EEG Emotion Recognition With Uncertainty-Aware Contrastive Learning and Frequency-Aware Self-Attention

Xu Xu, *Graduate Student Member, IEEE*, Junxin Chen, *Senior Member, IEEE*, Qiang He, *Member, IEEE*, Yongfei Wu, *Member, IEEE*, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*—Electroencephalography (EEG) emotion recognition plays a key role in improving human–machine interactions. Advanced algorithms have been proposed for this task. However, two challenges remain, i.e., unclear decision boundary in the embedded space and noise in physiological signals from various devices. To this end, we develop a novel framework, namely, UACL-Net, for EEG emotion recognition. It is based on uncertainty-aware contrastive learning (UACL) and frequency-aware self-attention (FASA). Specifically, UACL uses a multivariate Gaussian distribution to construct the latent space for different emotions. It is able to highlight interclass differences, thereby improving the robustness of model decisions. In addition, FASA generates learnable weights by applying self-attention (SA) to the real and imaginary components in the frequency domain. This helps adaptively reduce noise and capture global dependencies in temporal sequences. Our model is trained and tested on four benchmark datasets, achieving up to 94.88%, 98.71%, 96.91%, and 99.29% accuracy on SEED, DEAP, DREAMER, and FACED, respectively. Experimental results demonstrate that it is effective and has advantages over peer state-of-the-art (SOTA) methods.**

***Index Terms*—Contrastive learning, emotion recognition, frequency-aware attention, uncertainty confidence.**

I. Introduction

With the development of deep learning, emotion recognition is crucial for building friendly human–machine interactions [1]. It can be implemented from video, audio, texture, and physiological signals [such as electroencephalography (EEG) and electrocardiogram (ECG)] [2]. Among them, physiological signals can directly reflect the potential human nervous system with more reliable characteristics [3]. Specifically, EEG characterizes the neural activity in a more direct manner [4]. Following this, many studies have developed networks for EEG emotion recognition, which can be divided into two categories.

The first type typically relies on supervised learning to extract features and perform decision-making. Typically, Chen et al. [5] fused functional attention with prior spatial connections, then performed deep node-graph representation learning to capture long-range interactions and reduce oversmoothing. Yan et al. [6] developed a graph neural network (GNN) to improve feature distinctiveness by adaptively combining each convolution output. Chen et al. [7] proposed a graph disentanglement module to extract common and specific characteristics from both EEG graph connectivity and representation. However, these methods depend on a large amount of manual labeling.

To mitigate the above issues, the other type uses self-supervised learning (SSL) to encode EEG signals into embedded spaces. For instance, Kan et al. [8] proposed an SSL framework that incorporates advanced group data augmentation techniques to reduce the impact of random fluctuations. Ko et al. [9] developed an SSL architecture that uses feature normalization to mitigate intra/intersubject variability for robust EEG decoding. Zhang et al. [10] designed another SSL framework that can generate simulated EEG samples with high quality and diversity. However, two challenges remain unsolved, i.e., unclear decision boundary in the embedded space, and noise in physiological signals from various devices. Unclear decision boundaries stem in part from a low signal-to-noise ratio (due to muscle and environmental artifacts). As a result, feature clusters for adjacent emotional states (e.g., arousal and neutral) overlap in the learned feature space, reducing classifier accuracy.

Recently, the development of probabilistic representation and frequency-aware attention has introduced new approaches to address these challenges. First, probabilistic representation has been successfully applied to various computer vision tasks. Typically, Shi and Jain [11] represented each facial image as a Gaussian distribution, using the mean to estimate likely feature values and the variance to capture uncertainty.

Received 17 October 2025; revised 21 December 2025 and 28 January 2026; accepted 12 February 2026. Date of publication 25 February 2026; date of current version 18 March 2026. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant DUT25YG229, in part by Xiaomi Young Talents Program, and in part by the Youth Talent Support Program for Doctoral Students of China Association for Science and Technology. This article was recommended by Associate Editor M. Feroskhan. *(Corresponding author: Junxin Chen.)*

Xu Xu is with the School of Software, Dalian University of Technology, Dalian 116621, China, and also with the School of Computer Science and Engineering, Northeastern University, Shenyang 110004, China.

Junxin Chen is with the School of Software, Dalian University of Technology, Dalian 116621, China (e-mail: junxinchen@ieee.org).

Qiang He is with the College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110004, China.

Yongfei Wu is with the College of Artificial Intelligence, Taiyuan University of Technology, Taiyuan 030024, China.

Yicong Zhou is with the Department of Computer and Information Science, University of Macau, Macau, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2026.3665442.

Digital Object Identifier 10.1109/TCYB.2026.3665442

2168-2267 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Universidade de Macau. Downloaded on March 20,2026 at 23:55:09 UTC from IEEE Xplore. Restrictions apply.

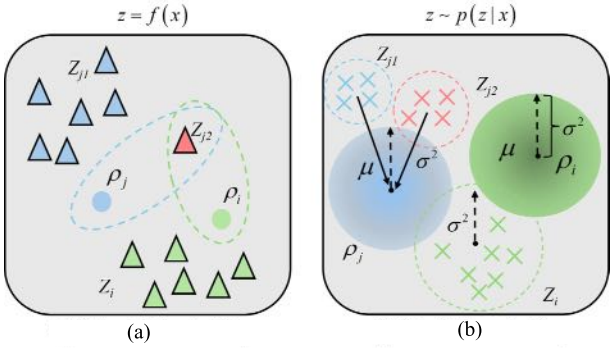

Fig. 1. Difference between (a) deterministic and (b) probabilistic embedding.

Based on the above model, Xie et al. [12] used multivariate Gaussian distributions for pixel-wise representations in image segmentation. On the other hand, some studies have shown that frequency-related attention is able to reduce noise impacts and capture global dependencies in temporal sequences. For instance, Xu et al. [13] designed a frequency enhancement branch to adaptively modulate the frequency information for imagined speech production. Zhou et al. [14] proposed a Fourier enhanced blocks to capture important structures in time series through frequency domain mapping.

Motivated by the above discussion, we develop a novel framework, namely, UACL-Net, for EEG emotion recognition. It is based on uncertainty-aware contrastive learning (UACL) and frequency-aware self-attention (FASA). First, contrastive learning with uncertainty-aware loss is designed to highlight interclass differences. It is illustrated in Fig. 1. Specifically, we use a multivariate Gaussian distribution to represent each category x_i, mapping it to a point z_i in the latent space, expressed as $z \sim p(z|x)$. This distribution includes the mean μ and variance σ^2. In particular, similar emotion categories x_{j1} and their label boundaries x_{j2} map to points z_{j1} and z_{j2}, respectively. When z_i is equidistant from both prototypes ρ_i and ρ_j, deterministic embedding $z = h(x)$ struggle to decide whether to map z_i to ρ_i or ρ_j. However, z_i consistently maps to ρ_i because ρ_i has a lower σ^2 than ρ_j in probabilistic embedding. Moreover, FASA is proposed to generate learnable weights with self-attention (SA) to the real and imaginary components in the frequency domain. It is able to adaptively reduce noise and capture global dependencies in temporal sequences.

Although UACL and FASA are inspired by probabilistic embeddings and frequency-domain transforms, they introduce substantial novelties. UACL represents both samples and class prototypes as Gaussian distributions and incorporates uncertainty-aware (confidence-weighted) prototype alignment, which reduces the influence of unreliable trials and alleviates prototype drift under noise or label corruption. FASA departs from heuristic band reweighting by applying SA to complex FFT features (real and imaginary parts) followed by inverse FFT (IFFT) reconstruction, enabling data-driven spectral emphasis and global dependency modeling within a unified block. To the best of our knowledge, this is the first study that jointly integrates probabilistic prototype-based contrastive learning and frequency-domain SA for EEG emotion recognition. Five benchmark datasets, i.e., SEED [15], DREAMER [16], DEAP [17], FACED [18], and AMIGOS [19], are utilized to train and test our framework. Their labels are based on different grain levels, which can cover the evaluation from coarse to fine. Extensive results show that UACL-Net outperforms other state-of-the-art (SOTA) algorithms, proving its effectiveness. Key contributions are listed as follows.

1) A novel framework is developed with UACL and FASA for EEG emotion recognition.
2) UACL has been found effective in highlighting interclass differences and further improving the recognition performance.
3) FASA has been found effective in adaptively reducing noise and capturing global dependencies in temporal sequences.

This article is organized as follows. First, Section II reviews past research on EEG emotion recognition. Next, Section III details our modules and network design, while Section IV explores our datasets and implementation methods. Following this, Section V discusses performance evaluation and provides an in-depth analysis of our work. Finally, Section VI summarizes the article and outlines our future research directions.

II. Related Work

A. EEG Emotion Recognition

EEG signals are nonstationary time series that contain rich temporal, spatial, and frequency information. Recently, many deep learning-based studies have focused on EEG emotion recognition. They can be divided into three types based on the utilized model, i.e., convolutional neural network (CNN) [24], [28], [33], time-series neural network (TNN) [22], [30], and hybrid model [25], [40], [41]. For example, Jin et al. [33] created an advanced graph convolutional network to aggregate features at three levels, thereby improving relationship modeling across the scalp. Xu et al. [22] created a network based on the Transformer architecture that utilizes a multidimensional global attention mechanism to exploit the complementary nature of temporal features in EEG data. In addition, some studies have explored hybrid models for emotion recognition. This is because different models are able to focus on different features. Typically, Li et al. [40] proposed a multiview network that jointly captures useful information, aiming to encode more discriminative features describing emotions. Wei et al. [41] combined CNN and Transformer to capture global features and extract local features, which are characterized by spatial relationships.

B. Contrastive Learning

Contrastive learning is able to map encoder output features to a latent space using a representation head. It can be divided into two types. The first type designs a specific strategy to constrain and push the distance between the representations of two instances [42]. The instances that need to be constrained and pushed are the positive sample and the negative sample, respectively. Well-known SSL algorithms include SimCLR [43], MoCo [44], and BYOL [45]. The other type focuses on semantic information in the training process. Specifically,

Authorized licensed use limited to: Universidade de Macau. Downloaded on March 20,2026 at 23:55:09 UTC from IEEE Xplore. Restrictions apply.

TABLE I
OVERVIEW OF TYPICAL WORK FOR EEG EMOTION RECOGNITION

Author	Year	Abbreviation	Model Type	Key Technology	Datasets
Yin *et al.* [20]	2021	ECLGCNN	Hybrid Network	Differential Entropy, Graph Domain Features	DEAP [17]
Pandey *et al.* [21]	2022	N/A	Hybrid Network	Variational Mode Decomposition	DEAP [17]
Xu *et al.* [22]	2023	AMDET	Transformer	Multi-Information Fusion, Self-Attention	DEAP [17], SEED [15], SEED-IV [23]
Liu *et al.* [24]	2023	GLFA-Net	GNN	Undirected Topology Graph	DEAP [17], SEED [15], DREAMER [16]
Zhou *et al.* [25]	2023	MTHNet	Hybrid Network	Multi-Information Fusion, Multi-Task Hybrid	SEED [15], SEED-IV [23], SEED-V [26], SEED-FRA [27]
Wang *et al.* [28]	2023	N/A	CNN	Self-Supervised Learning, Multi-Task Hybrid	DEAP [17], SEED [15]
Fan *et al.* [29]	2024	N/A	Hybrid Network	Information Fusion, Residual Connections	DEAP [17], DREAMER [16]
Ding *et al.* [30]	2024	MASA-TCN	TCN	Space-Aware Temporal Layer	DEAP [17], MAHNOB-HCI [31]
Yan *et al.* [6]	2024	BGAGCN	GCN	Multi-Head Attention, Graph Domain Features	SEED [15], SEED-IV [23], DREAMER [16]
Chen *et al.* [7]	2024	GDDN	GCN	Auxiliary Supervision, Graph Domain Features	SEED [15], SEED-IV [23], MPED [32]
Jin *et al.* [33]	2024	PGCN	GCN	Multi-Level Fusion, Graph Domain Features, 3D Spatial Distances	SEED [15], SEED-IV [23], MPED [32]
Xu *et al.* [34]	2024	EFSMDER	Hybrid Network	Multi-Dimension Input, Information Fusion	DEAP [17], DREAMER [16], HDED [35]
An *et al.* [36]	2025	LGDAAN-Nets	Hybrid Network	Multi-Feature Extraction, Local-Global Domain Discriminator	SEED [15], SEED-IV [23], DEAP [17]
Ding *et al.* [37]	2025	EmT	Hybrid Network	Temporal Graph Construction, Temporal Contextual Transformer	SEED [15], THU-EP [38], FACED [18], MAHNOB-HCI [31]
Liu *et al.* [39]	2025	MAS-DGAT-Net	Hybrid Network	Cross-Feature Fusion, Broad Learning System	SEED [15], SEED-IV [23], DEAP [17]
Ours	2026	UACL-Net	Hybrid Network	Uncertainty-Aware Contrastive Learning, Frequency-Aware Self-Attention	SEED [15], DREAMER [16], DEAP [17], FACED [18], AMIGOS [19]

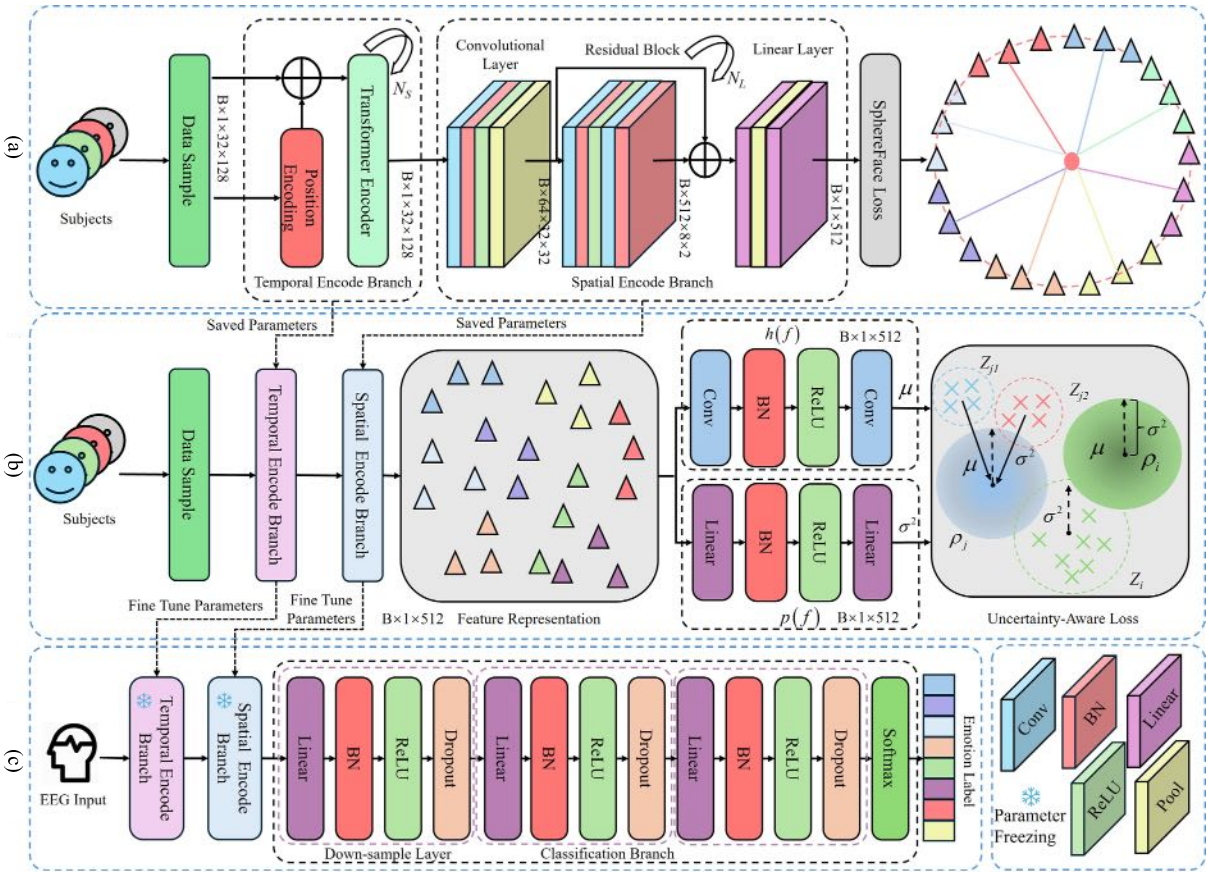

Fig. 2. Architecture of our framework. (a) Stage 1: pretrain learning. (b) Stage 2: uncertainty-aware fine-tune learning. (c) Stage 3: classification. Conv: convolutional. BN: batch normalization. Pool: pooling with a 1-D kernel. The latent embedding Z and its probabilistic parameters (μ, σ^2) are all 512-D.

each pixel is mapped and normalized into a representation domain. Among them, the same representations are aggregated, and different representations are formed. Regarding the instance segmentation, different instances correspond to different semantic information. The model is prone to classifying instance edge domains into the wrong category, which in turn leads to confusion in the representation domains. Many studies have focused on addressing this problem [46], [47]. Typically, Xie et al. [47] introduce pixel-level pretext tasks for learning feature representations.

Authorized licensed use limited to: Universidade de Macau. Downloaded on March 20,2026 at 23:55:09 UTC from IEEE Xplore. Restrictions apply.

C. Uncertainty Embedding

Uncertainty embedding commonly models feature embedding as a probabilistic distribution, such as the Gaussian distribution [48]. Following this, variational autoencoders (VAEs) are developed to calculate the mean and variance of these distributions. This method is now used in several computer vision applications [11], [12], [49]. For example, Li et al. [49] applied this technique to map facial images to a von Mises–Fisher distribution, which includes mean and concentration parameters. Traditional distance metrics, which usually compare distributions based on their means, do not effectively capture probabilistic similarities because they overlook variances. To address this, Kingma and Welling [50] used the reparameterization trick. This involves generating two sets of samples from the distributions using Monte–Carlo sampling and then combining the similarities of these samples to more accurately represent the similarities of the distributions.

III. Proposed Method

A. Framework Architecture

In the first stage, the EEG input $x \in R^{1\times 32\times 128}$ is fed into the data sampler to construct paired samples, where 32 is the number of channels, and 128 is the number of time points in a window. The samples are input into the teacher network with SphereFace loss [51] to generate feature representation. This network consists of a temporal encode branch (TEB) $T(\cdot)$ and a spatial encode branch (SEB) $S(\cdot)$, and is pretrained by Adam to minimize the loss function $Loss_s$ as

$$Loss_s = l_{sf}\left(S\left(T\left(x\right)\right), y\right) \tag{1}$$

where l_{sf} is SphereFace loss [51]. In addition, $y \in R^{1\times n}$ represents the one-hot encoding of the corresponding ground truth class label, where n is the number of emotion types. SphereFace loss normalizes both the input feature vectors and the class prototypes to lie on a unit hypersphere. It incorporates an angular margin to force the decision boundaries between classes to be farther apart. Therefore, we introduce this loss to extract feature representation $Z \in R^{1\times 512}$.

In the fine-tuning stage, we further optimize feature representation Z to improve the interclass differences. Specifically, two heads, i.e., the representation head $h(\cdot)$ and the probability head $p(\cdot)$, are developed to map the mean $\mu \in R^{1\times 512}$ and variance $\sigma_i^2 \in R^{1\times 512}$ from feature representation, respectively. The latent embedding Z and its probabilistic parameters (μ, σ^2) are all 512-D, ensuring a shared latent space for subsequent learning. Among them, $h(\cdot)$ consists of the convolution, batch normalization, ReLU, and convolution. In addition, $p(\cdot)$ consists of the fully connected layer, batch normalization, ReLU, and a fully connected layer. This distribution is constrained by the infoNCE $Loss_u$ [52] as

$$Loss_u = -\frac{1}{|C|\times|Z_c|}\sum_{c\in C}\sum_{z_{ci}\in Z_c} \times \log\left[\frac{e^{s(z_{ci},\rho_c)/\tau}}{e^{s(z_{ci},\rho_c)/\tau} + \sum_{\tilde{c}\in\tilde{C}}\sum_{\tilde{z}\in\tilde{Z}} e^{s(z_{ci},z_{\tilde{c}i})/\tau}}\right] \tag{2}$$

where Z_{ci} and $Z_{\tilde{c}i}$ are the sets including positive and negative sampled features representations belonging to anchor class C, respectively. In addition, $\tilde{C}$ and τ denote the set of classes other than c and the temperature control of the distribution softness. Particularly, s is a similarity measurement function, which is developed as a mutual likelihood score (MLS) [11].

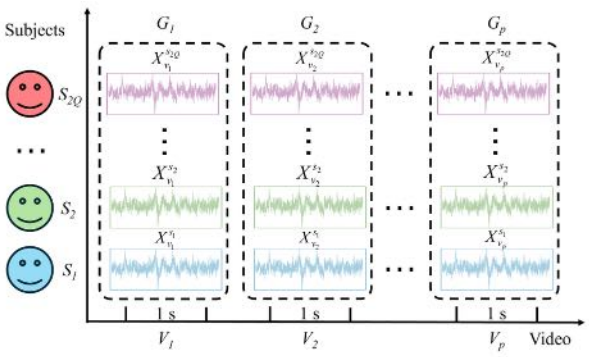

Fig. 3. Architecture of data sampler.

In the classification stage, we freeze the model parameters of TEB and SEB from the fine-tuning stage, and optimize the model parameters of the classification branch $C(\cdot)$ with cross-entropy loss l_{ce} [53] during training. The training label y of this loss is one-hot encoded for emotion classifications. This processing can be defined as

$$Loss_c = l_{ce}\left(C\left(S\left(T\left(x\right)\right)\right), y\right). \tag{3}$$

After training, the model parameters of TEB, SEB, and the classification branch are saved for model inference. First, the EEG input is fed into the TEB and SEB to extract feature representation. Then, this information is fed into the classification branch to generate the predicted emotion type. The cores of our framework, i.e., data sampler, UACL, and network architecture, are detailed in the following paragraphs (see Section II).

B. Data Sampler

This work develops a strategy to extract samples from individual EEG samples and is illustrated in Fig. 3. Each EEG sample is represented as $X_v^s \in R^{M\times C}$, capturing a 1-s signal from Subject s while watching video v. Here, M denotes the temporal length of the signal, and C indicates the number of electrode channels. Initially, the sampler randomly selects P video clips to construct a batch, labeled $v_1, v_2, \ldots, v_p$. It also selects $2Q$ subjects from two equal groups, labeled $s_1, s_2, \ldots, s_{2Q}$, to create pairs for each clip. Then, EEG signals are extracted from the chosen subjects and different clips, assembling a dataset $D = \{X_{v_i}^{s_k} | i = 1, 2, \ldots, P; k = 1, 2, \ldots, 2Q\}$. In addition, a group of samples $G_i = \{X_{v_i}^{s_1}, X_{v_i}^{s_2}, \ldots, X_{v_i}^{s_{2Q}}\}$ for each video clip v_i is contracted with similar characteristics.

A data augmentation is developed to enhance a group sample while maintaining the same clip-related features. It includes three steps, i.e., individual pairing, crossover, and separation. Regarding the original signals in the group $G_i = \{X_{v_i}^{s_k} | k = 1, 2, \ldots, 2Q\}$, individual signals are randomly paired to from Q pairs $\{X_{v_i}^{s_Q}, X_{v_i}^{s_{2Q}}\}$. Then, the split position N is randomly given and transformed to obtain pairs of signals

Authorized licensed use limited to: Universidade de Macau. Downloaded on March 20,2026 at 23:55:09 UTC from IEEE Xplore. Restrictions apply.

$\{\tilde{X}_{v_i}^{s_k}, \tilde{X}_{v_i}^{s_{k+Q}}\}$. Finally, the transformed signals are divided and paired into two groups, A and B, according to the split positions as

$$\tilde{G}_i^A = \left\{\tilde{X}_{v_i}^{s_k} | k = 1, 2, \ldots, Q\right\}$$
$$\tilde{G}_i^B = \left\{\tilde{X}_{v_i}^{s_k} | k = Q+1, Q+2, \ldots, 2Q\right\} \quad (4)$$

where $\tilde{G}_i^A \in R^{1\times 16\times 128}$ and $\tilde{G}_i^B \in R^{1\times 16\times 128}$ are combined as the EEG input $x \in R^{1\times 32\times 128}$.

C. Temporal Encode Branch

As we all know, Transformer is a typical sequence model and widely used to extract temporal information [54]. Based on the Transformer, we develop TEB that consists of a position encoder (PE) and frequency-aware Transformer encoder (FATE). First, the EEG signal $x \in R^{1\times 32\times 128}$ is fed into the PE to help the model understand the data layout. This encoder uses sine and cosine functions to generate unique vectors for each position as

$$PE_{(pos,2i)} = \sin\left(pos \big/ 10\,000^{2i/d}\right)$$
$$PE_{(pos,2i+1)} = \cos\left(pos \big/ 10\,000^{2i/d}\right) \quad (5)$$

where pos and i represent the position and dimension, respectively. Its output $F_{PE} \in R^{1\times 32\times 128}$ can be defined as

$$F_{PE} = Dropout\left(PE_{(pos,2i)}(x) + PE_{(pos,2i+1)}(x)\right). \quad (6)$$

Its output $F_{PE} \in R^{1\times 32\times 128}$ merges the original EEG signal x as $F_C \in R^{1\times 32\times 128}$, which is then input into the FATE to capture global dependencies in temporal sequences efficiently. This encoder is illustrated in Fig. 4. The process involves transforming the input sequence into the frequency domain via the fast Fourier transform (FFT), applying SA to the real and imaginary components, and then reconstructing the output in the time domain through the IFFT. Specifically, F_C is first transformed along the temporal dimension using the FFT as

$$F_f = \mathrm{FFT}(F_C, \mathrm{dim} = 1) \in R^{1\times 32\times 128}. \quad (7)$$

This results in complex-valued frequency representations, which are separated into real and imaginary components as

$$F_f = \mathcal{R}(F_f) + i \cdot \mathcal{I}(F_f). \quad (8)$$

Next, the real and imaginary parts are concatenated along the feature dimension to form a real-valued tensor suitable for standard attention modules as

$$F_{\mathrm{attn}} = \left[\mathcal{R}(F_f) \,\|\, \mathcal{I}(F_f)\right] \in R^{1\times 64\times 128}. \quad (9)$$

Multihead SA [54] is then applied to the concatenated representation, enabling the model to capture dependencies across both the real and imaginary parts of the frequency representation as

$$F_{\mathrm{attn_out}} = \mathrm{MultiHeadAttn}(F_{\mathrm{attn}}, F_{\mathrm{attn}}, F_{\mathrm{attn}}). \quad (10)$$

The output is split back into real and imaginary components

$$\mathcal{R}(\tilde{F}_f) = F_{\mathrm{attn_out}}[:, :, :D], \quad \mathcal{I}(\tilde{F}_f) = F_{\mathrm{attn_out}}[:, :, D:]. \quad (11)$$

The complex frequency-domain representation is reconstructed

$$\tilde{F}_f = \mathcal{R}(\tilde{F}_f) + i \cdot \mathcal{I}(\tilde{F}_f). \quad (12)$$

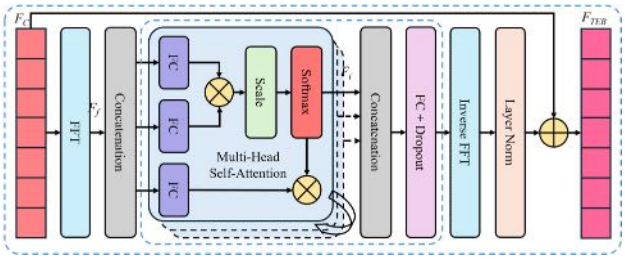

Fig. 4. Architecture of FATE. FC: fully connected layer.

The modified frequency-domain representation is projected back into the time domain as

$$\tilde{F}_C = \mathcal{R}\left(\mathrm{IFFT}(\tilde{F}_f, \mathrm{dim} = 1)\right). \quad (13)$$

A residual connection is added, followed by layer normalization to produce the final output F_{TEB} as

$$F_{TEB} = \mathrm{LayerNorm}(F_C + \tilde{F}_C) \in R^{1\times 32\times 128}. \quad (14)$$

D. Spatial Encode Branch

ResNet is able to capture features at multiple levels of abstraction while maintaining the integrity of spatial relationships [55]. Inspired by this, we developed an SEB to further extract the spatial information from the output of TEB. It is illustrated in Fig. 2(b). First, F_{TEB} is fed into the convolutional layer $CL(\cdot)$ to extract spatial features $F_{CL} \in R^{64\times 32\times 32}$ with a 1-D convolutional, batch normalization, ReLU, and max-pooling layer. Then, these features are input into the residual blocks $RB(\cdot)$ to generate residual features $F_{RB} \in R^{512\times 8\times 2}$, which are able to prevent spatial feature loss during the training stage [55]. Each residual block contains two convolutional layers with the same number and length of kernels. Among them, the kernels of the first layer are parallel to the time axis of the input EEG tensor, and the second layer is parallel to the channel axis. Finally, the output of SEB $Z \in R^{1\times 512}$, i.e., feature representation, is generated from F_{RB} by a linear layer $LL(\cdot)$, which consists of two fully connected layers and avg-pooling layers as

$$Z = LL\left(RB\left(CL\left(F_{TEB}\right)\right)\right). \quad (15)$$

E. Uncertainty-Aware Contrastive Learning

Inaccurate labels can lead to categories being wrongly assigned to incorrect classes due to the model's limited recognition capabilities. Specifically, class A might be incorrectly identified as belonging to class B, which disrupts the calculation of L_c. Although it is difficult to completely eliminate inaccurate labels, modeling the mapping probability to evaluate confidence levels can help reduce issues with low-confidence mappings.

1) Distribution Definition: This work models the probability of mapping an emotion category x_i to a representation z_i with $p(z_i|x_i)$, treating the representation as a random variable. Specifically, we employ a multivariate Gaussian distribution, $N(\mu, \sigma^2 I)$, to describe this relationship as

$$p\left(z_i|x_i\right) = N\left(z; \mu_i, \sigma_i^2 I\right) \quad (16)$$

Authorized licensed use limited to: Universidade de Macau. Downloaded on March 20,2026 at 23:55:09 UTC from IEEE Xplore. Restrictions apply.

where I denotes the unit diagonal matrix. In this setup, μ represents the most likely representation values, and σ_i^2 indicates variability within these representation values. Notably, σ_i^2 is inversely related to probability, meaning higher σ_i^2 values correspond to lower probabilities, aligning with the model in MLS. Both dimensions of μ and σ_i^2 remain consistent. The mean μ is derived from the representation head $h(\cdot)$, while the variance σ_i^2 is estimated by a parallel probability head $p(\cdot)$. Such a configuration aids in enhancing representation learning for downstream emotion recognition, ensuring class representations cluster around their prototypes in the latent space. Prototypes are typically calculated using the mean of the class representations in deterministic methods as $\rho = (1/n)\sum_{i=1}^{n} z_i$, where n is the count of representations per category. A significant limitation of this method is that if any representations are mislabeled, they could skew the prototype construction, leading to inaccuracies.

To address the above issues, this work develops a novel method for building representations. Each prototype represents the posterior distribution of n observed representations, assuming all are independently and identically distributed. The composite prototype can be modeled as $p(\rho|z_1, z_2, \ldots, z_n)$. Under the assumption that all the observations are conditionally independent, the distribution prototype can be derived as

$$p\left(\rho|z_1, z_2, \ldots, z_{n+1}\right) = \beta \frac{p\left(\rho|z_{n+1}\right)}{p\left(\rho\right)} p\left(\rho|z_1, z_2, \ldots, z_n\right) \tag{17}$$

where β is a normalization factor. After analyzing (proof details refer to study [11]), the prototype of (16) can be expressed as

$$\rho \sim N\left(\hat{\mu}, \hat{\sigma}^2 I\right) \tag{18}$$

where

$$\hat{\mu} = \sum_{i=1}^{n} \frac{\hat{\sigma}^2}{\hat{\sigma}_i^2} \mu_i \tag{19}$$

$$\frac{1}{\hat{\sigma}^2} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \tag{20}$$

where σ_i^2 denotes the per-sample variance predicted by the probability head $p(\cdot)$. From (19), it is clear that representations with different probabilities contribute differently to the prototype. Specifically, a smaller σ_i^2 (indicating higher probability) results in a greater contribution. This approach reduces the impact of inaccurate representations, thus improving the reliability of the prototype estimation. In addition, σ_i^2 acts as a radius in the latent space, indicating the potential location of the precise prototype with a certain probability. As the number of representations increases during training, the σ_i^2 of the prototype decreases, thereby narrowing the distribution region and refining the estimate of the prototype.

2) Distribution Measurement: This work defines z and ρ in (17). However, the traditional L_2 distance is inadequate for measuring the similarity between two distributions. To this end, the MLS is introduced to evaluate the similarity between the distributions z_i and z_j. It is defined as (21), where $\mu_i^{(l)}$ denotes the lth dimension of μ_i, and similarly for $\sigma_i^{(l)}$

$$\begin{aligned} MLS\left(z_i, z_j\right) &= \log(p(z_i = z_j)) \\ &= -\frac{1}{2}\sum_{l=1}^{D}\left(\frac{\left(\mu_i^{(l)} - \mu_j^{(l)}\right)^2}{\sigma_i^{2(l)} + \sigma_j^{2(l)}} + \log\left(\sigma_i^{2(l)} + \sigma_j^{2(l)}\right)\right) \\ &\quad - \frac{D}{2}\log 2\pi. \end{aligned} \tag{21}$$

The MLS combines a weighted L_2 distance with a log regularization term to overcome the limitations of the traditional L_2 distance, which only assesses similarity based on mappings in the latent space and ignores their reliability. Inaccurate labels can mislead optimization efforts and disrupt the structure of the latent space. A previous study [12] found that these inaccurate labels often correlate with low probabilities. The MLS handles these inaccuracies by considering the probabilities of z_i and z_j from two perspectives. First, as σ^2 increases, the weight assigned to the L_2 distance decreases, indicating a lower perceived similarity between z_i and z_j due to their low probabilities, despite potential similarities by L_2 standards. This method ensures that probability considerations extend beyond basic similarity measures in the representation learning process. Second, σ^2 is modified to penalize low probability representations, thereby increasing their reliability. Furthermore, σ^2 and μ are interdependent. The learnable σ^2 is directly related to the L_2 distance, meaning it can be adjusted based on the relationships among representations. Moreover, μ can also be optimized via changes in σ^2, which is consistent with logical reasoning.

Besides MLS, we consider two probabilistic distances under the same diagonal-Gaussian assumption for $p_i = \mathcal{N}(\mu_i, \mathrm{diag}(\sigma_i^2))$ and $p_j = \mathcal{N}(\mu_j, \mathrm{diag}(\sigma_j^2))$. The symmetric KL is

$$\begin{aligned} \mathrm{SKL}(p_i \| p_j) = \tfrac{1}{2}\sum_{l=1}^{D}\Bigg[&\frac{\left(\mu_i^{(l)} - \mu_j^{(l)}\right)^2}{\sigma_{i,l}^2} + \frac{\left(\mu_i^{(l)} - \mu_j^{(l)}\right)^2}{\sigma_{j,l}^2} \\ &+ \frac{\sigma_{i,l}^2}{\sigma_{j,l}^2} + \frac{\sigma_{j,l}^2}{\sigma_{i,l}^2} - 2\Bigg]. \end{aligned} \tag{22}$$

The 2-Wasserstein distance between diagonal Gaussians is

$$W_2^2(p_i, p_j) = \sum_{l=1}^{D}\left[\left(\mu_i^{(l)} - \mu_j^{(l)}\right)^2 + \left(\sqrt{\sigma_{i,l}^2} - \sqrt{\sigma_{j,l}^2}\right)^2\right]. \tag{23}$$

For use inside InfoNCE, we define similarities as

$$\begin{aligned} \tilde{s}_{\mathrm{SKL}}(i, j) &= -\frac{1}{\tau_k}\mathrm{SKL}(p_i \| p_j) \\ \tilde{s}_{W_2}(i, j) &= -\frac{1}{\tau_w} W_2^2(p_i, p_j). \end{aligned} \tag{24}$$

Empirically, both alternatives improve over nonprobabilistic baselines. However, MLS remains the most effective in our setting, owing to its variance-aware weighting and stabilizing log term that naturally down-weights low-confidence samples and provides well-behaved gradients within InfoNCE.

Authorized licensed use limited to: Universidade de Macau. Downloaded on March 20,2026 at 23:55:09 UTC from IEEE Xplore. Restrictions apply.

3) Classification Branch: In the interference stage, the classification branch extracts emotional features and predicts emotion labels based on the representations. It is illustrated in Fig. 2(c). First, the representations are fed into the multidropout layer with 512, 256, and 128 hidden units to down-sampling the features. This layer consists of three fully connected layers, batch normalization, ReLU, and dropout with 0.5. Multidropout has been determined to be able to prevent models from overfitting and enhance their robustness. Finally, a sigmoid activation function is used to predict the emotion category.

IV. Experiment Configuration

A. Datasets

This work employs five publicly available datasets, i.e., SEED [15], DREAMER [16], DEAP [17], FACED [18], and AMIGOS [19], for evaluating our framework. These datasets have different emotion labels, which can support our evaluation covering from coarse to fine.

To ensure fair cross-dataset comparison, all EEG datasets are processed using a unified preprocessing pipeline that is implemented in MNE-Python. Each raw EEG signal is bandpass filtered to 1–50 Hz using a fourth-order Butterworth filter, resampled to 128 Hz, normalized to zero mean and unit variance per channel, and segmented into 1-s nonoverlapping windows. All datasets are harmonized to a 32-channel spatial layout by spherical interpolation, ensuring consistent input dimensions across datasets. Dataset-specific operations (e.g., channel name remapping and removal of auxiliary EOG/EMG channels) are performed only for structural alignment and do not affect the underlying signal content. No dataset-specific parameter tuning is applied. This standardized pipeline eliminates preprocessing bias and ensures the fairness of all comparative evaluations.

B. Evaluation Criteria

Four typical evaluation criteria, i.e., accuracy (ACC), prediction (PRE), sensitivity (SEN), and $F1$-score ($F1$), are utilized to evaluate our framework [56]. In addition, we compute the area under the ROC curve (AUC) and Cohen's kappa (CK) across all four datasets. These metrics provide a more comprehensive view of model behavior, particularly under class imbalance conditions. Each criterion is computed and counted as mean and standard deviation (STD) with different folds (fivefold cross-validation). STD is able to measure the robustness of the model. Furthermore, the model's complexity and performance are evaluated by floating-point operations (FLOPs), parameters, and frames per second (FPS) [57].

To objectively quantify class separability in the learned embedding space, we employ three complementary metrics.

1) *Silhouette Coefficient (SC):* Measures how similar each sample is to its own cluster compared to other clusters. It is defined as

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{25}$$

where a_i is the average intraclass distance, and b_i is the nearest interclass distance for sample i. A higher value ($s_i \to 1$) indicates better separability and compactness.

2) *DB Index:* Evaluates the average ratio of intraclass scatter to interclass separation as

$$\mathrm{DB} = \frac{1}{K}\sum_{i=1}^{K}\max_{j\neq i}\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \tag{26}$$

where σ_i is the mean distance between samples and their cluster center c_i, and $d(c_i, c_j)$ is the distance between cluster centers. A lower DB index indicates more distinct and compact clusters.

3) *Interclass Margin:* Quantifies the mean pairwise distance between class centroids as

$$\mathrm{IM} = \frac{2}{K(K-1)}\sum_{i<j}\|c_i - c_j\|_2 \tag{27}$$

where K is the number of classes. A larger margin implies clearer decision boundaries.

C. Implementation

The hardware and software environments of our experiment are the NVIDIA RTX 3090 and PyTorch (Version 11.1), respectively. Pretrained models are saved at multiple epochs, and the one achieving the highest average emotion recognition accuracy over fine-tuning runs is selected to gauge the effectiveness of the pretraining framework. Training in both stages uses the Adam optimizer to minimize the respective loss functions. The pretraining phase runs for 1000 epochs with a learning rate of $1e^{-4}$ and a batch size of 64, while fine-tuning lasts 50 epochs with a learning rate of $1e^{-3}$ and a batch size of 512. The numbers of heads N_S and layers N_L are 8 and 4, respectively. Mathematically, given the total number of training epochs T_{total} and the initial weight $\lambda_u(0)$, the weight λ_u at epoch t is calculated as

$$\lambda_u(t) = \lambda_u(0)\cdot\exp\left(\alpha\cdot\left(\frac{t}{T_{total}}\right)^2\right) \tag{28}$$

where α represents a negative constant that controls the rate of weight reduction. It is set to 1×10^{-4}, which provides stable training across all datasets.

To further evaluate the cross-domain generalization of our framework, we conduct both leave-one-subject-out (LOSO) cross-validation and cross-dataset experiments. In the LOSO setting, each subject's data is used as the test set exactly once, while the remaining subjects' data forms the training set, ensuring that training and evaluation are performed on completely disjoint subject populations. For cross-dataset evaluation, we train the model on one dataset and test it on another, thereby assessing the model's robustness under domain shifts across different recording setups.

During training, peak GPU memory is dominated by intermediate activations rather than parameters or optimizer states. To reduce activation footprint, we enable mixed precision (AMP) and apply selective gradient checkpointing (SGC) to the deepest Transformer blocks in the temporal encoder and to residual stacks in the spatial encoder. This trades limited recomputation for a substantial activation reduction while preserving throughput. When GPU memory is constrained, we further use gradient accumulation to emulate large batch sizes

Authorized licensed use limited to: Universidade de Macau. Downloaded on March 20,2026 at 23:55:09 UTC from IEEE Xplore. Restrictions apply.

TABLE II
CROSS-VALIDATION PERFORMANCE OF OUR METHOD FOR EMOTION RECOGNITION

Metrics	SEED (%)		DREAMER (%)		DEAP (%)		FACED (%)	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD
ACC	94.88	3.86	98.71	2.33	96.91	2.70	99.29	1.85
PRE	93.46	3.62	97.31	2.72	95.76	2.43	98.12	1.32
SEN	93.52	3.75	98.16	2.56	95.89	2.56	97.99	1.89
F1	93.49	3.66	97.73	2.82	95.82	2.21	98.05	1.72
AUC	98.12	1.24	99.41	0.85	98.76	1.05	99.72	0.48
CK	92.15	3.58	97.95	2.13	95.12	2.44	98.65	1.21

without increasing instantaneous memory. These practices make the framework practical on commodity 8–24-GB GPUs while retaining the reported accuracy and throughput.

V. RESULT AND DISCUSSION

A. Recognition Result

We train and evaluate our framework for EEG emotion recognition using four datasets based on the experimental setups described above. These results are listed in Table II, which contains mean and STD metrics.

As shown in Table II, our framework achieves a mean accuracy, prediction, sensitivity, and $F1$-score of 94.88%, 98.46%, 93.52%, and 93.49% on the SEED dataset, respectively. It yields the STD accuracy, prediction, sensitivity, and $F1$-score of 3.86%, 3.62%, 3.75%, and 3.66% on the SEED dataset, respectively. In addition, UACL-Net achieves mean $F1$-scores of 93.49%, 97.73%, 95.82%, and 98.05% on SEED, DREAMER, DEAP, and FACED datasets, respectively. It also yields STD $F1$-scores of 3.66%, 2.82%, 2.21%, and 1.72% on SEED, DREAMER, DEAP, and FACED datasets, respectively. From Table II, we found that our model does not perform as well on the SEED dataset as it does on other datasets.

To provide a more comprehensive evaluation, we further report AUC and CK. The framework consistently yields AUC above 98.0% and CK above 92.0% across all datasets, indicating strong discriminative ability and robust agreement beyond chance under potential class imbalance. In particular, FACED achieves the best performance with an AUC of 99.72% and a CK of 98.65%, while SEED remains relatively lower (AUC 98.12%, CK 92.15%). After analyzing the SEED dataset, it was collected in a naturalistic setting, with participants watching the stimuli on their own devices in a non-laboratory environment.

B. Comparison With Counterparts

To evaluate the advantages of our proposal, the following studies [5], [7], [8], [9], [39], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71] are introduced for peer comparison. These models can be divided into two types, i.e., supervised learning and transfer learning-based models. It is worth noting that the SEED dataset is used for the evaluation of different bands. DREAMER, DEAP, and FACED datasets are utilized in the evaluation of different continuous types. In addition, the P-value obtained from a paired t-test statistical evaluation is used to examine the statistical significance between different fold evaluations.

The performance records on the SEED dataset are listed in Table III, where optimum and suboptimum performance are highlighted in bold and underlined, respectively. The previous SOTA method, GDDN [7], reported mean accuracies of 71.62%, 76.11%, 73.97%, 83.08%, and 84.44% on the δ, θ, α, β, and γ bands, respectively, with an overall average of 92.54% ($\pm$3.65%). More recent approaches, such as Chen et al. [5] and Ko et al. [9], further boosted the accuracy to 93.41% and 93.85% across all bands. In contrast, our framework achieves the best overall performance of 94.88% ($\pm$3.86%), outperforming the latest published results by 1.03-2.34% in mean accuracy.

On the other hand, Table IV presents the recognition performance on the DREAMER, DEAP, and FACED datasets. For DREAMER, our model attains 99.11% (valence) and 99.28% (arousal), which is higher than Ko et al. [9] (98.9%) and Chen et al. [7] (99.07%). On DEAP, our method achieves 98.10% (valence) and 98.19% (arousal), clearly surpassing the prior top performance of 96.59%. For FACED, our model delivers 99.28% (valence) and 99.21% (arousal), exceeding all compared methods that ranged between 60% and 99%. Moreover, our standard deviations (1.5%–2.6%) are consistently lower than those of other strong baselines (3%–7%) The "*" indicates that, via paired t-tests against all baselines, our improvements are statistically significant at $p < 0.01$. These results confirm our approach improves accuracy and robustness to EEG noise and subject differences, ensuring strong generalization across diverse emotion recognition benchmarks.

It can be seen that the performance records of transfer learning-based models are better than supervised learning-based ones. This is because transfer learning can maximize the similarity between positive sample pairs while minimizing the similarity between negative sample pairs. This results in richer and more robust feature representations, which are crucial to our task due to the complex structures and subtle differences present. This work further boosts performance by incorporating spatial and temporal features with learnable weights. These performance records show that our model is effective and outperforms other SOTA methods.

C. Ablation Study

1) Contribution of UACL: This work develops UACL to decrease the class-internal difference, thereby enhancing the robustness of emotion recognition. To evaluate the contributions of this method, we first analyze the performance records of our model with different loss functions, i.e., cross-entropy (CE) [53], CosFace [72], ArcFace [73], and SphereFace [51]. In addition, we compare the results of our UACL-Net with other typical contrastive learning methods, i.e., SimCLR [43], MoCo [44], and BYOL [45]. Furthermore, we replace MLS in UACL with symmetric KL and 2-Wasserstein (W_2) while keeping all settings fixed. These performance records are listed in Table V.

As can be observed, our UACL-Net without the pretrain stage yields a mean accuracy of 87.82%, 91.37%, 89.77%, and 91.91% on the SEED, DREAMER, DEAP, and FACED datasets, respectively. These metrics are decreased by 7.60%, 7.34%, 7.14%, and 7.38%, respectively, compared to our framework. Our UACL-Net without the fine-tune stage also

Authorized licensed use limited to: Universidade de Macau. Downloaded on March 20,2026 at 23:55:09 UTC from IEEE Xplore. Restrictions apply.

TABLE III
ACCURACY OF OUR METHOD AND COUNTERPARTS FOR EEG EMOTION RECOGNITION ON SEED DATASET

Methods	δ Band (%)		θ Band (%)		α Band (%)		β Band (%)		γ Band (%)		All Bands (%)	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Supervised Learning												
Suykens *et al.* [58]	43.06	8.27	40.07	6.50	43.97	10.89	48.63	10.29	51.59	11.83	56.73	16.29
Song *et al.* [59]	49.79	10.94	46.36	12.06	48.29	12.28	56.15	14.01	54.87	17.53	79.95	9.02
Li *et al.* [60]	58.92	7.41	61.38	9.05	59.74	8.66	66.81	9.12	68.43	8.97	83.15	7.12
Li *et al.* [61]	70.37	7.68	76.00	6.92	66.22	11.52	72.54	8.97	71.70	8.03	86.81	5.79
Li *et al.* [62]	61.43	6.33	62.56	10.53	66.58	10.56	72.64	8.87	74.88	7.36	87.04	6.52
Song *et al.* [63]	64.35	6.63	63.52	10.60	62.01	9.07	70.73	4.85	74.13	7.43	86.30	6.91
Li *et al.* [64]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	86.52	6.22
Liu *et al.* [39]	68.35	6.11	66.77	10.48	65.51	8.77	75.16	4.65	77.93	7.03	89.55	6.79
Transfer Learning												
Pan *et al.* [65]	44.10	8.22	41.26	9.21	42.93	14.33	43.93	10.06	48.43	9.73	63.64	14.88
Fernando *et al.* [66]	53.23	7.47	50.60	8.31	55.06	10.60	56.72	10.78	64.47	14.96	69.00	10.89
Li *et al.* [67]	63.01	7.49	63.22	7.52	63.50	9.50	73.59	9.12	73.72	8.67	84.14	6.87
Zhong *et al.* [68]	64.88	6.87	60.69	5.79	60.84	7.57	74.96	8.94	77.50	8.10	85.30	6.72
Chen *et al.* [69]	64.60	6.76	70.06	8.28	60.94	9.23	73.72	10.34	76.55	8.95	89.63	6.79
Ma *et al.* [70]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	89.30	9.18
Li *et al.* [71]	63.40	7.58	62.13	6.59	66.35	11.62	75.10	10.14	78.56	8.74	88.10	6.54
Kan *et al.* [8]	65.15	7.45	71.35	7.58	73.49	10.52	72.19	7.45	80.78	7.36	88.55	7.72
Chen *et al.* [7]	71.62	6.54	76.11	6.78	73.97	9.87	83.08	6.87	84.44	7.75	92.54	**3.65**
Ko *et al.* [9]	73.25	6.72	75.10	6.95	75.42	9.35	84.26	6.72	85.03	7.14	93.41	5.12
Chen *et al.* [5]	74.02	6.45	**76.28**	6.83	75.51	9.18	85.07	6.51	84.62	7.03	93.85	5.04
Ours*	**79.09**	**6.38**	75.02	**6.52**	**77.31**	**8.52**	**86.64**	**6.45**	**85.12**	**6.36**	**94.88**	3.86

TABLE IV
ACCURACY OF OUR METHOD AND COUNTERPARTS FOR EEG EMOTION RECOGNITION ON DREAMER, DEAP, AND FACED DATASETS

Methods	DREAMER (%)				DEAP (%)				FACED (%)			
	Valence		Arousal		Valence		Arousal		Valence		Arousal	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Supervised Learning												
Suykens *et al.* [58]	60.46	8.12	60.54	7.05	58.73	8.30	58.77	6.22	60.52	7.45	60.57	7.38
Song *et al.* [59]	85.25	7.89	85.32	7.95	82.73	7.10	82.83	7.02	85.34	6.98	85.37	6.05
Li *et al.* [60]	88.12	6.10	88.25	6.05	85.10	5.40	85.22	5.35	88.65	5.20	88.72	5.18
Li *et al.* [61]	92.51	7.72	92.64	7.68	89.83	6.85	89.93	6.79	92.62	7.65	92.69	7.71
Li *et al.* [62]	92.76	5.75	92.88	5.70	90.07	5.80	90.17	5.76	92.86	4.68	92.94	4.72
Song *et al.* [63]	91.97	6.82	92.09	6.78	89.35	5.90	89.41	5.88	92.07	4.74	92.15	4.80
Li *et al.* [64]	92.23	4.70	92.33	4.66	89.53	4.83	89.63	4.79	92.31	3.69	92.38	3.73
Liu *et al.* [39]	95.01	4.65	95.13	3.60	93.10	3.75	93.20	3.70	95.50	4.62	95.58	3.68
Transfer Learning												
Pan *et al.* [65]	67.82	7.25	67.91	7.18	65.85	6.30	65.93	7.22	67.95	5.40	67.95	5.35
Fernando *et al.* [66]	73.53	7.15	73.63	6.08	71.43	7.20	71.48	5.12	73.62	5.25	73.68	5.18
Li *et al.* [67]	89.67	4.95	89.79	5.90	87.07	4.05	87.17	5.98	89.77	4.92	89.84	4.88
Zhong *et al.* [68]	90.90	4.88	91.02	4.82	88.27	4.95	88.37	4.89	91.01	4.90	91.08	4.85
Chen *et al.* [69]	95.52	4.68	95.64	4.62	92.75	4.80	92.86	4.75	95.63	4.70	95.74	4.65
Ma *et al.* [70]	95.17	4.70	95.29	4.65	92.41	4.82	92.51	4.78	95.27	4.72	95.35	4.68
Li *et al.* [71]	93.89	4.80	94.01	4.75	91.17	4.90	91.27	4.85	93.99	4.82	94.07	4.78
Kan *et al.* [8]	94.37	3.77	94.49	4.72	91.63	4.88	91.74	3.83	94.47	3.80	94.55	3.76
Chen *et al.* [7]	98.62	3.54	98.75	3.50	95.76	3.65	95.87	3.60	98.73	2.55	98.81	2.52
Ko *et al.* [9]	98.81	3.12	98.93	3.08	96.22	3.35	96.33	3.30	98.91	2.43	98.97	2.40
Chen *et al.* [5]	98.95	3.05	99.07	3.01	96.48	3.28	96.59	3.24	99.05	2.38	99.11	2.35
Ours*	**99.11**	**2.58**	**99.28**	**2.54**	**98.10**	**2.65**	**98.19**	**2.60**	**99.28**	**1.53**	**99.21**	**1.57**

decreased by about 3.00%. In addition, our model with SphereFace loss also has a better performance than the other peer loss functions, likely due to its stronger angular margin that enforces clearer class boundaries. Nevertheless, our framework with ArcFace or CosFace also outperforms existing baselines, indicating that the reported improvements are not solely due to SphereFace but arise from the integration of UACL and FASA with the chosen pretraining strategy.

MLS attains the best or tied-best accuracy, W_2 is competitive but slightly lower, and SKL trails due to higher sensitivity to variance scale despite symmetrization. These findings reinforce that the uncertainty-aware similarity shaping is crucial, consistent with the significant degradation observed when UACL is removed. From Table V, removing UACL results in mean accuracies of 82.25%, 85.57%, 84.01%, and 86.08% on SEED, DREAMER, DEAP, and FACED, respectively, 9.08%, 9.44%, 9.27%, and 9.49% lower than our complete framework. Furthermore, integrating our model with other contrastive learning methods further enhances performance.

On the other hand, we visualize the decision features with different edge-aware methods to evaluate our UACL, which can be illustrated in Fig. 5. It is evident that when

Authorized licensed use limited to: Universidade de Macau. Downloaded on March 20,2026 at 23:55:09 UTC from IEEE Xplore. Restrictions apply.

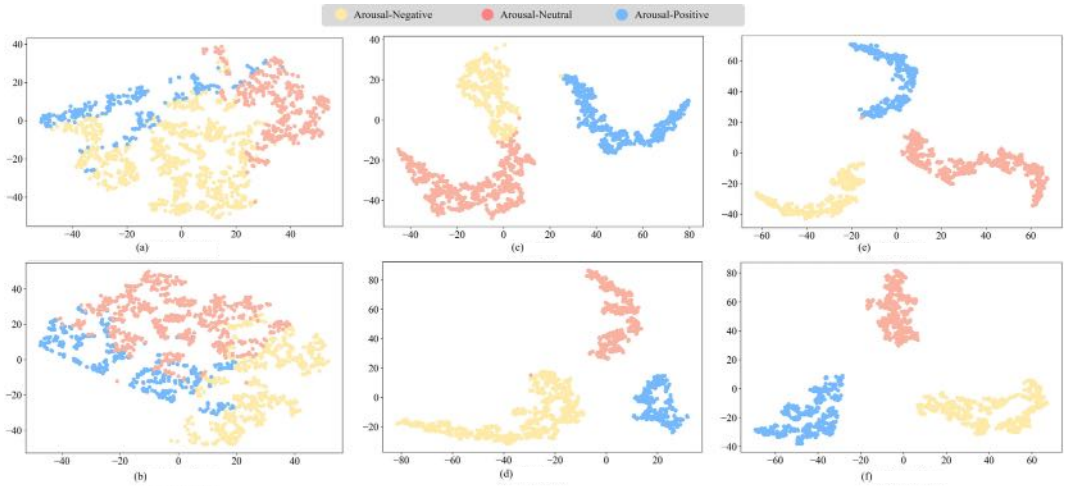

Fig. 5. Feature clustering maps of the ablation study for UACL on SEED dataset. (a) UACL-Net without a pretraining stage. (b) UACL-Net without the fine-tune stage. (c) UACL-Net with cross-entropy loss. (d) UACL-Net with ArcFace loss. (e) UACL-Net with CosFace loss. (f) UACL-Net with SphereFace loss.

TABLE V
ABLATION STUDY OF UACL FOR EEG EMOTION RECOGNITION

Methods	SEED (%)		DREAMER (%)		DEAP (%)		FACED (%)	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Loss								
w/ CE	87.82	4.17	91.37	2.52	89.77	2.92	91.91	2.09
w/ CosFace	91.15	4.02	94.83	2.43	93.11	2.81	95.39	1.93
w/ ArcFace	91.99	3.98	95.72	2.44	93.96	2.78	96.26	1.91
w/ SphereFace	**94.88**	**3.86**	**98.71**	**2.33**	**96.91**	**2.70**	**99.29**	**1.85**
MLS								
w/ W_2	94.10	3.90	98.20	2.40	96.20	2.74	99.00	1.88
w/ SKL	93.20	3.95	97.80	2.48	95.60	2.79	98.60	1.92
Contrastive								
w/o UACL	82.25	4.45	85.57	2.69	84.01	3.11	86.08	2.13
w/ BYOL	91.96	4.03	95.63	2.43	93.94	2.82	96.18	1.93
w/ SimCLR	89.26	4.15	92.82	2.56	91.15	2.91	93.36	1.99
w/ MoCo	92.33	4.01	96.01	2.42	94.28	2.81	96.57	1.92

TABLE VI
QUANTITATIVE BOUNDARY SEPARABILITY METRICS

Dataset	w/o UACL			UACL-Net		
	SC ↑	DB ↓	IM ↑	SC ↑	DB ↓	IM ↑
SEED	0.36	0.88	1.45	0.48	0.63	2.01
DEAP	0.31	0.93	1.32	0.42	0.68	1.84
DREAMER	0.38	0.82	1.57	0.50	0.59	2.12
FACED	0.33	0.90	1.40	0.44	0.65	1.93

combined with CE, the clustered features are not well-separated, leading to an increased prediction error rate. In addition, when combined with CosFace and ArcFace, there is misclassification between arousal-neutral and arousal-negative or arousal-positive, respectively. In contrast, our model effectively distinguishes these features. This experiment determined that UACL can effectively decrease the class-internal difference and enhance the robustness of emotion recognition.

From Table VI, it is evident that UACL-Net achieves higher Silhouette scores, lower Davies–Bouldin (DB) indices, and larger interclass margins (IMs) across all four datasets, compared to its ablated version (w/o UACL). This consistently demonstrates that uncertainty-aware modeling effectively sharpens decision boundaries and enhances cluster separability. Among the datasets, DREAMER and SEED exhibit the most significant improvement, which correlates with their higher intersubject consistency and balanced emotion labels. These quantitative results provide strong empirical support for the claim that UACL promotes more discriminative and well-structured embedding spaces.

As shown in Table VII, UACL-Net demonstrates strong robustness to variance estimation errors across all four datasets. When multiplicative Gaussian noise is applied to the predicted variance with perturbation strength $\delta \in \{0.05, 0.10, 0.20\}$, the accuracy degradation remains below 1.2% for $\delta \leq 0.10$ and within 3% even at $\delta = 0.20$. This indicates that the probabilistic embedding design effectively absorbs moderate variance fluctuations without compromising discriminability. In addition, introducing a variance regularization term during training further stabilizes model performance. A mild regularization factor ($\lambda = 0.01$) yields the best results, improving accuracy by approximately 0.6% across all datasets, while overly weak ($\lambda = 0.001$) or strong ($\lambda = 0.05$) regularization slightly reduces accuracy. These results confirm that UACL-Net maintains stable decision boundaries under uncertain or imprecise variance estimation, validating the robustness of the proposed probabilistic embedding framework.

2) Contribution of FASA: A hybrid network is developed that cascades TEB and SEB to extract spatial and temporal information. To evaluate the contributions of the multi-information encoder, we analyze the performance records of our model without TEB and SEB. In addition, we also evaluate the contributions of two branches. Regarding the TEB, we analyze the performance records of the TEB without PE and FATE. To evaluate the contributions of FATE, we analyze the performance records of FATE without FFT. On the other hand, this work analyzes the performance records of the SEB with MinPool, MaxPool, and AvgPool to explore the effects of different pooling methods. Furthermore, we analyze the effects of dilated operation for this model. Finally, peer typical attentions, i.e., SA [54], frequency enhanced

Authorized licensed use limited to: Universidade de Macau. Downloaded on March 20,2026 at 23:55:09 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
ROBUSTNESS TO VARIANCE ESTIMATION ERRORS ACROSS FOUR DATASETS

Setting	SEED	DEAP	DREAMER	FACED
	Acc. (%) / Δ (%)	Acc. (%) / Δ (%)	Acc. (%) / Δ (%)	Acc. (%) / Δ (%)
Original	94.88 / 0.00	96.91 / 0.00	98.71 / 0.00	99.29 / 0.00
Perturbation				
$\delta = 0.05$	94.31 / –0.57	96.32 / –0.59	98.18 / –0.53	98.78 / –0.51
$\delta = 0.10$	93.73 / –1.15	95.77 / –1.14	97.62 / –1.09	98.15 / –1.14
$\delta = 0.20$	92.12 / –2.76	94.10 / –2.81	96.00 / –2.71	97.00 / –2.29
Regularization				
$\lambda = 0.001$	95.08 / +0.20	97.02 / +0.11	98.83 / +0.12	99.40 / +0.11
$\lambda = 0.01$	95.48 / +0.60	97.51 / +0.60	99.31 / +0.60	99.89 / +0.60
$\lambda = 0.05$	94.76 / –0.12	96.55 / –0.36	98.40 / –0.31	98.90 / –0.39

TABLE VIII
ABLATION STUDY OF FASA FOR EEG EMOTION RECOGNITION

Methods	SEED (%)		DREAMER (%)		DEAP (%)		FACED (%)	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD
Hybrid								
w/o TEB	90.35	4.05	94.07	2.45	92.28	2.84	94.55	1.94
w/o SEB	89.76	4.08	93.38	2.46	91.68	2.85	93.93	1.96
TEB								
w/o PE	92.78	3.95	96.53	2.38	94.77	2.76	97.09	1.89
w/o FATE	91.25	4.01	94.93	2.42	93.26	2.81	95.49	1.92
w/o FFT	91.53	4.00	95.12	2.45	93.29	2.86	95.58	1.96
SEB								
w/ MinPool	92.58	3.96	96.32	2.39	94.56	2.77	96.88	1.98
w/ MaxPool	**94.88**	**3.86**	**98.71**	**2.33**	**96.91**	**2.70**	**99.29**	**1.85**
w/ AvePool	94.25	3.89	98.05	2.35	96.27	2.72	98.63	1.86
w/o Dilate	93.96	3.95	97.75	2.35	95.97	2.73	98.33	1.87
Peer								
w/ SA	92.15	3.95	97.50	2.45	95.06	2.80	97.20	1.90
w/ FED	93.21	3.90	97.62	2.38	96.12	2.72	98.52	1.87
w/ AFE	94.01	3.88	98.40	2.35	96.70	2.72	99.00	1.86

decomposition (FED) [14], and adaptive frequency enhancement (AFE) [13], are utilized to compare this module. These results are listed in Table VIII.

As seen in Table VIII, our model without TEB yields mean accuracies of 90.35%, 91.37%, 89.77%, and 91.91% on SEED, DREAMER, DEAP, and FACED datasets, respectively. These metrics are decreased by 4.53%, 4.64%, 4.63%, and 4.74%, respectively, compared to our framework. In addition, our model without SEB yields mean accuracies of 89.76%, 93.38%, 91.68%, and 93.93% on SEED, DREAMER, DEAP, and FACED datasets, respectively. These metrics are decreased by 5.12%, 5.33%, 5.23%, and 5.36%, respectively, compared to our framework. Their STD accuracies are all increased by about 1.00%. We can find that temporal information is more imparted than spatial information in this task. Excluding PE or FATE also leads to performance drops. Furthermore, FATE without FFT decreased by 3.35%, 3.59%, 3.62%, and 3.71% on SEED, DREAMER, DEAP, and FACED datasets, respectively, compared to our framework. Similarly, MaxPool performs better than AvgPool in SEB, and omitting the dilated operation slightly reduces accuracy. FATE also has a better performance than peer attentions.

To further validate the interpretability of the proposed FASA module, we analyzed the learned attention distributions across EEG frequency bands. By averaging the attention weights across channels and time, we obtained band-level importance scores. As shown in Fig. 6, the model consistently assigns

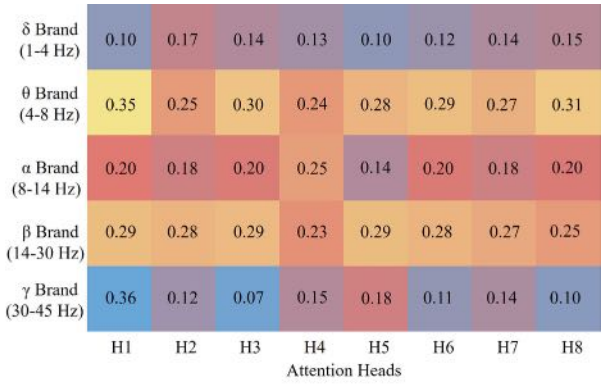

Fig. 6. Frequency-band attention distribution learned by FASA.

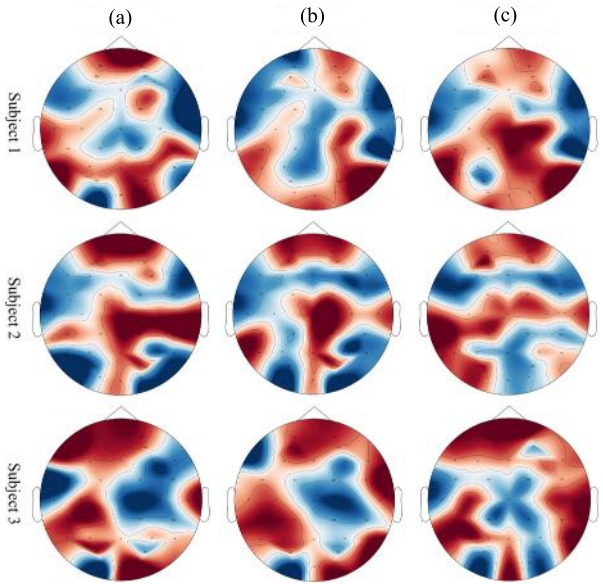

Fig. 7. Visualization of topographic maps for the hybrid model on the SEED dataset. (a) UACL-Net without TEB. (b) UACL-Net without SEB. (c) UACL-Net.

higher importance to the θ and β bands, which have been widely reported as key correlates of emotional processing in EEG literature. This visualization confirms that FASA indeed attends to meaningful frequency-specific patterns rather than acting as a input-output-only module.

The multi-information cascade is further analyzed in relation to neurophysiological findings. Multiscale entropy (MSE)

Authorized licensed use limited to: Universidade de Macau. Downloaded on March 20,2026 at 23:55:09 UTC from IEEE Xplore. Restrictions apply.

TABLE IX
CONFUSION MATRIX ANALYSIS ON SEED DATASET. "0": NEGATIVE/PASSIVE. "1": NEUTRAL. "2": POSITIVE/ACTIVE

Methods	w/o UACL			w/o TEB			w/o SEB			UACL-Net		
	0	1	2	0	1	2	0	1	2	0	1	2
Arousal												
0	90.12	7.28	2.60	91.83	6.09	2.08	92.45	5.45	2.10	94.57	4.03	1.40
1	6.95	86.38	6.67	5.82	88.06	6.12	5.52	88.86	5.62	4.18	92.72	3.10
2	3.82	8.09	88.09	2.43	7.57	90.00	2.08	6.69	91.23	1.63	5.09	93.28
Valence												
0	91.31	6.79	1.90	92.18	5.91	1.91	92.98	5.02	2.00	95.18	3.82	1.00
1	7.38	85.19	7.43	6.23	86.99	6.78	5.83	88.37	5.80	4.47	92.00	3.53
2	3.34	7.68	88.98	2.72	7.08	90.20	2.39	6.32	91.29	1.73	5.19	93.08

[74] is applied to EEG feature maps, and results are visualized in Fig. 7. The cascade enhances functional connectivity in specific brain regions. For our model without TEB, connections in prefrontal regions are strengthened, with fewer changes in the left parietal and right parietal lobes. Without SEB, the prefrontal lobes show more pronounced strengthening, while internal connections within the left parietal regions display increased strength and expanded connectivity. Multi-information cascades highlight enhanced strength and extent of internal connections in the prefrontal lobes, along with improved internal connectivity between both the right and left parietal lobe regions. The cascade mechanism improves internal connections in prefrontal and parietal regions, aligning with existing neuroscience discoveries.

D. Confusion Matrix Analysis

To further examine potential class bias and intercategory confusion, we analyze the normalized confusion matrices on the SEED dataset, as shown in Table IX. The objective of this experiment is to evaluate whether UACL-Net maintains balanced discrimination across emotional categories and reduces bias toward dominant affective states. As shown in the results, UACL-Net exhibits more uniform recognition performance compared to its ablated variants. Without UACL, the model tends to confuse neutral (1) and positive (2) emotions, leading to higher off-diagonal elements in the upper right region of the matrix. The inclusion of the TEB and the SEB progressively improves class separability, while the complete UACL-Net further reduces cross-class confusion, especially for the neutral class, whose recall increases by 4.1% over "w/o UACL." Misclassifications primarily occur between adjacent affective states (e.g., neutral versus positive) that share overlapping EEG spectral characteristics. Overall, this analysis confirms that UACL-Net effectively mitigates class bias and achieves more consistent discrimination across emotion categories.

E. Discussion

The proposed UACL-Net demonstrates strong potential for clinical applications, particularly in mood disorder diagnosis, neurorehabilitation, and assistive technologies, due to its high accuracy and robustness in EEG-based emotion recognition. The core innovation lies in the integration of UACL and FASA. UACL models latent representations as multivariate Gaussian distributions, effectively capturing interclass differences and mitigating the influence of noisy or ambiguous labels. FASA, on the other hand, enhances the model's ability to capture global dependencies and reduce signal noise by applying SA in the frequency domain. Unlike prior works that mainly apply advanced attention mechanisms to EEG, our contributions lie in introducing uncertainty-aware prototype learning and frequency-structured spatial attention. These innovations address fundamental issues of noise robustness and spectral–spatial representation, representing a significant step beyond contextual adaptation. Compared to existing SOTA methods, UACL-Net consistently outperforms both supervised and self-supervised baselines across four benchmark datasets, achieving up to 99.29% accuracy on the FACED dataset. These results highlight the effectiveness and generalizability of the proposed framework in complex emotion recognition tasks. Our classifier could serve as a richer affective input channel for next-generation brain–computer interfaces (BCIs). This advance would boost BCI responsiveness and user experience in areas like neurorehabilitation, smart assistants, and immersive entertainment, accelerating real-world adoption.

It is worth noting that our uncertainty-aware prototype formulation assumes conditional independence among observed representations. While EEG signals are naturally correlated over time and across channels, this approximation enables closed-form prototype aggregation and efficient confidence interval estimation. Importantly, such dependencies are already captured by the temporal encoder and spatial encoder before prototype construction, making the independence assumption a reasonable simplification in practice. Nevertheless, future work may consider more flexible probabilistic formulations, such as Gaussian process priors for temporal continuity or graphical models for interchannel dependencies, to further relax this assumption.

There are some challenges, and future directions are summarized as follows. First, emotional states are inherently complex and multidimensional, often manifested through mixed cues across facial expressions, speech, and physiological responses. Traditional EEG-based models may oversimplify this complexity; thus, fusing EEG with other physiological or contextual signals could provide a more comprehensive representation of emotions. Second, real-time emotion recognition requires lightweight and efficient models, but deep learning frameworks often entail high computational costs. Future work should explore model compression and feature selection to enhance efficiency, enabling deployment in real-world, resource-limited scenarios. Finally, EEG data raises serious privacy concerns due to its sensitive nature.

Authorized licensed use limited to: Universidade de Macau. Downloaded on March 20,2026 at 23:55:09 UTC from IEEE Xplore. Restrictions apply.

Ensuring secure data sharing and storage is essential, particularly in multi-institution collaborations. Robust anonymization methods, differential privacy, and federated learning frameworks offer promising solutions for safeguarding user privacy while maintaining model utility.

VI. Conclusion

This study develops a novel framework for EEG emotion recognition. Specifically, UACL is based on a multivariate Gaussian distribution to model the embedding space and is able to highlight interclass differences. In addition, a hybrid model that combines the Transformer and CNN is further developed. It is able to efficiently extract spatial and temporal information from EEG signals and therefore yields accurate emotion recognition. Our model is trained and tested on four benchmark datasets, which are able to cover coarse and fine validation. Both the qualitative and quantitative results demonstrate that our proposed model is effective and has advantages over peer methods. In the future, we plan to enhance our framework with cross-modal inputs and multicenter clinical data to further improve its applicability.

References

[1] X. Wu, L. Lu, M. Dotoli, G. Fortino, and M. Chen, "A novel agent-based approach for dynamic emotion modeling in social networks," *IEEE Trans. Cybern.*, early access, Nov. 7, 2025, doi: 10.1109/TCYB.2025.3625166.

[2] K. Hou et al., "Emotion recognition from multimodal physiological signals via discriminative correlation fusion with a temporal alignment mechanism," *IEEE Trans. Cybern.*, vol. 54, no. 5, pp. 3079–3092, May 2024.

[3] C.-H. Lee, H. Kim, B. C. Yoon, and D.-J. Kim, "Toward foundational model for sleep analysis using a multimodal hybrid-self-supervised learning framework," *IEEE Trans. Cybern.*, vol. 55, no. 12, pp. 5619–5632, Dec. 2025.

[4] S. Lin, C. Fan, D. Han, Z. Jia, Y. Peng, and S. Kwong, "HATNet: EEG-based hybrid attention transfer learning network for train driver state detection," *IEEE Trans. Cybern.*, vol. 55, no. 5, pp. 2437–2450, May 2025.

[5] C. L. Philip Chen, B. Chen, and T. Zhang, "AdamGraph: Adaptive attention-modulated graph network for EEG emotion recognition," *IEEE Trans. Cybern.*, vol. 55, no. 5, pp. 2038–2051, May 2025.

[6] H. Yan, K. Guo, X. Xing, and X. Xu, "Bridge graph attention based graph convolution network with multi-scale transformer for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 15, no. 4, pp. 2042–2054, Apr. 2024.

[7] B. Chen, C. L. P. Chen, and T. Zhang, "GDDN: Graph domain disentanglement network for generalizable EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 15, no. 3, pp. 1739–1753, Feb. 2024.

[8] H. Kan, J. Yu, J. Huang, Z. Liu, H. Wang, and H. Zhou, "Self-supervised group meiosis contrastive learning for EEG-based emotion recognition," *Int. J. Speech Technol.*, vol. 53, no. 22, pp. 27207–27225, Nov. 2023.

[9] W. Ko, S. Jeong, S.-K. Song, and H.-I. Suk, "EEG-oriented self-supervised learning with triple information pathways network," *IEEE Trans. Cybern.*, vol. 54, no. 11, pp. 6495–6508, Nov. 2024.

[10] Z. Zhang, Y. Liu, and S.-H. Zhong, "GANSER: A self-supervised data augmentation framework for EEG-based emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2048–2063, Jul. 2023.

[11] Y. Shi and A. Jain, "Probabilistic face embeddings," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6902–6911.

[12] H. Xie et al., "Boosting semi-supervised semantic segmentation with probabilistic representations," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 3, pp. 2938–2946.

[13] X. Xu, C. Fu, J. Chen, G. Jeon, and D. Camacho, "Robust imagined speech production from electrocorticography with adaptive frequency enhancement," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Sep. 2024, pp. 1–7.

[14] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "FedFormer: Frequency enhanced decomposed transformer for long-term series forecasting," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 27268–27286.

[15] 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.

[16] S. Katsigiannis and N. Ramzan, "DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 1, pp. 98–107, Jan. 2018.

[17] S. Koelstra et al., "DEAP: A database for emotion analysis;Using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan. 2012.

[18] J. Chen, X. Wang, C. Huang, X. Hu, X. Shen, and D. Zhang, "A large finer-grained affective computing EEG dataset," *Sci. Data*, vol. 10, no. 1, p. 740, Oct. 2023.

[19] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A dataset for affect, personality and mood research on individuals and groups," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 479–493, Apr. 2021.

[20] Y. Yin, X. Zheng, B. Hu, Y. Zhang, and X. Cui, "EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM," *Appl. Soft Comput.*, vol. 100, Mar. 2020, Art. no. 106954.

[21] P. Pandey and K. R. Seeja, "Subject independent emotion recognition from EEG using VMD and deep learning," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 5, pp. 1730–1738, May 2022.

[22] Y. Xu et al., "AMDET: Attention based multiple dimensions EEG transformer for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 15, no. 3, pp. 1067–1077, Jul. 2024.

[23] 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.

[24] S. Liu, Y. Zhao, Y. An, J. Zhao, S.-H. Wang, and J. Yan, "GLFANet: A global to local feature aggregation network for EEG emotion recognition," *Biomed. Signal Process. Control*, vol. 85, Aug. 2023, Art. no. 104799.

[25] Q. Zhou, C. Shi, Q. Du, and L. Ke, "A multi-task hybrid emotion recognition network based on EEG signals," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105136.

[26] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 14, no. 2, pp. 715–729, Jun. 2022.

[27] W. Liu, W.-L. Zheng, Z. Li, S.-Y. Wu, L. Gan, and B.-L. Lu, "Identifying similarities and differences in emotion recognition with EEG and eye movements among Chinese, German, and French people," *J. Neural Eng.*, vol. 19, no. 2, Apr. 2022, Art. no. 026012.

[28] X. Wang, Y. Ma, J. Cammon, F. Fang, Y. Gao, and Y. Zhang, "Self-supervised EEG emotion recognition models based on CNN," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1952–1962, 2023.

[29] C. Fan et al., "ICaps-ResLSTM: Improved capsule network and residual LSTM for EEG emotion recognition," *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105422.

[30] Y. Ding, S. Zhang, C. Tang, and C. Guan, "MASA-TCN: Multi-anchor space-aware temporal convolutional neural networks for continuous and discrete EEG emotion recognition," *IEEE J. Biomed. Health Informat.*, vol. 28, no. 7, pp. 3953–3964, Jul. 2024.

[31] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[32] T. Song, W. Zheng, C. Lu, Y. Zong, X. Zhang, and Z. Cui, "MPED: A multi-modal physiological emotion database for discrete emotion recognition," *IEEE Access*, vol. 7, pp. 12177–12191, 2019.

[33] 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.

[34] X. Xu et al., "Embedded EEG feature selection for multi-dimension emotion recognition via local and global label relevance," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 514–526, 2024.

[35] X. Xu, T. Jia, Q. Li, F. Wei, L. Ye, and X. Wu, "EEG feature selection via global redundancy minimization for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 421–435, Jan. 2023.

[36] Y. An, S. Hu, S. Liu, X. Wang, Z. Gu, and Y. Zhang, "LGDAAN-Nets: A local and global domain adversarial attention neural networks for EEG emotion recognition," *Knowl.-Based Syst.*, vol. 318, Jun. 2025, Art. no. 113613.

[37] Y. Ding et al., "EmT: A novel transformer for generalized cross-subject EEG emotion recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 6, pp. 10381–10393, Jun. 2025.

Authorized licensed use limited to: Universidade de Macau. Downloaded on March 20,2026 at 23:55:09 UTC from IEEE Xplore. Restrictions apply.

[38] X. Hu, F. Wang, and D. Zhang, "Similar brains blend emotion in similar ways: Neural representations of individual difference in emotion profiles," *NeuroImage*, vol. 247, Feb. 2022, Art. no. 118819.
[39] S. Liu et al., "MAS-DGAT-Net: A dynamic graph attention network with multibranch feature extraction and staged fusion for EEG emotion recognition," *Knowl.-Based Syst.*, vol. 305, Dec. 2024, Art. no. 112599.
[40] W. Li, Y. Tian, B. Hou, J. Dong, S. Shao, and A. Song, "A bi-stream hybrid model with MLPBlocks and self-attention mechanism for EEG-based emotion recognition," *Biomed. Signal Process. Control*, vol. 86, Sep. 2023, Art. no. 105223.
[41] Y. Wei, Y. Liu, C. Li, J. Cheng, R. Song, and X. Chen, "TC-Net: A transformer capsule network for EEG-based emotion recognition," *Comput. Biol. Med.*, vol. 152, Jan. 2023, Art. no. 106463.
[42] H. Wang et al., "Learning contrastive evolving micro-clusters for robust semi-supervised data stream classification," *IEEE Trans. Cybern.*, vol. 56, no. 1, pp. 334–346, Jan. 2026.
[43] 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.
[44] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.
[45] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.
[46] W. Li, W. W. Y. Ng, H. Wang, J. Zhang, C. Zhong, and L. Yang, "ERMAV: Efficient and robust graph contrastive learning via multiadversarial views training," *IEEE Trans. Cybern.*, vol. 55, no. 5, pp. 2188–2201, May 2025.
[47] Z. Xie, Y. Lin, Z. Zhang, Y. Cao, S. Lin, and H. Hu, "Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16684–16693.
[48] F. Chen, J. Shao, A. Zhu, D. Ouyang, X. Liu, and H. T. Shen, "Modeling hierarchical uncertainty for multimodal emotion recognition in conversation," *IEEE Trans. Cybern.*, vol. 54, no. 1, pp. 187–198, Jan. 2024.
[49] L. Shen, J. Xu, X. Xu, P. Shen, S. Li, and B. Hooi, "Spherical confidence learning for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 15624–15632.
[50] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
[51] W. Liu, Y. Wen, B. Raj, R. Singh, and A. Weller, "SphereFace revived: Unifying hyperspherical face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2458–2474, Feb. 2023.
[52] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.
[53] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 23803–23828.
[54] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. ECCV*, Oct. 2016, pp. 630–645.
[56] X. Xu, J. Chen, C. Fu, and Z. Lyu, "Hypercomplex neural network and cross-modal attention for multi-modal emotion recognition using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 16, no. 4, pp. 3523–3536, Oct. 2025.
[57] X. Xu, F. Cong, Y. Chen, and J. Chen, "Sleep stage classification with multi-modal fusion and denoising diffusion model," *IEEE J. Biomed. Health Informat.*, vol. 29, no. 12, pp. 9233–9244, Dec. 2025.
[58] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.
[59] 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.
[60] J. Li, S. Qiu, Y.-Y. Shen, C.-L. Liu, and H. He, "Multisource transfer learning for cross-subject EEG emotion recognition," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3281–3293, Jul. 2020.
[61] J. Li, S. Li, J. Pan, and F. Wang, "Cross-subject EEG emotion recognition with self-organized graph neural network," *Frontiers Neurosci.*, vol. 15, Jun. 2021, Art. no. 611653.
[62] Q. Li, T. Zhang, C. L. P. Chen, K. Yi, and L. Chen, "Residual GCB-net: Residual graph convolutional broad network on emotion recognition," *IEEE Trans. Cognit. Develop. Syst.*, vol. 15, no. 4, pp. 1673–1685, Dec. 2023.
[63] T. Song, S. Liu, W. Zheng, Y. Zong, and Z. Cui, "Instance-adaptive graph for EEG emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 2701–2708.
[64] Y. Li et al., "GMSS: Graph-based multi-task self-supervised learning for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2512–2525, Jul. 2023.
[65] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.
[66] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2960–2967.
[67] Y. Li, W. Zheng, Y. Zong, Z. Cui, T. Zhang, and X. Zhou, "A bi-hemisphere domain adversarial neural network model for EEG emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 12, no. 2, pp. 494–504, Apr. 2021.
[68] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1290–1301, Jul. 2022.
[69] H. Chen, M. Jin, Z. Li, C. Fan, J. Li, and H. He, "MS-MDA: Multisource marginal distribution adaptation for cross-subject and cross-session EEG emotion recognition," *Frontiers Neurosci.*, vol. 15, Dec. 2021, Art. no. 778488.
[70] B.-Q. Ma, H. Li, W.-L. Zheng, and B.-L. Lu, "Reducing the subject variability of EEG signals with adversarial domain generalization," in *Proc. Int. Conf. Neural Inf. Process.*, 2019, pp. 30–42.
[71] Z. Li et al., "Dynamic domain adaptation for class-aware cross-subject and cross-session EEG emotion recognition," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 12, pp. 5964–5973, Dec. 2022.
[72] H. Wang et al., "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.
[73] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.
[74] S. Li and X.-J. Wang, "Hierarchical timescales in the neocortex: Mathematical mechanism and biological insights," *Proc. Nat. Acad. Sci. USA*, vol. 119, no. 6, Feb. 2022, Art. no. 2110274119.

Authorized licensed use limited to: Universidade de Macau. Downloaded on March 20,2026 at 23:55:09 UTC from IEEE Xplore. Restrictions apply.