Tang, X. Liu, X. Zheng, W. Zhang, E. Zhu, Consensus graph learning for multi-view clustering, IEEE Trans. Multimed. 24 (2021) 2461–2472.

[36] C. Tang, X. Liu, X. Zhu, E. Zhu, Z. Luo, L. Wang, W. Gao, CGD: Multi-view clustering via cross-view graph diffusion, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 5924–5931.

[37] Z. Tao, H. Liu, S. Li, Z. Ding, Y. Fu, Marginalized multiview ensemble clustering, IEEE Trans. Neural Netw. Learn. Syst. 31 (2) (2019) 600–611.

[38] C. Tang, X. Zheng, W. Zhang, X. Liu, X. Zhu, E. Zhu, Unsupervised feature selection via multiple graph fusion and feature weight learning, Sci. China Inf. Sci. 66 (5) (2023) 1–17.

[39] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, D. Testuggine, Supervised multimodal bitransformers for classifying images and text, 2019, arXiv preprint arXiv:1909.02950.

[40] W. Wang, D. Tran, M. Feiszli, What makes training multi-modal classification networks hard? in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12695–12705.

[41] Y. Gan, R. Han, L. Yin, W. Feng, S. Wang, Self-supervised multi-view multi-human association and tracking, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 282–290.

[42] J. Wang, Y. Zheng, J. Song, S. Hou, Cross-view representation learning for multi-view logo classification with information bottleneck, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4680–4688.

[43] W. Liu, X. Yue, Y. Chen, T. Denoeux, Trusted multi-view deep learning with opinion aggregation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 7585–7593.

[44] C. Chen, C.-M. Vong, S. Wang, H. Wang, M. Pang, Easy domain adaptation for cross-subject multi-view emotion recognition, Knowl.-Based Syst. 239 (2022) 107982.

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

[46] Y. Li, B. Fu, F. Li, G. Shi, W. Zheng, A novel transferability attention neural network model for EEG emotion recognition, Neurocomputing 447 (2021) 92–101.

[47] Z. Wang, Y. Wang, C. Hu, Z. Yin, Y. Song, Temporal-spatial representation learning transformer for EEG-based emotion recognition, 2022, arXiv preprint arXiv:2211.08880.

[48] X. Si, D. Huang, Y. Sun, D. Ming, Temporal aware mixed attention-based convolution and transformer network (MACTN) for EEG emotion recognition, 2023, arXiv preprint arXiv:2305.18234.

[49] Y. Hao, H. Cao, A new attention mechanism to classify multivariate time series, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 2020.

[50] T. Song, W. Zheng, P. Song, Z. Cui, EEG emotion recognition using dynamical graph convolutional neural networks, IEEE Trans. Affect. Comput. 11 (3) (2018) 532–541.

[51] X. Du, C. Ma, G. Zhang, J. Li, Y.-K. Lai, G. Zhao, X. Deng, Y.-J. Liu, H. Wang, An efficient LSTM network for emotion recognition from multichannel EEG signals, IEEE Trans. Affect. Comput. (2020).

[52] J.A. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (1999) 293–300.

[53] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1) (1967) 21–27.

[54] G. Zhang, M. Yu, Y.-J. Liu, G. Zhao, D. Zhang, W. Zheng, Sparsedgcnn: Recognizing emotion from multichannel EEG signals, IEEE Trans. Affect. Comput. (2021).

[55] H. Chen, M. Jin, Z. Li, C. Fan, J. Li, H. He, MS-MDA: multisource marginal distribution adaptation for cross-subject and cross-session EEG emotion recognition, Front. Neurosci. 15 (2021) 778488.

[56] Z. He, Y. Zhong, J. Pan, An adversarial discriminative temporal convolutional network for EEG-based cross-domain emotion recognition, Comput. Biol. Med. 141 (2022) 105048.

[57] W. Guo, G. Xu, Y. Wang, Horizontal and vertical features fusion network based on different brain regions for emotion recognition, Knowl.-Based Syst. 247 (2022) 108819.

[58] Y. Wang, S. Qiu, D. Li, C. Du, B.-L. Lu, H. He, Multi-modal domain adaptation variational autoencoder for eeg-based emotion recognition, IEEE/CAA J. Autom. Sin. 9 (9) (2022) 1612–1626.

[59] W. Guo, G. Xu, Y. Wang, Multi-source domain adaptation with spatio-temporal feature extractor for EEG emotion recognition, Biomed. Signal Process. Control 84 (2023) 104998.

[60] W. Li, B. Hou, X. Li, Z. Qiu, B. Peng, Y. Tian, TMLP+ SRDANN: A domain adaptation method for EEG-based emotion recognition, Measurement 207 (2023) 112379.

[61] Z. Wang, Y. Wang, J. Zhang, Y. Tang, Z. Pan, A lightweight domain adversarial neural network based on knowledge distillation for EEG-based cross-subject emotion recognition, 2023, arXiv preprint arXiv:2305.07446.

[62] A. Etkin, T. Egner, R. Kalisch, Emotional processing in anterior cingulate and medial prefrontal cortex, Trends Cogn. Sci. 15 (2) (2011) 85–93.

[63] T. Canli, J.E. Desmond, Z. Zhao, G. Glover, J.D. Gabrieli, Hemispheric asymmetry for emotional stimuli detected with fMRI, Neuroreport 9 (14) (1998) 3233–3239.