Disentangled Representation Learning for Robust Brainprint Recognition

Chuhang Zheng, Qi Zhu, *Member, IEEE*, Lunke Fei, *Senior Member, IEEE*, Shengrong Li, Xiangping Bryce Zhai, *Member, IEEE*, David Zhang, *Life Fellow, IEEE*, and Daoqiang Zhang, *Senior Member, IEEE*

***Abstract*—Electroencephalography (EEG) biometrics draws increasing attention in high-security requirements due to its advantages of anti-spoofing, live traits, and non-duplicated. However, existing EEG datasets, which rely on external stimuli or task-specific instructions for data collection, often intertwine identity-related information with biases such as emotional states, cognitive tasks, and disease markers. Besides, EEG signals are time-varying, while identity information within EEG signals is relatively fixed, which poses challenges for extracting identity features from EEG to perform accurate person identification. This high correlation hampers the promotion of brainprint recognition in real-life applications. In this paper, we propose a disentangled representation learning based identity recognition framework, which disentangles the EEG signal into intrinsic identity-related information and biased identity-invariant information, thus enhancing the performance of EEG biometrics. First, two parallel encoders are used to extract intrinsic identity-relevant and bias identity-irrelevant factors, respectively, and each encoder consists of a temporal filter module and a novel spatial-temporal attention module. Then, we further refine the disentanglement process through a correlation-driven loss that minimizes factor similarity across spatial-temporal and global representational domains. Adversarial training and reconstruction regularization are introduced to facilitate the identity and biased representations to be independent and complementary to each other. Additionally, we extend supervised contrastive learning to the component level, minimizing cross-component similarity and encouraging each component to independently reflect its unique information, thereby improving the disentanglement efficacy. Our proposed framework achieves state-of-the-art performance on diverse datasets encompassing emotional, motor imagery, and pathological conditions, demonstrating the robustness and effectiveness of our proposed brainprint identity recognition model.**

***Index Terms*—EEG, biometrics, brainprint, contrastive learning, adversarial training.**

I. Introduction

BIOMETRICS technique plays a vital role in security scenarios, such as smart devices, internet information security, etc. The commonly used biometric traits include face [1], fingerprint [2], palmprint [3], iris [4], etc. Recently, EEG based brainprint [5] has emerged as a promising biometric technology, offering distinct advantages over other traits by utilizing electrical signals generated by brain activity for individual recognition. First, under conditions of stress or fear, the neural patterns of the human brain can be altered. This alteration can prevent attackers from using threats to gain access to users, thereby affirming the robustness of EEG-based security measures against spoof attacks [6], [7]. Besides, unlike other biometric traits that can be mimicked or replicated, the unique neural patterns captured by EEG signals are inherently tied to living biological processes, rendering them inimitable and enhancing security. Consequently, EEG biometrics can be applicable to scenarios with high-security requirements [8], such as the military and financial industries, where the development of precise brainprint recognition models is of paramount importance.

With the rapid development of artificial intelligence and EEG sensing, EEG-based person identification approaches have drawn increasing attention in recent years. Behrouzi and Hatzinakos [9] developed a novel graph variational auto-encoder (GVAE) method to extract nodal features of brain functional connections to obtain more accurate features from a graph perspective and achieved promising performance on three biometric databases. Wilaiprasitporn et al. [10] proposed a cascade of deep learning using a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs), where CNNs are used to handle spatial information and RNNs are adopted to extract temporal information. Li et al. [11] utilized a tensorial scheme to extract the effective tensorial representation from multichannel EEG, and the scheme performs the designed tensorial learning to improve the discriminability of the feature space and finally carries out the devised tensorial measurement in the learned metric

Received 11 June 2024; revised 15 June 2025; accepted 17 August 2025. Date of publication 25 August 2025; date of current version 5 September 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62076129, Grant 62371234, and Grant 62136004; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20231438; and in part by the Key Research and Development Plan of Jiangsu Province under Grant BE2022842. The associate editor coordinating the review of this article and approving it for publication was Dr. Naser Damer. *(Corresponding author: Qi Zhu.)*

Chuhang Zheng, Qi Zhu, Shengrong Li, Xiangping Bryce Zhai, and Daoqiang Zhang are with the College of Artificial Intelligence and the Key Laboratory of Brain-Machine Intelligence Technology, Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China (e-mail: h.zheng@nuaa.edu.cn; zhuqinuaa@163.com; Lisrong@nuaa.edu.cn; blueicezhaixp@nuaa.edu.cn; dqzhang@nuaa.edu.cn).

Lunke Fei is with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China (e-mail: flksxm@126.com).

David Zhang is with the School of Data Science, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: davidzhang@cuhk.edu.cn).

Digital Object Identifier 10.1109/TIFS.2025.3602266

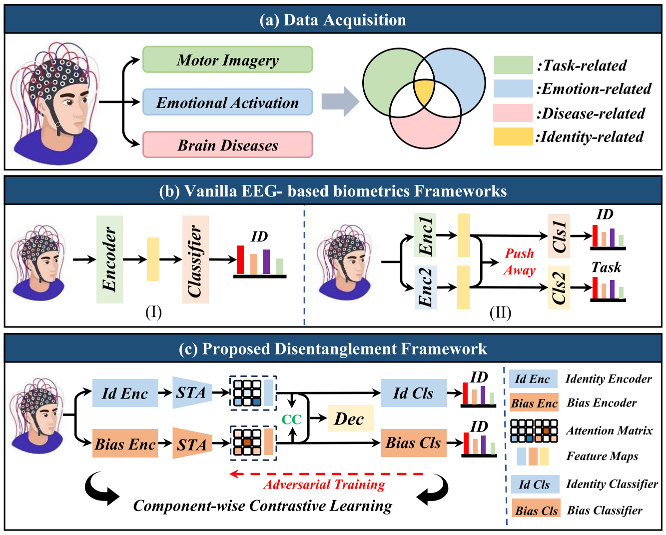


Fig. 1. Diagrams of data acquisition and comparison of our proposed method with existing EEG-based biometrics methods. (a) Workflow of existing EEG data acquisition. (b) Existing EEG-based biometrics frameworks. (c) The proposed disentanglement-based EEG-biometrics frameworks.

space for classification. However, the above methods focus on specific cognitive tasks or activation, which presents a significant obstacle to the promotion of EEG-based person identification (PI).

As can be seen from Fig. 1(a), the acquisition of EEG signals may require the induction of specific tasks, such as emotional activation, motor imagery, etc. This entanglement can compromise the accuracy of identification processes, as biased information may obscure the underlying identity signals that are crucial for recognition. Disentangled Representation Learning (DRL) emerges as a potent solution for disentangling the complex, intertwined factors present within data. Kong et al. [12] assumed that task-related EEG can be divided into a background EEG (BEEG) and a residual EEG (REEG), and further proposed a fast LRMD-based EEG identification algorithm with maximum correntropy criterion (MCC) and rational quadratic kernel, which can efficiently extract BEEG out and deliver a high accuracy classification. Özdenizci et al. [13] proposed an adversarial inference approach to extend such deep learning models to learn task-invariant person-discriminative representations that can provide robustness in terms of longitudinal usability. Bollens et al. [14] adopted factorized hierarchical variational autoencoders to exploit parallel EEG recordings of the same stimuli. Jin et al. [15] designed a disentangled adversarial generalization network (DAGN) for stable task-independent brainprint recognition, which effectively decorrelated the disentangling task-relevant and identity-relevant features within EEG signals. It is worth noting that the researches mentioned above have primarily been validated on a single type of dataset. This limitation can result in suboptimal robustness when the models are confronted with diverse tasks or activation data, thereby constraining the generalizability of EEG-based PI systems.

The spatial topology and temporal dynamics are both crucial for extracting meaningful patterns from dynamic signals. Therefore, researchers have implemented techniques to exploit complex spatial-temporal patterns within EEG signals. Song et al. [16] proposed DGCNN to dynamically learn the intrinsic relationship between different channels, which enables more discriminative spatial topology feature extraction. Du et al. [17] implemented long short-term memory (LSTM) to extract the crucial temporal features for emotion recognition. These methods captured only spatial or temporal correlations, which limits the model's performance. Song et al. [18] proposed an EEG classification model named ConFormer, which utilized the self-attention mechanism and CNN to couple local and global features of EEG signals. Ding et al. [19] designed a multi-scale convolutional neural network that extracts dynamic temporal and asymmetric spatial features for accurate emotion recognition. The above methods extract the spatial and temporal features in different stages, falling short in learning the complex coupling of spatial-temporal patterns within EEG signals. However, the spatial-temporal coupling can better describe the invariant individual characteristics in time-varying EEG signals.

The existing EEG biometric methods are illustrated in Fig. 1(b). For example, methods like (I) do not consider the entangled information within EEG signals, while the methods in (II) require labels corresponding to specific tasks, which brings challenges to model training and real-life application promotion. To address the above issues, as can be seen in Fig. 1(c), we propose an EEG-based biometrics framework via adversarial disentanglement and component-wise contrastive learning, which effectively extracts identity-related and task-related information from EEG signals to achieve accurate identification. Specifically, we design two parallel encoders, i.e. temporal filter module and spatial-temporal attention module, to extract intrinsic identity-related and bias task-related representations simultaneously. Then, due to the different temporal sensitivities of identity and bias information, we adopt correlation coefficient minimization to the cross-factor spatial-temporal attention matrix and global latent representations for the disentanglement of the identity-related features and identity-irrelevant features. What's more, a decoder is employed to reconstruct the combination of two latent representations, which ensures these two components are complementary to each other. For the identity branch, a classifier is trained to accomplish accurate person identification while an adversarial training strategy is introduced to the bias branch, which further eliminates the task-related information from EEG to a great extent. Finally, we extend supervised contrastive learning to component-wise feature extraction, which not only alleviates the cross-subject similarity in some tasks or activations that affect the performance of PI but also encourages the disentanglement efficiency of our model. In brief, the main contributions of our proposed method are as follows:

1) We propose a generalized feature disentanglement framework for the adaptive extraction of intrinsic identity-related discriminative representation against bias representation, such as emotion, motor imagery, and brain disease characteristics, within EEG signals for accurate and robust EEG biometrics.

2) We develop a novel spatial-temporal attention module to capture the spatial-temporal patterns within EEG signals, and further present a correlation-driven loss to preliminarily disentangle the entangled identity-bias factors. Then, the disentanglement is further improved by complementary constraint and adversarial training.

3) To learn better disentangled representations, the supervised contrastive learning strategy is extended to the component level to induce independence between the latent representations.

4) We conduct extensive experiments on various identity recognition tasks under various interferences, including motor imagery, emotion, and brain disease EEG datasets, to verify the effectiveness of our proposed method. The experimental results demonstrate that our model achieves state-of-the-art PI performance with the help of multi-stage disentanglement.

The rest of this paper is organized as follows. In Section II, we introduce the related works. Then we present the details of the proposed method and its optimization in Section III. The experimental results are shown in Section IV. Finally, we draw a summary in Section V.

II. Related Works

A. Disentangled Representation Learning

Disentangled representation learning (DRL) can factorize the data into discriminative latent representations of variations. It has been widely used in classification [20], generative models [21], etc. Early research on learning disentangled representation is mostly based on generative adversarial networks (GAN) [22] and auto-encoders [23]. Xu et al. [24] proposed a novel algorithm with asymmetric adversarial-based feature disentanglement learning to extract domain-invariant and discriminative representation. Sanchez et al. [25] adopted mutual information maximization to capture the attributes of data in the shared and exclusive representations while minimizing the mutual information between the shared and exclusive representations to enforce disentanglement. Xu et al. [26] designed an adversarial feature disentanglement network, which contains intra-class reconstruction and inter-class adversaries to disentangle the identity-related and identity-unrelated features. Jin et al. [15] proposed an adversarial framework named DAGN to remove the remaining label-related features from dominant features for the disentanglement of task-relevant and identity-relevant features in brainprint recognition. This goal of separating underlying factors of variation shares conceptual similarities with multi-view learning (MVL), where different data modalities or feature subsets are treated as distinct views. The objective in MVL is often to learn a common underlying representation or to leverage complementary information across views. For instance, recent MVL methods [27], [28], [29] explore multi-view complementarity to enhance the model's performance on downstream tasks. In comparison with the existing DRL and MVL methods, we combine adversarial training and contrastive learning to disentangle the identity-related and identity-irrelevant information to enhance EEG biometrics.

B. Contrastive Learning

Contrastive learning (CL) is a popular representation learning method, which aims at making similar instances closer and dissimilar instances far from each other in the latent space [30], [31]. This method can be applied to both supervised and unsupervised settings. SupCon [32] extends SimCLR [33] to a fully-supervised manner, which is a well-known self-supervised framework, leading to state-of-the-art performance for image classification. The SimCLR loss can be formulated as follows:

$$L = -\sum_{i \in I} \log \frac{\exp\left(z_i \cdot z_{j(i)}/\tau\right)}{\sum_{a \in A(i)} \exp\left(z_i \cdot z_a/\tau\right)} \tag{1}$$

where z_i and $z_{(j(i))}$ are the embeddings from input data, which are generated from two encoders, $\cdot$ denotes the inner product, index i is called the anchor, index $j(i)$ is called positive, and the other $2(N-1)$ indices are called negatives. Its positive samples are drawn from samples of the same class as the anchor, rather than being data augmentations of the anchor. SupCon loss extends SimCLR to a fully supervised manner by leveraging label information. The use of many positives and negatives for each anchor demonstrates state-of-the-art performance without the need for hard-negative mining, which can be difficult to tune properly. With the great effect of CL, many recent studies applied it to the field of physiological signal analysis, i.e. EEG. Lee et al. [34] proposed a deep learning framework that adopts both self-supervised contrastive learning and supervised contrastive learning to force the robustness of the latent representations for automatic sleep staging. Ho and Armanfard [35] introduced a contrastive learning module to learn the local topological structure of graphs, which enhances the detection of seizures. In our framework, we eliminate cross-subject similarity information and promote feature disentanglement efficiency by extending supervised contrastive learning to the component level.

III. Proposed Method

The proposed adversarial disentanglement and contrastive learning-based person identification framework mainly consists of two modules: adversarial disentanglement network (ADN) and component-wise contrastive learning. EEG signal is disentangled into two complementary parts: intrinsic identity-related feature and bias feature by the ADN with the help of correlation-driven loss, complementary constraint and adversarial training strategy. What's more, we further introduce a component-wise contrastive learning method to encourage the disentanglement performance of ADN and eliminate the cross-subject invariance. The proposed framework is shown in Fig. 2.

A. Adversarial Disentanglement Network

The architecture of ADN is shown in the first half of Fig. 2. We first designed two encoders to handle intrinsic identity-related representation and bias task-related representation separately. Specifically, the identity or bias encoder consists of a temporal filter (TF) module and a spatial-temporal attention module, the architecture of the temporal filter module

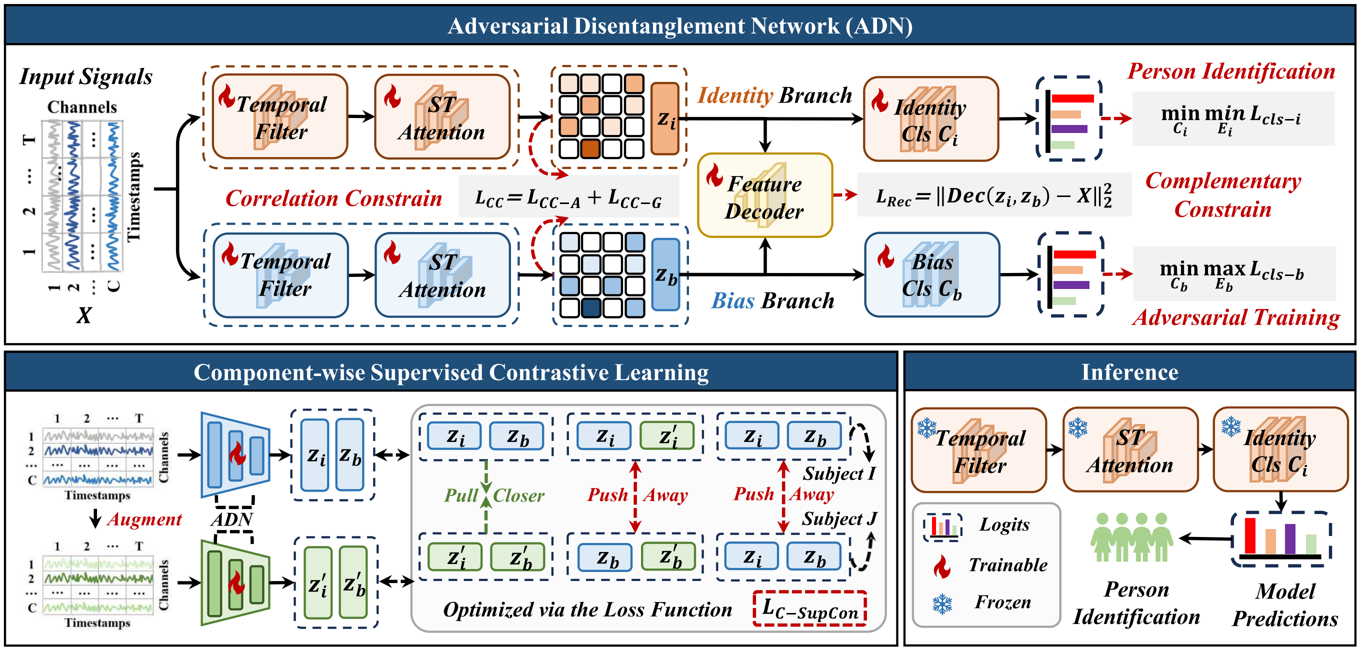


Fig. 2. The framework of our proposed model. The pre-processed EEG signal and its augmented version are fed into a shared Adversarial Disentanglement Network (ADN). ADN consists of two separate channels, E_i and E_b encode the identity and bias information of EEG, which consists of a temporal filter module (TF) and a spatial-temporal attention module (STA), respectively. A correlation-driven loss is introduced to both spatial-temporal and global disentanglement. A decoder is adopted to reconstruct the combination of two latent representations, which ensures these two components are complementary to each other. For E_i, we train a classifier to accomplish accurate EEG biometrics. In contrast with E_i, an adversarial training scheme is used to maximize the CE loss of E_b to further eliminate the biased information. The supervised contrastive learning framework is extended to component-wise to improve the disentanglement performance of the proposed model.

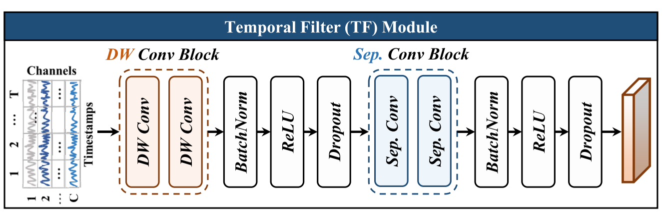


Fig. 3. The proposed temporal filter module, which contains a depthwise convolutional block and a separable convolutional block.

is illustrated in Fig. 3, it first extracts the local temporal information of EEG signals and can reduce the model's parameters. The depthwise convolutional block contains two depth-wise convolution layers, a batchnorm layer, a ReLU activation function layer and a dropout layer, while the separable convolutional block contains two separable convolution layers, a batchnorm layer, a ReLU activation function layer and a dropout layer. For a given input $X \in \mathbb{R}^{C \times T}$, where C is the number of electrodes and T denotes the number of timepoints, we can obtain two latent representations $X_i \in \mathbb{R}^{C \times T_1}$ and $X_b \in \mathbb{R}^{C \times T_1}$ by the separate encoders, where T_1 represents the new temporal dimension, which can be formulated as:

$$X_i = TF(X), X_b = TF(X) \tag{2}$$

where $TF(\cdot)$ and $TF(\cdot)$ are temporal filter modules for identity channel and bias channel, respectively. Recent research has drawn the conclusion that identity-related information is less sensitive than that of task-related information [12], [36]. Therefore, we present a novel spatial-temporal attention

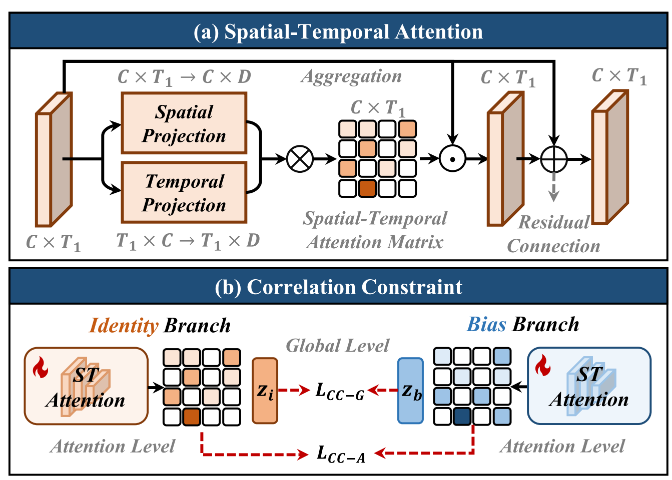


Fig. 4. The proposed (a) spatial-temporal attention module and (b) correlation-driven constraint, where $\otimes$, $\odot$, and $\oplus$ denote the matrix multiplication, dot product and element-wise addition, respectively.

(STA) to capture the spatial-temporal patterns of identity-related information or bias information within EEG signals, the architecture of STA is illustrated in Fig. 4(a). Specifically, for a given representation X_i or X_b obtained by the local temporal feature extractor, we first perform Spatial Projection (SP) over the spatial dimension and Temporal Projection (TP) over the temporal dimension, which is implemented with a simple linear transformation:

$$\begin{aligned} S_i &= SP(X_i), T_i = TP(\text{Reshape}(X_i)) \\ S_b &= SP(X_b), T_b = TP(\text{Reshape}(X_b)) \end{aligned} \tag{3}$$

where Reshape(·) is adopted to transpose the dimensions of X_i or X_b from $C \times T_1$ to $T_1 \times C$. $S_i, S_b \in \mathbb{R}^{C\times D}$ denote the spatial information, while T_i, $T_b \in \mathbb{R}^{T_1\times D}$ represent the temporal information. Then, we integrate the spatial information and temporal information to obtain a spatial-temporal attention matrix by a matrix multiplication operation, which is then input to the SoftMax function:

$$A_i = SoftMax\left(S_i T_i^T\right), A_b = SoftMax\left(S_b T_b{}^T\right) \tag{4}$$

where A_i and A_b are the spatial temporal attention scores of X_i and X_b. Finally, the latent global spatial-temporal representations z_i and z_b are obtained by the product of spatial-temporal attention weight and X_i, X_b:

$$z_i = A_i X_i + X_i, z_b = A_b X_b + X_b \tag{5}$$

where the residual connection is adopted to avoid gradient vanishing, $z_i, z_b \in \mathbb{R}^{C\times T_1}$ are the final identity-related representation and bias representation. Then, we introduce a correlation-driven loss to minimize the correlation coefficient between identity and bias representations:

$$\begin{aligned} \mathcal{L}_{CC} &= \mathcal{L}_{CC-A} + \mathcal{L}_{CC-G} \\ \mathcal{L}_{CC-A} &= CC\left(A_i, A_b\right) \\ \mathcal{L}_{CC-G} &= CC\left(z_i, z_b\right) \end{aligned} \tag{6}$$

where $CC(\cdot)$ denotes the correlation coefficient operator, $\mathcal{L}_{CC-A}$ is adopted to minimize the similarity between A_i, A_b, and $\mathcal{L}_{CC-G}$ is used to disentangle the global identity and bias features.

After the feature extraction stage, we design a decoder Dec to reconstruct the original input by concatenation of (z_i, z_b). Here, we measure the similarity between the original input X and the reconstructed output via the mean squared error (MSE) loss:

$$\mathcal{L}_{\text{Recon}} = \|Dec\left(z_i, z_b\right) - X\|_2^2 \tag{7}$$

This complementary constraint contributes to the disentanglement performance, enabling each component to contain enough information. For the identity encoder E_i, our goal is to achieve accurate PI. Therefore, an objective of E_i is to ensure the classification accuracy of the PI task, in which we adopt a multi-class cross-entropy (CE) loss to update the parameters of E_i:

$$\mathcal{L}_{cls-i} = -\sum_{k=1}^{K} y_k \log\left(p_k^i\right) \tag{8}$$

where K is the number of categories, y_i represents the ground truth, and p_k^i denotes the probability of class k obtained by the SoftMax function after the identity classifier C_i. To train the bias encoder E_b, inspired by Generative Adversarial Network (GAN) [37], we design an adversarial classifier to perform a min-max game by training C_b to make correct prediction while the E_b is trained to eliminate the identity-related information within EEG signals by fooling C_b to make false predictions. The min-max game can be formulated as:

$$\min_{C_b} \max_{E_b} \mathcal{L}_{cls-b} \tag{9}$$

By the adversarial training strategy, the identity-related information is eliminated from the EEG signals while the E_b is closer to extracting task-related information as the training goes on, which further demonstrates the disentanglement.

B. Component-Wise Supervised Contrastive Learning

Supervised contrastive loss (SupCon) [32] is an extension of self-supervised contrastive loss (SimCLR) [33], which extends the SimCLR to the fully-supervised setting by leveraging label information, leading to state-of-the-art image classification performance. Contrastive representation learning requires one positive sample (e.g., an augmented version of the anchor sample) and many negative samples (e.g., randomly chosen samples from the mini-batch). The original SupCon loss is as follows:

$$\mathcal{L}_{SupCon} = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp\left(v_i \cdot v_p/\tau\right)}{\sum_{a\in A(i)} \exp\left(v_i \cdot v_a/\tau\right)} \tag{10}$$

where $P(i) = \left\{p \in P(i) : y_p = y_i\right\}$ are the indices of positive samples in the augmented batch with respect to v_i, and $|P(i)|$ is the cardinality of $P(i)$. v_i, v_p, v_a denote the anchor, positive and negative samples, respectively. $A(i)$ represents the index sets of negative samples and τ is the temperature of similarity functions, i.e., cosine similarity. To further contribute to the disentanglement of ADN, we proposed a new representation learning objective via extending SupCon to component-wise. Specifically, at the beginning of the proposed model, a data augmentation module $Aug(\cdot)$ transforms the original pre-processed input X into a slightly different but semantically identical signal as follows:

$$X' = Aug(X) \tag{11}$$

where X' denotes the augmented version of X. For a given batch of data $\{x_k, y_k\}_{k=1\ldots N}$, where N_b is the batch size. The corresponding augmented batch used for training consists of $2N_b$ pairs, $\{\widetilde{x_l}, \widetilde{y_l}\}_{l=1\ldots 2N}$. Different from the original SupCon, to better handle the similarity redundant information that affects PI across subjects, we only perform one random augmentation to the original data. Augmentation is crucial for contrastive representation learning, following the previous studies [38], [39], we apply two random augmentations, namely, ***jittering*** and ***bandpass filtering*** with a probability of 0.5. The details of the augmentation will be discussed in Section IV. For the original data X and augmented data X' that can be formed as a positive pair, are fed into a shared ADN:

$$z_i, z_b = ADN(X), z_i', z_b' = ADN(X') \tag{12}$$

where z_i, z_b and z_i', z_b' are the disentangled representations of original data and augmented data, respectively. Our goal is to maximize the agreement of z_i, z_i' or z_b, z_b', while minimizing the similarity between z_i, z_i' and z_b, z_b', which not only demonstrates the disentanglement of the previous ADN, but also eliminates the similarity information within different subjects. Formally, we cannot directly calculate the contrastive loss via these four components. Therefore, for latent vectors (z_i, z_b) or latent vectors (z_i', z_b'), we obtain two latent vectors that need to be pulled closer by a concatenation operator:

$$z_{c-i} = Concat(z_i, z_b), z_{c-p} = Concat(z_i', z_b') \tag{13}$$

where z_{c-i} and z_{c-p} denote the component-wise anchor vector and component-wise positive vector, respectively. Besides, the latent vectors that need to be pushed away are computed by an average operation:

$$z_{c-i'} = Mean(z_i, z_i'), z_{c-a'} = Mean(z_b, z_b') \tag{14}$$

where $z_{c-i'}$ and $z_{c-a'}$ denote the disentangled representations that need to be pushed away. Then, the component-wise supervised contrastive learning objective can be formulated as:

$$\mathcal{L}_{C-SupCon} = -\log \frac{\exp(z_{c-i} \cdot z_{c-p}/\tau)}{\sum_{a \in A(i)} \exp(z_{c-i} \cdot z_a/\tau) + \exp(z_{c-i'} \cdot z_{c-a'}/\tau)} \tag{15}$$

where $(\cdot)$ is the inner product function with temperature τ, which calculates the similarity between two latent vectors. As can be seen from the Eq. (15), compared to the original SupCon, an extra term $\exp\varphi(z_{c-i'} \cdot z_{c-a'}/\tau)$ is added to the denominator, which facilitates the disentangled representations to be pushed away. At the same time, the component-wise agreement of original samples and augmented samples is maximized to further encourage the disentanglement of ADN. What's more, the similarity between different subjects is also minimized to eliminate cross-subject invariant information, which contributes to the better PI performance of our proposed framework.

C. Model Optimization

Our proposed framework consists of two channels: the identity channel and the bias channel. The identity channel aims at learning identity-related information and achieving accurate person identification, while the bias channel extracts task-related information via an adversarial training game. For the identity channel, we optimize the parameter by minimizing the loss function below:

$$\mathcal{L}_{identity} = \mathcal{L}_{cls-i} + \alpha(\mathcal{L}_{CC} + \mathcal{L}_{Recon}) + \beta\mathcal{L}_{C-SupCon} \tag{16}$$

where $\mathcal{L}_{CC} = \mathcal{L}_{CC-A} + \mathcal{L}_{CC-G}$. The bias channel is trained by minimizing the following objective:

$$\mathcal{L}_{bias} = -\mathcal{L}_{cls-b} + \alpha(\mathcal{L}_{CC} + \mathcal{L}_{Recon}) + \beta\mathcal{L}_{C-SupCon} \tag{17}$$

where α and β are the hyperparameters, which balance the weight of $\mathcal{L}_{CC}+\mathcal{L}_{Recon}$ and $\mathcal{L}_{C-SupCon}$, respectively. By training the two channels separately, the proposed framework can finally push the identity-related and task-related information away to improve disentanglement performance. Furthermore, the cross-subject invariant information is also eliminated by the component-wise supervised contrastive learning strategy. Algorithm 1 summarizes the algorithm description of our proposed method.

Algorithm 1 Training of the Proposed Framework

Input:
The pre-processed EEG samples $X = \{x_n, y_n\}_{n=1...N}$
Batch size N_B, hyperparameters α, β
Output:
Learned model parameters
for a random sampled mini-batch $\mathcal{B} = \{x_k, y_j\}_{k=1}^{|\mathcal{B}|}$ **do**
  for $\{x_k, y_k\} \in \mathcal{B}$ **do** ▹Feature Extraction
    $x_k' = Aug(x_k)$
    $X_{k,i}, X_{k,b} = TF_i(x_k), TF_b(x_k)$
    $A_{k,i}, A_{k,b} = STA_i(X_{k,i}), STA_b(X_{k,b})$
    $z_{k,i}, z_{k,b} = ADN(x_k), ADN(x_k)$
    $z_{k,i}', z_{k,b}' = ADN(x_k'), ADN(x_k')$
  end for
  for $\{x_k, y_k\} \in \mathcal{B}$ **do** ▹Objective Optimization
    $\mathcal{L}_{CC-A} \leftarrow$ calculated by Eq. (6) with $A_{k,i}, A_{k,b}$
    $\mathcal{L}_{CC-G} \leftarrow$ calculated by Eq. (6) with $z_{k,i}, z_{k,b}$
    $\mathcal{L}_{CC} = \mathcal{L}_{CC-A} + \mathcal{L}_{CC-G}$
    $\mathcal{L}_{Recon} \leftarrow$ calculated by Eq. (7) with $x_k, z_{k,i}, z_{k,b}$
    $\mathcal{L}_{cls-i} \leftarrow$ calculated by Eq. (8) with y_k, $p_{k,i}$
    $\mathcal{L}_{cls-b} \leftarrow$ calculated by Eq. (8) with y_k, $p_{k,b}$
    $\mathcal{L}_{C-SupCon} \leftarrow$ calculated by Eq. (15)
  end for
  $\theta_{E_i} \leftarrow \nabla_{\theta_{E_i}}\big(\mathcal{L}_{identity} = \mathcal{L}_{cls-i} + \alpha(\mathcal{L}_{CC} + \mathcal{L}_{Recon}) + \beta\mathcal{L}_{C-SupCon}\big)$
  $\theta_{E_b} \leftarrow \nabla_{\theta_{E_b}}\big(\mathcal{L}_{bias} = -\mathcal{L}_{cls-b} + \alpha(\mathcal{L}_{CC} + \mathcal{L}_{Recon}) + \beta\mathcal{L}_{C-SupCon}\big)$
end for

IV. Experiments

A. Dataset and Pre-Processing

To evaluate our proposed method, we conduct experiments on three different datasets including motor imagery, emotion, and mental disease.

1) PhysioNet Motor Movement/Imagery: The PhysioNet Motor Movement/Imagery dataset [40] is one of the most popular and publicly available datasets produced by the BCI2000 system [41]. The EEG signal was collected while 109 healthy participants were performing 6 different tasks: eyes open (EO), eyes closed (EC), left or right fist open and closed (MOV-lr), imagining MOV-lr (IMG-lr), both fists open and closed (MOV-b), and imagining MOV-b (IMG-b). The movement of a fist was based on the appearance of a target on the screen. The 64-channel EEG signals are recorded with a sampling rate of 160Hz. In our study, the 64-channel EEG signals are bandpass filtered between 0.5 and 42Hz.

2) SEED-Joint: SEED [42] was collected by Shanghai Jiao Tong University for emotion recognition. The 62-channel EEG signals of 15 subjects were recorded while watching 15 emotional film clips at a sampling rate of 1000Hz. The duration of each film clip is about 4 minutes. The data was downsampled to 200Hz and a bandpass filter from 0-75Hz was applied. SEED-IV [43] comprises 62-channel EEG signals of 15 subjects, which include four types of emotions: neutral, sad, fear, and happy. Three sessions of data are collected, each comprising 24 trials for each subject. Similar to SEED, the data was downsampled to 200Hz and then processed with a bandpass filter between 0 and 75Hz. Due to the limited subjects, we combined SEED and SEED-IV to build the SEED-Joint dataset.

3) TUH-EEG: This dataset is a subset of the Temple University Hospital EEG data corpus (TUH EEG) [44]. The

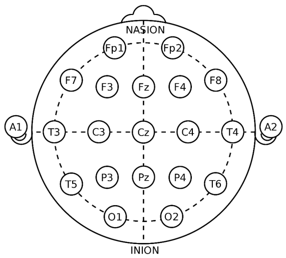


Fig. 5. Selected channels topomap of the TUH-EEG dataset.

TUH-EEG corpus v.1.1.0 contains over 20000 clinical EEG recordings collected from about 14000 patients. The dataset was variedly collected using 24 to 36 sensors. For further analysis, 1321 subjects were chosen using the following criteria owing to the higher number of subjects: 1) Containing at least two sessions. 2) 22 channels, shown in Fig. 5, are available in the EEG signals. For the 1321 subjects, one trial of the first two sessions was selected to generate samples. The EEG signals were downsampled to 200Hz, and independent component analysis was adopted for the removal of artifacts. Then, a bandpass filter between 0 and 75Hz was implemented.

B. Experimental Settings and Comparison Methods

1) Experiment Settings: To verify the effectiveness of our proposed method, 10-fold cross-validation is implemented. Specifically, data augmentation is first adopted by splitting each trial into smaller non-overlapping segments for the training in the deep neural networks, a 1s non-overlapping time window is implemented to generate samples for further experiments. Before training, the dataset is divided into 10 subsets with the same size, one subset is used as the testing set, and the other 9 subsets are used as the training set. This process is repeated 10 times. Finally, the average performance of the 10-fold cross-validation is taken as the final result.

2) Implementation Details: The proposed framework is implemented using Pytorch on a GeForce GTX 3080 GPU device with 12GB memory. The Adam optimization algorithm is used to optimize the objective function. The model is trained for a total of 200 epochs. In our proposed model, the grid search is adopted to select the hyperparameters, the parameter of the loss function, α and β are both searched from the set of $\{0.1, 0.2, 0.3, \ldots, 1\}$. The embedding size of the proposed method is tuned over the set $\{32, 64, 128, 256, 512\}$, the dropout rate is set to 0.5 and ReLU is used as the activation function. In the proposed framework, we consider two different augmentation methods: ***jittering*** and ***bandpass filtering***. ***Jittering*** adds additional perturbations to each sample. We consider both high and low-frequency noise on each channel independently. For ***bandpass filtering***, we use scipy.signal.butter to preserve only the within-band frequency information. The high-pass and low-pass are (1Hz,30Hz) and (10Hz,40Hz) for the PhysioNet dataset, (1Hz,40Hz) and (20Hz,70Hz) for the SEED-Joint and TUH-EEG dataset. Low-pass or high-pass or both are selected with equal probability. Also, the bandpass filtering is applied to each channel independently. Our code is available at: https://github.com/xbrainnet/ADN.

3) Evaluation Metrics and Comparison Methods: The proposed method and other comparison methods are evaluated using three evaluation metrics: top-1 classification accuracy (Top-1 Acc), macro-F1, and equal error rate (EER). Top-1 Acc and macro-F1 are used to evaluate the identification performance. EER is used to evaluate the performance of the models in authentication scenarios, where models determine whether a user is who he or she claims to be. To further validate the effectiveness of our method, we compare our method with several methods that can be categorized into three distinct groups: EEG biometric recognition methods, motor-imagery classification methods and EEG-based emotion recognition methods. The EEG biometric recognition methods encompass LRMD [12], CNN-Attention [45], GVAE [9], GCNN [46], DAGN [15], iCaRL [47], and W-ResNet [48]. The motor-imagery classification methods include EEGNet [49], ATCNet [50], FBCNet [51], ConFormer [18], and CSPNet [52]. In addition, the EEG-based emotion recognition methods are DGCNN [16], TSCeption [19], ArjunViT [53], and PGCN [54]. Note that all these methods used the same experimental settings, such as training-testing data split and the evaluation metrics were consistent. Then, we briefly introduce the above methods.

- **LRMD:** LRMD is a machine learning method that can efficiently extract background EEG containing subjects' unique intrinsic features for accurate EEG biometrics.
- **CNN-Attention:** CNN-Attention is a unified model that integrates CNN and Attention mechanism for person identification and classification of driving fatigue state.
- **GVAE:** GVAE can extract nodal features of brain functional connections, which considers the node neighborhood of the reconstructed adjacency matrix to demonstrate more accurate person identification.
- **GCNN:** GCNN represents EEG signals as a graph based on within-frequency and cross-frequency functional connectivity estimates for person identification.
- **DAGN:** DAGN introduces a nonlinear decorrelation method and an adversarial self-challenging strategy to disentangle identity-relevant features from task-relevant components for person identification.
- **iCaRL:** iCaRL extracts spatial-temporal-frequency domain features and adopts continual learning for robust person identification.
- **W-ResNet:** W-ResNet implements continuous wavelet transform to obtain the 2D spatial images of EEG signals and adopts ResNet for feature extraction to demonstrate effective person identification.
- **EEGNet:** EEGNet is a compact convolutional neural network for EEG-based BCIs, which introduces depthwise and separable convolutions to construct an EEG-specific model that encapsulates well-known EEG feature extraction concepts for BCI.
- **ATCNet:** ATCNet is an attention-based temporal convolutional network for EEG-based motor imagery classification, which utilizes multiple techniques to boost the MI

classification performance with a relatively small number of parameters.

- **FBCNet:** FBCNet employs a multi-view data representation followed by spatial filtering to extract spectro-spatially discriminative features and proposes a novel variance layer that effectively aggregates the EEG time-domain information for MI classification.
- **ConFormer:** ConFormer introduces convolution module and self-attention module to extract the low-level local features and global correlation to encapsulate local and global features in a unified EEG classification framework.
- **CSPNet:** CSPNet integrates knowledge-driven common spatial patterns (CSP) filters with convolutional neural networks to perform robust MI classification.
- **DGCNN:** DGCNN can dynamically learn the intrinsic relationship between different EEG channels so as to benefit from more discriminative EEG feature extraction for emotion recognition.
- **TSCeption:** TSCeption is a multi-scale convolutional neural network for emotion recognition that consists of dynamic temporal, asymmetric spatial, and high-level fusion layers, which learns discriminative representations in the time and channel dimensions simultaneously.
- **ArjunViT:** ArjunViT implements a variation of the Transformer, the Vision Transformer (ViT) to process EEG signals for accurate emotion recognition.
- **PGCN:** PGCN aggregates features at local, mesoscopic, and global levels, which achieves state-of-the-art EEG-based emotion recognition performance.

Notably, DAGN employs cognition task labels for adversarial training [15]. Therefore, for the PhysioNet dataset, 6 different tasks were adopted as the cognition task label. For the SEED-Joint dataset, we implemented the emotion labels as task labels. To unify the emotion labels, ‘fear’ and ‘sad’ in the SEED-IV dataset are mapped to ‘negative’, and ‘happy’ is mapped to ‘positive’, thereby remaining consistent with the emotion categories of the SEED dataset. Since the EEG signals in the TUH-EEG dataset are mostly the resting-state data of patients, DAGN was not evaluated on this dataset. Besides, since PGCN [54] requires the spatial distance between different electrodes to construct the initial adjacency matrix, the experiment was only conducted on the SEED-Joint dataset.

C. Performance Comparison With Other Methods

We compared the performance of our method and other comparison methods on PhysioNet, SEED-Joint, and TUH-EEG datasets, and the experiment results are illustrated in Table I, Table II and Table III. As can be seen from the tables, our proposed method achieves the best performance among all methods on the three datasets. Specifically, the average Top-1 Acc of our method reaches 99.55%, 99.85%, and 81.28%, which surpasses suboptimal results by 0.98%, 1.04%, and 4.94% on PhysioNet, SEED-Joint, and TUH-EEG datasets, respectively. The EER of our proposed model achieves 0.51%, 0.16%, and 5.56%, which are obviously lower than other comparison methods. It can be seen from the tables that not only does our method achieve the highest average Top-1 Acc and F1-score, but it also performs the lowest average EER compared to other models. This demonstrates the effectiveness of our approach in EEG- biometrics recognition.

TABLE I

IDENTIFICATION AND AUTHENTICATION PERFORMANCE (MEAN±STD) OF OUR METHOD AND THE COMPARISON METHODS ON THE PHYSIONET DATASET (%). THE BEST PERFORMANCES ARE IN BOLD AND THE SECOND RUNNERS ARE UNDERLINED

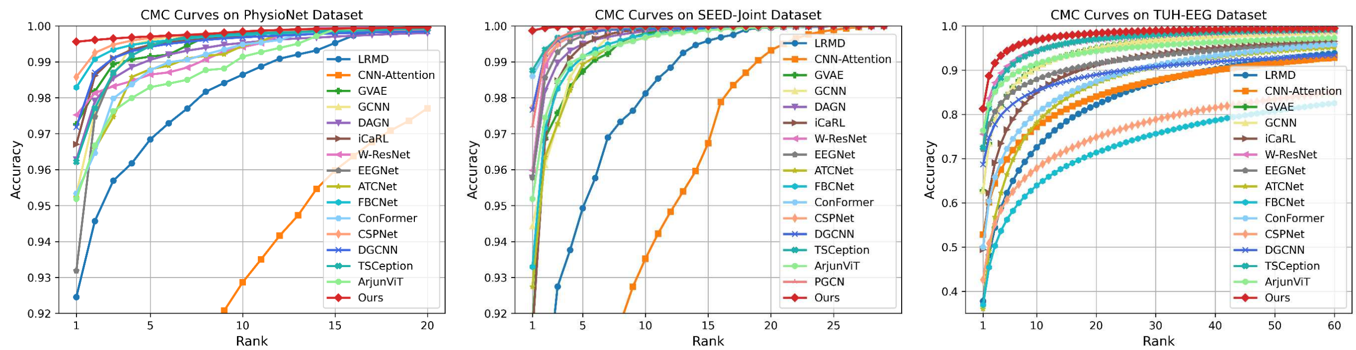
Method	Metrics		
	ACC (↑)	macro-F1 (↑)	EER (↓)
LRMD	92.45±2.77	92.04±3.45	9.81±6.82
CNN-Attention	83.62±6.60	82.75±4.18	14.59±7.64
GVAE	97.26±0.28	95.25±0.20	1.94±0.18
GCNN	95.31±0.33	95.28±0.42	4.17±3.51
DAGN	96.42±0.27	96.41±0.26	2.88±0.59
iCaRL	96.70±0.29	96.68±0.31	2.92±1.10
W-ResNet	97.52±0.46	97.51±0.46	2.39±0.22
EEGNet	93.18±4.32	93.33±3.83	6.83±4.14
ATCNet	95.23±0.33	96.05±2.21	2.45±0.19
FBCNet	98.29±0.29	98.28±0.29	<u>1.71±0.27</u>
ConFormer	95.33±0.20	96.72±0.20	3.83±0.32
CSPNet	<u>98.57±0.19</u>	<u>98.56±0.20</u>	1.72±0.09
DGCNN	97.18±0.17	97.18±0.16	2.80±0.16
TSCeption	96.21±2.37	97.28±2.74	3.11±0.37
ArjunViT	95.18±0.25	95.04±0.23	3.09±0.45
Ours	**99.55±0.17**	**99.41±0.11**	**0.51±0.12**

TABLE II

IDENTIFICATION AND AUTHENTICATION PERFORMANCE (MEAN±STD) OF OUR METHOD AND THE COMPARISON METHODS ON THE SEED-JOINT DATASET (%). THE BEST PERFORMANCES ARE IN BOLD AND THE SECOND RUNNERS ARE UNDERLINED

Method	Metrics		
	ACC (↑)	macro-F1 (↑)	EER (↓)
LRMD	87.31±3.10	88.15±2.48	12.11±4.55
CNN-Attention	76.25±3.29	70.36±3.37	25.75±5.30
GVAE	91.65±1.21	91.49±0.98	9.75±1.78
GCNN	94.42±0.85	94.35±0.88	6.70±0.68
DAGN	95.80±0.34	95.57±0.37	3.22±0.29
iCaRL	91.43±2.13	90.76±2.33	8.62±2.17
W-ResNet	95.98±0.38	95.73±0.41	3.92±0.36
EEGNet	95.77±0.60	95.60±0.59	4.51±0.59
ATCNet	92.74±0.78	92.72±0.66	8.39±0.92
FBCNet	93.29±0.80	92.10±0.97	8.02±1.00
ConFormer	98.62±0.06	98.60±0.02	0.85±0.20
CSPNet	97.71±0.20	97.44±0.22	1.24±0.22
DGCNN	97.66±0.25	97.36±0.31	2.62±0.31
TSCeption	<u>98.81±0.06</u>	<u>98.76±0.04</u>	<u>0.76±0.21</u>
ArjunViT	95.10±0.39	94.78±0.55	3.47±0.39
PGCN	97.21±0.25	97.03±0.26	2.27±0.24
Ours	**99.85±0.05**	**99.83±0.06**	**0.16±0.05**

The main reason for the superiority of our method is that, compared with other methods, our proposed method not only effectively extracts the spatial-temporal patterns within EEG signals, but also considers the entangled components such as identity-related information and task-related information, which limits the identification performance. What's more, our model implements component-wise contrastive



(a) (b) (c)

Fig. 6. CMC curves of the proposed method and other comparison methods on three datasets: (a), (b), and (c) are CMC curves on the PhysioNet dataset, SEED-Joint dataset, and TUH-EEG dataset, respectively.

TABLE III

IDENTIFICATION AND AUTHENTICATION PERFORMANCE (MEAN±STD) OF OUR METHOD AND THE COMPARISON METHODS ON THE TUH-EEG DATASET (%). THE BEST PERFORMANCES ARE IN BOLD AND THE SECOND RUNNERS ARE UNDERLINED

Method	Metrics		
	ACC (↑)	macro-F1 (↑)	EER (↓)
LRMD	37.79±2.37	35.50±2.48	18.82±0.35
CNN-Attention	52.80±6.45	51.70±6.47	12.04±0.78
GVAE	73.53±1.11	71.94±1.03	8.82±0.51
GCNN	62.75±2.15	60.97±2.09	6.96±0.25
iCaRL	49.95±6.54	47.42±6.64	13.29±1.25
W-ResNet	75.53±2.17	74.15±2.29	<u>6.72±0.94</u>
EEGNet	72.12±2.30	71.53±2.25	11.67±0.27
ATCNet	37.05±4.70	34.28±4.87	27.94±1.05
FBCNet	36.93±1.61	34.86±1.68	10.38±0.37
ConFormer	49.96±4.04	47.98±3.88	22.59±0.46
CSPNet	42.60±2.18	42.48±1.97	20.78±1.06
DGCNN	68.67±2.82	68.14±2.81	9.47±0.56
TSCeption	72.42±1.28	70.87±1.49	7.62±0.77
ArjunViT	<u>76.34±0.72</u>	<u>74.90±0.79</u>	10.35±0.26
Ours	**81.28±1.99**	**79.99±1.07**	**5.56±0.29**

learning to promote the disentanglement of EEG signals for EEG-biometrics recognition. EEG-biometrics methods such as CNN-Attention, GCNN, iCaRL, and W-ResNet face challenges in capturing the spatial-temporal patterns within EEG signals, and exhibit limitations in disentangling the identity and bias information in EEG. Motor-imagery classification methods and EEG-based emotion recognition methods such as Conformer, CSPNet, TSCeption, and PGCN, learn the spatial or spatial-temporal features for EEG classification, but also ignore the disentanglement of identity-related features and bias features, thus falling short in biometric recognition. LRMD decomposes the identity and task-related features for EEG biometrics. However, the limited feature extraction ability of machine learning methods for high-dimensional signals limits the recognition performance of the model. Although the DAGN method utilizes cognition task labels for the feature disentanglement, the considerable training difficulty results in suboptimal biometric recognition performance. Besides, its inherent dependence on cognition task labels fundamentally restricts the model's generalizability in real-world scenarios. In contrast, our model can effectively extract the spatial-temporal information within EEG signals and disentangle the entangled latent factors in EEG to a great extent, thus achieving the best identification performance.

Fig. 6(a)-(c) show the cumulative match characteristic (CMC) curves on PhysioNet, SEED-Joint, and TUH datasets, respectively. The CMC curves confirmed that our method outperforms the comparison methods. It can be observed from the figure that as the number of ranks increases, our method consistently performs the highest accuracy across the three datasets, which also demonstrates the robustness of our model.

D. Ablation Experiments

The proposed model contains the following four modules, including spatial-temporal attention (STA), correlation-inspired similarity constraint ($\mathcal{L}_{CC}$) and complementarity constraint and ($\mathcal{L}_{Recon}$), adversarial classifier training (Adv-Cls) and component-wise supervised contrastive learning (C-SupCon). To further verify the effectiveness of each module of our proposed model, we introduce several variants of our model for ablation experiments on the three datasets. Besides, we conduct a baseline experiment of retaining only a single encoder to demonstrate the necessity of the dual encoder structure in our model. The experimental results are shown in Fig. 7. Specifically, we utilize several simplified versions of our proposed model:

- **w/o STA:** The model removes the spatial-temporal attention module.
- **w/o SCC:** The model removes the correlation-inspired similarity and complementarity constraint (SCC).
- **w/o BiasEnc:** The model removes the bias encoder (BiasEnc) with only the identity encoder.
- **w/o Adv-Cls:** The model without adversarial classifier training.
- **w/o C-SupCon:** The model without component-wise supervised contrastive learning.

The experimental results illustrated in Fig. 7(a), (b) and (c) show the ablation results on the PhysioNet dataset, SEED-Joint

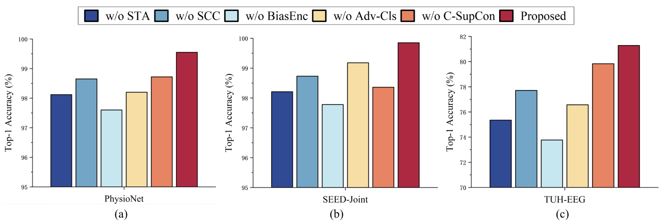


Fig. 7. The ablation study of our model on the PhysioNet dataset, SEED-Joint dataset, and TUH-EEG dataset, respectively, where the vertical axis represents the Top-1 Acc (%).

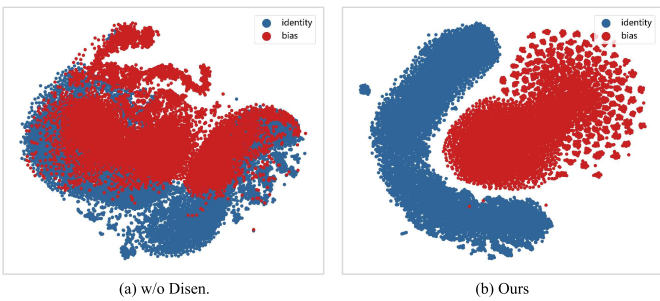


Fig. 8. The t-SNE visualization of the identity and bias representations of our method. (a) denotes the representations generated with the model w/o Disen., (b) is the representations extracted by our model.

dataset and TUH-EEG dataset, respectively, which demonstrate that our proposed model attains the best identification and authentication performance by effectively integrating all modules across different datasets. First, the model with STA outperforms the baseline model, showing the STA's ability to extract spatial-temporal patterns. Second, the complementarity of identity features and bias features is maintained by applying $\mathcal{L}_{CC}$ and $\mathcal{L}_{Recon}$ for the disentanglement of entangled factors, leading to better recognition performance. Besides, the bias features can be extracted without task labels, and the interference noise irrelevant to biometric recognition can be removed from EEG signals to a certain extent with the help of the adversarial classifier in our proposed model. Furthermore, the ablation experiment of the bias encoder evaluates the effectiveness of the dual encoder structure in our proposed framework. Finally, the C-SupCon enables the model to better disentangle the identity factor and bias factor within EEG signals to learn a more robust identity-related feature, thus achieving better identification performance.

To verify the disentanglement performance of our proposed framework, we remove the disentanglement losses, i.e., $\mathcal{L}_{CC}$, $\mathcal{L}_{Recon}$, and $\mathcal{L}_{C-SupCon}$, and adversarial training, denoted as w/o Disentanglement (Disen.) for further ablation experiments. Firstly, we visualize the feature extracted by w/o Disen. and our model on the PhysioNet dataset, and the result shown in Fig. 8 demonstrates that our model effectively disentangles the identity and bias factors, thus enhancing the model's performance. Furthermore, we employ Mutual Information (MI) and the Correlation Coefficient (CC) to quantify the separation between identity and bias features. As illustrated in Table IV, our proposed method achieves significant reductions in both MI and CC across the three datasets. These results confirm its strong disentanglement capability, thereby contributing to improved person identification accuracy.

TABLE IV

MUTUAL INFORMATION AND CORRELATION COEFFICIENT BETWEEN THE IDENTITY AND BIAS REPRESENTATIONS GENERATED WITH THE MODEL W/O DISEN, AND THE REPRESENTATIONS EXTRACTED BY OUR MODEL

Model	Mutual Information (MI)			Correlation Coefficient (CC)		
	PhysioNet	SEED-Joint	TUH-EEG	PhysioNet	SEED-Joint	TUH-EEG
w/o Disen.	0.580	0.526	0.695	0.417	0.440	0.775
Ours	**0.092**	**0.068**	**0.129**	**0.075**	**0.061**	**0.151**

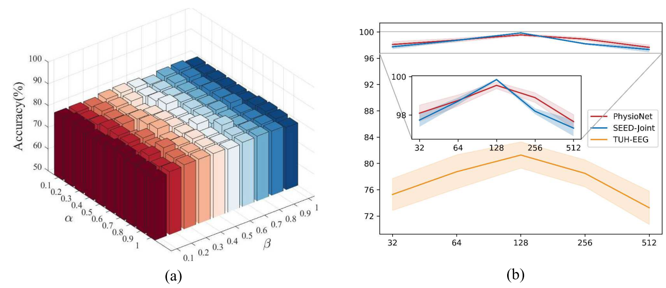


Fig. 9. The parameter sensitive analysis of the proposed method on the TUH-EEG dataset. (a) the weight of α and β on the TUH-EEG dataset, (b) the embedding size on the three datasets.

E. Parameter Settings and Sensitivity Analysis

We further perform extensive experiments to evaluate the sensitivity of the hyperparameters of our model. Firstly, in our proposed framework, we introduce hyperparameters α and β to control the weight of $\mathcal{L}_{CC}+\mathcal{L}_{Recon}$ and $\mathcal{L}_{C-SupCon}$ in the overall loss function, respectively. We tune over the hyperparameters α and β with the value $\{0.1, 0.2, 0.3, \ldots, 1\}$, grid search is adopted to determine the optimal parameters through the validation Top-1 Acc under the combinations of these two hyperparameters, and the result on the TUH-EEG dataset is shown in Fig. 9(a). As can be seen from the figure, we can find that the classification results were relatively stable, when the two parameters were slightly tuned. In particular, the proposed method achieved better recognition performance when α and β are 0.4 and 0.6. Besides, we choose the embedding size over the set $\{32, 64, 128, 256, 512\}$ on the PhysioNet dataset. The results on the PhysioNet, SEED-Joint, and TUH-EEG datasets are shown in Fig. 9(b). The performance of our proposed method initially improves with an increase in embedding size due to enhanced fitting ability. However, if the embedding size becomes too large, it can lead to overfitting, causing a degradation in identification performance.

F. Visualizations

To further investigate the effectiveness of our proposed method, we visualize the learned features of our model and several comparison methods with t-SNE [55] for the first 100 subjects on the TUH-EEG dataset, shown in Fig. 10. We first downscale the representation vectors learned by our method with t-SNE and then project the two-dimensional vectors onto a space for visualization as shown in Fig. 10(j), and

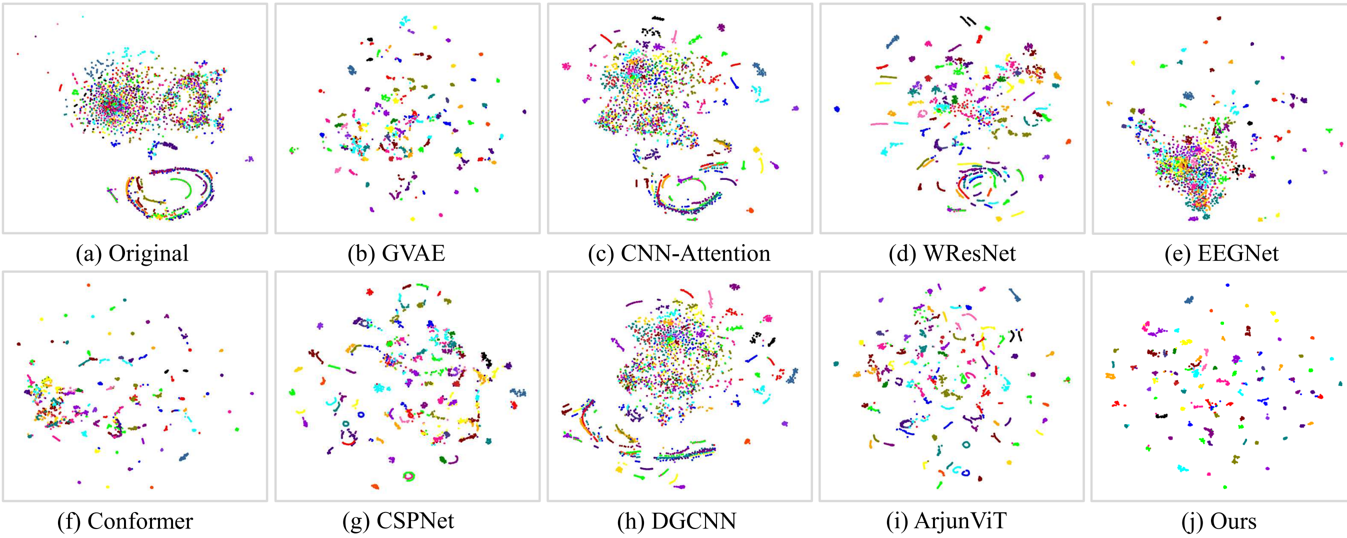


Fig. 10. The t-SNE visualization of the distribution of original data and features extracted by different deep learning models on the TUH-EEG dataset.

the results of the original sample data and other comparison methods are shown in Fig. 10 (a)-(i). As can be seen from Fig. 10, first, compared with the original sample data, the features extracted by CNN-Attention, WResNet, EEGNet, and DGCNN are highly mixed without clear decision boundaries, as they fail to capture complex spatial-temporal patterns. While GVAE, Conformer, CSPNet, and ArjunViT show a better performance in integrating intra-class samples, there are still some cases in which the subjects are difficult to distinguish. Existing methods are limited by their failure to address entangled factors in EEG signals, which compromises identification performance. In contrast, our model not only mines complex patterns but, more importantly, disentangles identity information from detrimental biases, leading to more compact intra-subject clusters, greater inter-subject separation, and superior Top-1 accuracy.

V. Conclusion

In this paper, we present a generalized framework for feature disentanglement designed to extract EEG-based brainprint information and identity-bias information, specifically those arising from task-related information, disease indicators, and signal noise. First, a novel spatial-temporal attention is developed to extract the complex spatial-temporal patterns within EEG signals. Second, a correlation-driven loss is implemented to minimize the similarity between identity and bias representations for disentanglement. Then, an adversarial classifier is trained to maximize the bias encoder's ability to extract bias features within EEG signals, and a reconstruction loss is adopted to constrain the complementarity of identity-bias features. Furthermore, we implement SupCon at the component level to further disentangle the identity-related information and identity-irrelevant information, thus obtaining a more robust feature representation for biometric recognition. The experimental results on three datasets show that our proposed method not only achieves promising recognition results for EEG biometrics compared to state-of-the-art methods but also explicitly extracts robust identity-related features under distinct types of interference, including emotion, motor imagery, and brain disease characteristics.

References

[1] Y. Huang et al., "CurricularFace: Adaptive curriculum learning loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5901–5910.

[2] G. Shen, J. Zhang, A. Marshall, L. Peng, and X. Wang, "Radio frequency fingerprint identification for LoRa using spectrogram and CNN," in *Proc. IEEE Conf. Comput. Commun.*, May 2021, pp. 1–10.

[3] Z. Yang, H. Huangfu, L. Leng, B. Zhang, A. B. J. Teoh, and Y. Zhang, "Comprehensive competition mechanism in palmprint recognition," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 5160–5170, 2023.

[4] C. Wang, J. Muhammad, Y. Wang, Z. He, and Z. Sun, "Towards complete and accurate iris segmentation using deep multi-task attention network for non-cooperative iris recognition," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2944–2959, 2020.

[5] E. Debie, N. Moustafa, and A. Vasilakos, "Session invariant EEG signatures using elicitation protocol fusion and convolutional neural network," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 4, pp. 2488–2500, Jul. 2022.

[6] S. N. A. Seha and D. Hatzinakos, "EEG-based human recognition using steady-state AEPs and subject-unique spatial filters," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3901–3910, 2020.

[7] M. V. Ruiz-Blondet, Z. Jin, and S. Laszlo, "CEREBRE: A novel method for very high accuracy event-related potential biometric identification," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 7, pp. 1618–1629, Jul. 2016.

[8] M. Wang, S. Wang, and J. Hu, "Cancellable template design for privacy-preserving EEG biometric authentication systems," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 3350–3364, 2022.

[9] T. Behrouzi and D. Hatzinakos, "Graph variational auto-encoder for deriving EEG-based graph embedding," *Pattern Recognit.*, vol. 121, Jan. 2022, Art. no. 108202.

[10] T. Wilaiprasitporn, A. Ditthapron, K. Matchaparn, T. Tongbuasirilai, N. Banluesombatkul, and E. Chuangsuwanich, "Affective EEG-based person identification using the deep learning approach," *IEEE Trans. Cognit. Develop. Syst.*, vol. 12, no. 3, pp. 486–496, Sep. 2020.

[11] W. Li, Y. Yi, M. Wang, B. Peng, J. Zhu, and A. Song, "A novel tensorial scheme for EEG-based person identification," *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 2500717.

[12] X. Kong, W. Kong, Q. Fan, Q. Zhao, and A. Cichocki, "Task-independent EEG identification via low-rank matrix decomposition," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 412–419.

[13] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdogmus, "Adversarial deep learning in EEG biometrics," *IEEE Signal Process. Lett.*, vol. 26, no. 5, pp. 710–714, May 2019.

[14] L. Bollens, T. Francart, and H. V. Hamme, "Learning subject-invariant representations from speech-evoked EEG using variational autoencoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 1256–1260.

[15] X. Jin et al., "Disentangled adversarial generalization network for cross-session task-independent brainprint recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 16, no. 4, pp. 1248–1259, Aug. 2024.

[16] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, Jul. 2020.

[17] X. Du et al., "An efficient LSTM network for emotion recognition from multichannel EEG signals," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1528–1540, Jul. 2022.

[18] 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, 2022.

[19] Y. Ding, N. Robinson, S. Zhang, Q. Zeng, and C. Guan, "TSception: Capturing temporal dynamics and spatial asymmetry from EEG for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2238–2250, Jul. 2023.

[20] E. Creager et al., "Flexibly fair representation learning by disentanglement," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 1436–1445.

[21] N. Hadad, L. Wolf, and M. Shahar, "A two-step disentanglement method," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 772–780.

[22] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2642–2651.

[23] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 343–351.

[24] S. Xu, Z. Zhou, and J. Shang, "Asymmetric adversarial-based feature disentanglement learning for cross-database micro-expression recognition," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 5342–5350.

[25] E. H. Sanchez, M. Serrurier, and M. Ortner, "Learning disentangled representations via mutual information estimation," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, 2020, pp. 205–221.

[26] W. Xu, H. Liu, W. Shi, Z. Miao, Z. Lu, and F. Chen, "Adversarial feature disentanglement for long-term person re-identification," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1201–1207.

[27] Z. Shu, Y. Luo, Y. Huang, C. Mao, and Z. Yu, "View-interactive attention information alignment-guided fusion for incomplete multi-view clustering," *Expert Syst. Appl.*, vol. 252, Oct. 2024, Art. no. 124258.

[28] T. Sun, Z. Shu, Y. Huang, H. Wang, and Z. Yu, "Semantic feature graph consistency with contrastive cluster assignments for multilingual document clustering," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 24, no. 1, pp. 1–22, Jan. 2025.

[29] Z. Shu, T. Sun, Y. Luo, and Z. Yu, "Ambiguous instance-aware contrastive network with multi-level matching for multi-view document clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2025, vol. 39, no. 19, pp. 20479–20487.

[30] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.

[31] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193907–193934, 2020.

[32] P. Khosla et al., "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 18661–18673.

[33] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[34] S. Lee, T.-H. Pham, and P. Zhang, "DREAM: Domain invariant and contrastive representation for sleep dynamics," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2022, pp. 1029–1034.

[35] T. K. K. Ho and N. Armanfard, "Self-supervised learning for anomalous channel detection in EEG graphs: Application to seizure analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 7866–7874.

[36] E. Gibson, N. J. Lobaugh, S. Joordens, and A. R. McIntosh, "EEG variability: Task-driven or subject-driven signal of interest?," *NeuroImage*, vol. 252, May 2022, Art. no. 119034.

[37] R. Labaca-Castro, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 73–76.

[38] C. Yang et al., "ATD: Augmenting CP tensor decomposition by self supervision," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2021, pp. 32039–32052.

[39] C. Yang, M. B. Westover, and J. Sun, "ManyDG: Many-domain generalization for healthcare applications," in *Proc. 11th Int. Conf. Learn. Represent.*, 2023, pp. 1–24.

[40] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[41] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "BCI2000: A general-purpose brain–computer interface (BCI) system," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1034–1043, Jun. 2004.

[42] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auto. Mental Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[43] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "EmotionMeter: A multimodal framework for recognizing human emotions," *IEEE Trans. Cybern.*, vol. 49, no. 3, pp. 1110–1122, Mar. 2019.

[44] I. Obeid and J. Picone, "The temple university hospital EEG data corpus," *Frontiers Neurosci.*, vol. 10, May 2016, Art. no. 195498.

[45] T. Xu et al., "E-key: An EEG-based biometric authentication and driving fatigue detection system," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 864–877, Apr. 2023.

[46] M. Wang, H. El-Fiqi, J. Hu, and H. A. Abbass, "Convolutional neural networks using dynamic functional connectivity for EEG-based person identification in diverse human states," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 12, pp. 3259–3272, Dec. 2019.

[47] J. Jin, Z. Chen, H. Cai, and J. Pan, "Affective EEG-based person identification with continual learning," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–16, 2024.

[48] A. Seyfizadeh, R. L. Peach, P. Tovote, I. U. Isaias, J. Volkmann, and M. Muthuraman, "Enhancing security in brain–computer interface applications with deep learning: Electroencephalogram-based user identification," *Expert Syst. Appl.*, vol. 253, Nov. 2024, Art. no. 124218.

[49] 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, Oct. 2018, Art. no. 056013.

[50] H. Altaheri, G. Muhammad, and M. Alsulaiman, "Physics-informed attention temporal convolutional network for EEG-based motor imagery classification," *IEEE Trans. Ind. Informat.*, vol. 19, no. 2, pp. 2249–2258, Feb. 2023.

[51] R. Mane et al., "FBCNet: A multi-view convolutional neural network for brain–computer interface," 2021, *arXiv:2104.01233*.

[52] X. Jiang, L. Meng, X. Chen, Y. Xu, and D. Wu, "CSP-Net: Common spatial pattern empowered neural networks for EEG-based motor imagery classification," *Knowl.-Based Syst.*, vol. 305, Dec. 2024, Art. no. 112668.

[53] A. Arjun, A. S. Rajpoot, and M. R. Panicker, "Introducing attention mechanism for EEG signals: Emotion recognition with vision transformers," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 5723–5726.

[54] M. Jin, C. Du, H. He, T. Cai, and J. Li, "PGCN: Pyramidal graph convolutional network for EEG emotion recognition," *IEEE Trans. Multimedia*, vol. 26, pp. 9070–9082, 2024.

[55] L. V. D. Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.