Heterogeneous Modality Dynamic Trustworthy Fusion Network for Cross-Subject Sleep Stage Classification

Kun Wang, Qi Zhu, *Member, IEEE*, Junyong Zhao, Chuhang Zheng, Wei Shao, and Daoqiang Zhang, *Senior Member, IEEE*

***Abstract*—Sleep stage classification plays a pivotal role in clinical practice, significantly enhancing the efficiency of sleep assessment and assisting in the diagnosis of sleep disorders. However, existing sleep stage classification methods are susceptible to heterogeneous multi-modal sleep data and the distribution drift among subjects. Additionally, sleep is a dynamic process, inherently encompassing uncertainty at both temporal and sample levels, which presents challenges for achieving highly trustworthy identification of sleep stages. To address these issues, we propose a Heterogeneous Modality Dynamic Trustworthy fusion network, called HMDT-Net, for cross-subject sleep stage classification. Specifically, we first develop a heterogeneous modality dynamic trustworthy fusion strategy with self-attention and cross-attention mechanisms. It effectively integrates heterogeneous multi-modal data and adaptively estimates uncertainty across different time windows. Second, our approach incorporates adversarial learning to extract a common representation of features across subjects while enhancing the discriminative power of these features. This mitigates the influence of inter-subject variations on the model. Additionally, we introduce a self-paced learning (SPL) mechanism to mitigate the influence of challenging training samples, identified through alignment loss and domain discrimination loss, on the classification model. Finally, we utilize a multi-layer perceptron (MLP) to classify the fused features extracted. Extensive experiments conducted on three multi-modal sleeping datasets containing EEG and EOG data, namely Sleep-EDF-20, Sleep-EDF-78, and SHHS, demonstrate that our proposed method surpasses several state-of-the-art sleep stage classification techniques.**

***Index Terms*—Sleep stage classification, heterogeneous modality fusion, cross-subject alignment, temporal uncertainty, self-paced learning.**

Received 27 October 2025; accepted 14 December 2025. Date of publication 12 January 2026; date of current version 26 March 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62136004, Grant 62276130, Grant 62376123, Grant 62006115, Grant 62076129, and Grant 62371234, in part by the National Key RD Program of China under Grant 2023YFF1204803, in part by the Key Research and Development Plan of Jiangsu Province under Grant BE2022842, in part by Key Research and Development Plan in Jiangsu (Social Development) under Grant BE2022677, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20231438. *(Corresponding authors: Qi Zhu; Daoqiang Zhang.)*

The authors are with the Key Laboratory of Brain-Machine Intelligence Technology, College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Ministry of Education, Nanjing 211106, China (e-mail: zhuqinuaa@163.com; dqzhang@nuaa.edu.cn).

Recommended for acceptance by Y. Zhang.

Digital Object Identifier 10.1109/TETCI.2025.3647679

I. Introduction

SLEEP stage classification is essential for understanding the physiological mechanisms of sleep, evaluating sleep quality, and diagnosing sleep disorders [1], garnering increasing attention in the intersection of artificial intelligence and medicine. Most of the sleep stage classification techniques rely on analyzing physiological signals, including electroencephalography (EEG) [2], electrooculography (EOG) [3], and electromyography (EMG) [4]. According to the rules established by the American Academy of Sleep Medicine (AASM) [5], experts divide standard polysomnography (PSG) data into five sleep stages: Wake (W), Non-Rapid Eye Movement (N1, N2, and N3), and Rapid Eye Movement (REM). However, heterogeneous sleep data and uncertainty at both temporal and sample levels affect the robustness and generalization ability of sleep stage classification models [6], [7], [8]. Therefore, it is crucial to investigate effective multi-modal cross-subject classification models to enhance the accuracy and reliability of sleep stage classification.

In recent years, deep learning has shown remarkable advantages and has been widely used in sleep stage classification tasks. For example, Eldele et al. [8] utilized a multi-resolution convolutional neural network to extract time-invariant features and capture temporal dependencies. Supratak et al. [9] proposed a convolutional neural network to extract local features and bidirectional long short-term memory networks to automatically learn transition rules between sleep stages. Fu et al. [10] introduced a deep neural network model based on time-frequency fusion and attention mechanisms, extracting feature information from both the time and frequency domains. EOG signals can independently capture sleep-related eye movement patterns and provide competitive performance for sleep staging [11], [12], [13]. Fan et al. [11] proposed an automatic sleep staging method based on EOG signals using a two-scale convolutional neural network combined with a recurrent neural network. Maiti et al. [12] proposed a SE-Resnet-Transformer model for automated sleep stage classification using EOG signals, improving accessibility and reducing reliance on EEG for sleep monitoring. Furthermore, Zhou et al. [14] proposed using a single-modal EOG to generate feature representations for automatic sleep monitoring. Rahman et al. [15] exploited a method utilizing discrete wavelet transform and neighborhood component analysis to achieve automatic sleep stage classification of single-channel

EOG. However, these methods rely solely on extracting features from single-modal sleep stage data, thus being unable to fully leverage the complementary information present in different modalities such as EEG and EOG, to promote the classification performance.

Multi-modal signal fusion often exhibits complementary information, where information from one modality can compensate for the limitations of others [16]. Although EEG and EOG are recorded from the same subject, their distinct physiological origins and feature subspaces can cause misalignment when directly fused, impairing sleep state classification accuracy. Therefore, more and more methods have been proposed to solve sleep stage classification tasks by using multi-modal physiological signals (i.e., EEG, EOG, and so on) [17], [18]. For example, Li et al. [19] proposed learning spatiotemporal representations by combining time-frequency images and graph learning with VGG-16 and gated recurrent unit networks. Lu et al. [20] developed a lightweight U-Net network that adaptively recalibrates features using joint channels and spatial attention. Dai et al. [21] introduced MultiChannelSleepNet based on the transformer encoder for sleep stage classification. The discrepancies among multiple datasets could be due to various factors, including heterogeneous patient populations, different signal channels, diverse types or methods of data collection equipment [22]. In clinical settings, a sleep staging model should be well-generalized to unseen datasets from any new populations in any environment.

Numerous studies have focused on enhancing cross-subject classification performance [23], [24]. In multi-subject EEG analysis, individual variability often leads to differences in signal amplitude, temporal patterns, and feature distributions, making traditional single-subject training methods difficult to generalize to other subjects. To address these challenges, cross-subject alignment (CA) strategies are employed to minimize inconsistencies in feature distributions across subjects and ensure data consistency [22], [25]. Therefore, cross-subject alignment methods can improve the model's performance on data from multiple centers [26], which includes introducing adversarial training, domain alignment, or sample re-labeling during the feature extraction stage. For instance, Eldele et al. [27] built an adversarial learning framework, with an attention mechanism to preserve domain-specific features, and used an iterative self-training strategy to enhance classification performance. However, these methods are all based on fixed signal lengths for sleep cycle training and without considering the granularity of time series. Moreover, they fail to consider how the ease or difficulty of sample learning affects the model.

Sleep is a dynamic biological process [28], multiple brain areas work together to maintain periods of sleep and wake [29]. Traditional methods considering temporal dependencies between sleep stages require encoding long temporal sequences, leading to temporal uncertainty [30], [31]. The temporal dependency and representative feature of raw data will be preserved in the small windows slide sequentially [32]. Additionally, sample uncertainty stems from multiple sources of experimental procedures, equipment variations, and missing information [33], [34]. With the presence of a considerable portion of ambiguous samples in large-scale sleep stage classification datasets, the model is unable to learn the sleep characteristics associated with a particular stage. Traditional methods input all samples from the same setting into the model, disregarding the impact of highly uncertain samples on the model.

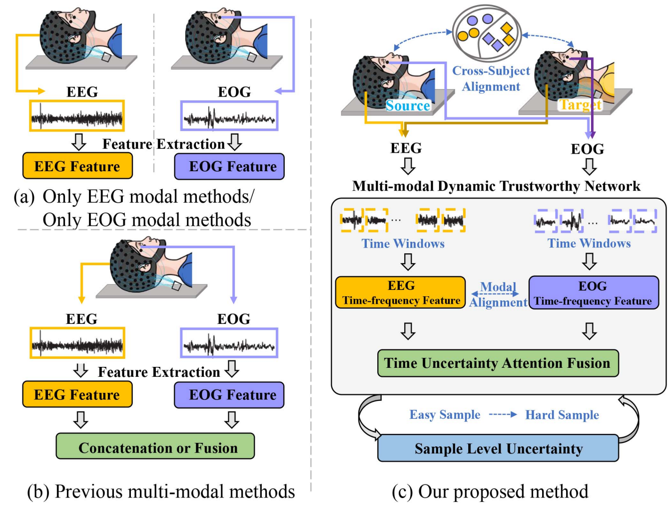

Fig. 1. Illustration of the previous methods and our method. (a) the only EEG modal methods and only EOG modal methods, (b) the previous multi-modal methods, and (c) the proposed heterogeneous modality dynamic trustworthy fusion network.

To address the above challenges, we propose a Heterogeneous Modality Dynamic Trustworthy fusion network, called HMDT-Net, for cross-subject sleep stage classification, which not only effectively integrates heterogeneous sleep data but also mitigates the impact of uncertainty on recognition. Initially, we develop a heterogeneous modality dynamic trustworthy fusion strategy with self-attention mechanisms and cross-attention mechanisms to extract crucial features within modalities and between modalities, respectively. Simultaneously, it divides dynamic sleep data into fine-grained time windows and assesses their uncertainties to guide robust fusion. Secondly, through adversarial learning, we compel the domain classifier to confuse the representation of cross-modal features and enhance the discriminative power of the features through contrastive learning. Moreover, we adaptively estimate uncertainty at the sample level by self-paced learning, gradually feeding samples into the model from low to high uncertainty, thereby reducing the influence of complex samples on the model. Finally, the fused multi-modal sleeping features are classified by multi-layer perceptron. Experimental results on several multi-modal sleeping datasets demonstrate that the proposed method outperforms existing competing methods in sleep stage classification. Fig. 1 illustrates the primary differences between existing sleep stage classification methods and our approach. Fig. 1(a) shows the single-modal sleep stage classification models utilizing either EEG or EOG, which struggle to capture the complementary information between different modalities. In contrast, Fig. 1(b) and (c) respectively depict the existing multi-modal sleep stage classification approaches and our proposed method. Unlike existing multi-modal methods, our approach embeds calibration across subjects within the fusion model and is capable of mitigating the uncertainty at multiple levels of sleep data, thereby facilitating more accurate and robust recognition performance. In brief, the main contributions of our method are summarized as follows:

1) We propose a heterogeneous modality dynamic trustworthy fusion network for cross-subject sleep stage classification. To mitigate the temporal uncertainty inherent in multi-modal sleep data, we adaptively assess the uncertainty across different windows using an attention mechanism. Simultaneously, we employ a cross-attention mechanism to derive a common spatial representation that integrates heterogeneous modalities.
2) An adversarial learning-based cross-subject sleep alignment strategy is devised, which not only projects the sleep data of subjects from the source domain and the target domain into a common space but also enhances the discriminative power of the features through a contrastive learning mechanism.
3) We adopt self-paced learning to adaptively assess the difficulty of the distribution of sleep samples, avoiding the premature involvement of challenging sleep samples in training to alleviate uncertainty at the sleep sample layer and prevent their impact on the robustness of the initial classification model.
4) Experiments across multiple datasets show that our method not only provides promising recognition results but also significantly enhances the robustness of recognition in the challenging clinical scenario of cross-subject multi-modal sleep stage classification.

The remainder of the paper is organized as follows, Section II provides an elaboration on the related works; Section III details the model composition; Section IV presents the experimental settings and results; Section V discusses the experimental findings; Finally, in Section VI, we conclude our work in this paper.

II. Related Works

A. Domain Adaptation

The essence of the problem of cross-subject sleep recognition lies in transferring data from different domains to expand the model's generalization capability and applicability. Domain adaptation typically assumes that the model is pre-trained on a source domain where the target domain data is either completely unlabeled or has very few labeled samples [35]. Then, through various feature alignment techniques, the model adapts to the target domain, learning the mapping between domains without retraining the entire model.

Due to the distribution differences between the source and target domains, a model trained on the source domain may perform poorly when applied to the target domain. These differences may include variations in data marginal distribution [36], class distribution [37], label distribution [38], etc. One common approach in domain adaptation methods is to reduce the feature discrepancy between domains through feature alignment or mapping [39]. This can be achieved by introducing specific layers or mapping functions into the model. Methods for reducing feature distribution discrepancy in domain adaptation primarily fall into two categories: one is by measuring the difference between different feature distributions using distance metrics, such as maximum mean discrepancy [40]. Another method is adversarial training, which reduces the domain differences by introducing adversarial losses. For example, conditional domain adversarial networks [41] and adversarial discriminative domain adaptation [42] fine-tune the target domain feature extractor from the source domain feature extractor to remove specific information from the target domain. Domain-adversarial neural networks create an adversarial relationship between the feature extractor and domain discriminator to eliminate domain drift.

$$E\left(\theta_f, \theta_y, \theta_d\right) = \frac{1}{n}\sum_{i=1}^{n} L_y^i\left(\theta_f, \theta_y\right) - \lambda\left(\frac{1}{n}\sum_{i=1}^{n} L_d^i\left(\theta_f, \theta_d\right) + \frac{1}{n'}\sum_{i=n+1}^{N} L_d^i\left(\theta_f, \theta_d\right)\right) \tag{1}$$

where L_y^i represents the prediction loss on the i-th example, L_d^i denotes domain classification loss, θ_f is the parameter of feature extractor, θ_y is the parameter of label prediction output layer, θ_d is the parameter of domain prediction output layer, λ is the parameter to tune the trade-off between two quantities, n denotes the number of source domain samples, n'represents the number of target domain samples, and $N = n + n'$ indicates the total number of samples from both domains.

B. Self-Paced Learning

Self-paced learning (SPL) [43] is a method that mimics the human learning mechanism, allowing models to dynamically adjust the learning process based on the difficulty and confidence level of the samples. The objective of this approach is to gradually increase the difficulty of samples, enabling the model to adapt to increasingly complex scenarios, thereby enhancing learning efficiency and generalization capability. With the increase in iteration, all samples are gradually included in the learning process to better adapt the model to the training data. Takase et al. [44] proposed a self-paced augmentation method, which dynamically selects appropriate samples for data augmentation while training neural networks.

Zhu et al. [45] employed a self-paced learning strategy to optimize the emotion recognition model, mitigating the influence of hard-learning samples.

$$(\omega^*, v^*) = \text{argmin} \sum_{i}^{n}(v_i L(x_i, y_i, \omega))) + \Omega(\omega) - \lambda f(v) \tag{2}$$

where (ω^*, v^*) denotes the optimal values of the model parameters and the sample selection variables, $F(v)$ represents the self-paced regularizer, which encourages the progressive inclusion of samples into training as the pacing parameter λ increases. The training set consists of n samples, $v \in [0, 1]$, ω represents the parameters of the model to be optimized, where x_i and y_i represent the sample and its class label, respectively, $L(x_i, y_i, \omega)$ denotes the loss function, and $\Omega(\omega)$ represents the regularization term, and λ is the self-learning parameter, used to determine which samples to include for learning in the current

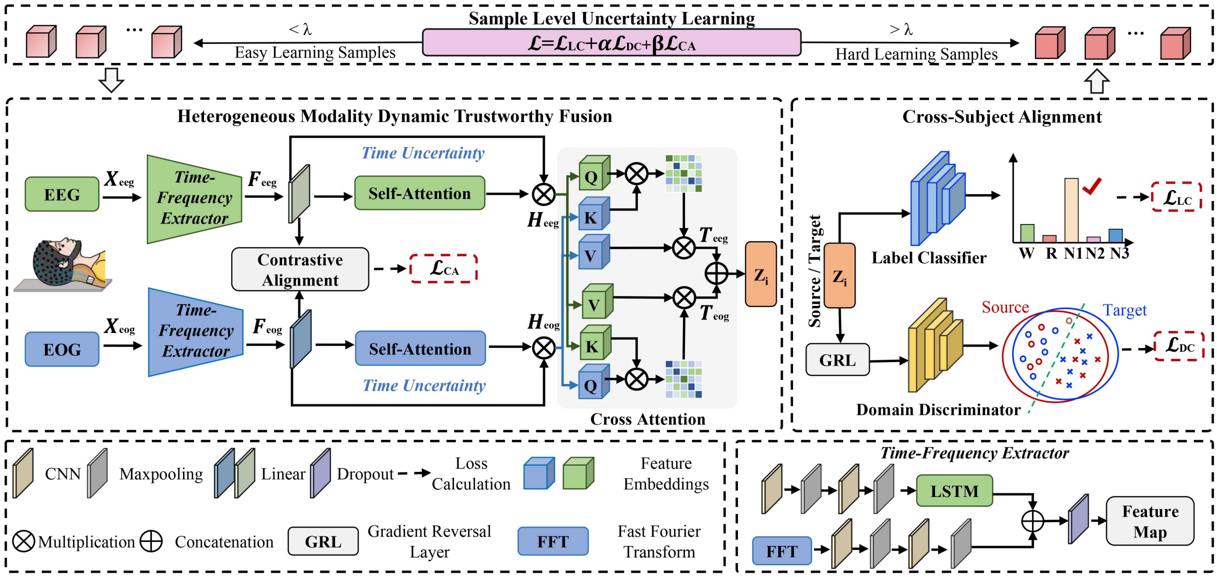

Fig. 2. The proposed architecture for sleep stage classification. EEG (X_{eeg}) and EOG (X_{eog}) signals are processed by time-frequency extractors to obtain intra-modal temporal-spectral features F_{eeg} and F_{eog}. A time uncertainty attention module reweights reliable temporal windows, and cross-attention fusion generates shared representations Z_i. The fused features are then fed into a label classifier for sleep stage prediction and a domain discriminator via a GRL for cross-subject alignment. Meanwhile, self-paced learning adaptively incorporates samples from easy to hard, thereby enhancing model robustness and reliability.

iteration. By gradually increasing the value of λ, more and more samples are selected until all samples are included in the model.

III. Proposed Method

In this section, we present a cross-subject multi-modal dynamic trustworthy network tailored for sleep stage classification. As shown in Fig. 2, the proposed network framework encompasses heterogeneous modality dynamic trustworthy fusion, cross-subject alignment, and sample level uncertainty learning. The heterogeneous modality dynamic trustworthy fusion module contains time-frequency feature extractor and time uncertainty attention fusion module. The time-frequency feature extractor to extract the intra-modal feature. The time uncertainty attention fusion module dynamically evaluates the reliability of each temporal segment through a self-attention mechanism, reweighting time windows to suppress unstable or noisy intervals. Meanwhile, the cross-attention fusion module integrates complementary EEG and EOG features. In addition, cross-subject alignment and sample level uncertainty learning enhance the model's robustness. These specific modules are detailed as follows.

A. Heterogeneous Modality Dynamic Trustworthy Fusion

1) Time-Frequency Feature Extractor: To better obtain the multi-modal feature, we propose a time-frequency feature extractor to extract the intra-modal features. The input multi-modal signals (EEG signal X_{eeg} and EOG signal X_{eog}) are first fed into the time-frequency feature extractor to obtain EEG features F_{eeg} and EOG features F_{eog}, respectively. Specifically, the time-frequency feature extractor includes two parallel pathways. On one pathway, the signal is fed into a series of convolutional layers and pooling layers, followed by a long short-term memory (LSTM) network to capture long-term dependencies in the time dimension effectively. On the other pathway, the signal is fed into the Fast Fourier Transform (FFT) layer to obtain frequency domain signal and then is fed into a series of convolutional layers and pooling layers to obtain frequency feature. Finally, the time and frequency domain features are concatenated, followed by Dropout layers to obtain the EEG feature F_{eeg} and EOG feature F_{eog}, respectively. All convolutional layers are followed by the batch normalization layer and Gaussian Error Linear Unit (GELU) activation function layer. To enhance the generalization performance of the model, batch normalization is employed to normalize the data, and the GELU activation function is utilized to mitigate the vanishing gradient issue. Therefore, $F_m F_{eog}$ is defined as,

$$F_m = [LSTM(conv_{maxpool}(X_m), conv_{maxpool}(FFT(X_m))] \tag{3}$$

where $conv_{maxpool}(.)$ represents a series of convolutional and pooling layers, and $[.]$ denotes the concatenation operator. m represents eeg and eog, respectively.

2) Time Uncertainty Attention Fusion Module: Time window is a strategy which is widely used to capture the dynamics characteristics of brain activity within a defined period and extract the temporally resolved features [46], [47]. Sleep monitoring is a long-term dynamic process, and with the intricate complexities of brain neural activity mechanisms, uncertainties may emerge within specific time windows. Specifically, the dynamic variations in sleep-related EEG/EOG signals across different time windows present a challenge for the model to consistently capture stable temporal dependencies. To address this issue, our model employs a temporal uncertainty attention fusion module based on a self-attention mechanism, which adaptively adjusts the contribution of each time window based on its reliability.

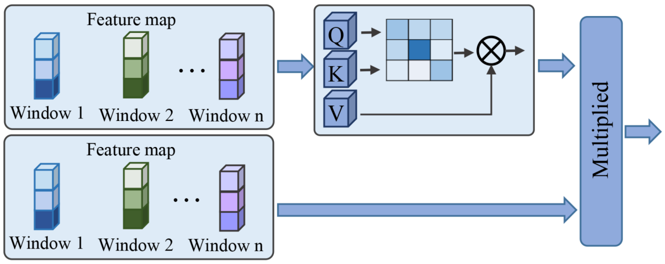

Fig. 3. Illustration of time uncertainty.

This mechanism is shown in Fig. 3, which allows the model to adaptively estimate the significance of sleep stage classification in different time windows, thereby reducing the negative impact caused by highly uncertain time series features. Thus, the EEG uncertainty feature H_{eeg} and the EOG uncertainty feature H_{eog} are defined as follows,

$$Q_m = F_m W^i_{Qm}, K_m = F_m W^i_{Km}, V_m = F_m W^i_{Vm} \tag{4}$$

$$H_m = Softmax\left(\frac{Q_m K_m^T}{\sqrt{d}}\right) V_m \tag{5}$$

where, Q_m, K_m, V_m are learned Query, Key, and Value, respectively, $W^i_{Qm}, W^i_{Km}, W^i_{Vm}$ are parameter matrices, respectively, and d represents the feature dimension. In addition, we adopt the cross-attention to fuse the multi-modal feature. Thus, the cross-attention T_{eeg} and T_{eog} are calculated by:

$$T_m = Softmax\left(\frac{Q_m K_m^T}{\sqrt{d}}\right) V_m \tag{6}$$

Then, the T_{eeg} and T_{eog} are added to obtain final fusion features Z_i.

Meanwhile, to overcome the heterogeneity from different modalities and further improve classification performance, we propose a modal contrastive alignment loss L_{CA} to map the features extracted from two modalities to a common latent space. The modal similarity is defined as,

$$P_{eeg}(F_{eeg}) = \frac{exp(sim(F_{eeg}, F^a_{eog})/\tau)}{\sum_{i=1}^{A} exp(sim(F_{eeg}, F^i_{eog})/\tau)} \tag{7}$$

$$P_{eog}(F_{eog}) = \frac{exp(sim(F_{eog}, F^a_{eeg})/\tau)}{\sum_{i=1}^{A} exp(sim(F_{eog}, F^i_{eeg})/\tau)} \tag{8}$$

$$\begin{aligned} L_{CA} = \frac{1}{2} E_{(F_{eeg}, F_{eog})\sim D}[&H(P'_{eeg}(F_{eeg}), P_{eeg}(F_{eeg})) \\ &+ H(P'_{eog}(F_{eog}), P_{eog}(F_{eog}))] \end{aligned} \tag{9}$$

where H represents the cross-entropy loss function, $sim(.)$ denotes the cosine similarity function, τ represents the temperature parameter, F^a_{eeg}, F^a_{eog} denote positive samples, F^i_{eeg}, F^i_{eog} represent negative samples, D denotes data distribution, and $P'_{eeg}(F_{eeg})$ and $P'_{eog}(F_{eog})$ represent the ground-truth similarity. The positive pair between the EEG and EOG is set to 1; otherwise, the negative pair is set to 0.

B. Cross-Subject Alignment

Cross-subject scenarios are prevalent in real applications of sleep stage classification, and the resulting differences in data distribution across different subjects pose challenges for classification. Here, we introduce cross-subject alignment to minimize variations in data distribution while enhancing feature discriminability. Specifically, this alignment method incorporates a label classifier and a domain discriminator, both utilizing Multi-Layer Perceptron (MLP) techniques.

Label classifier is employed to decode the feature representation learned by the feature fusion. As training progresses, labeled data from the source domain and unlabeled data from the target domain are used for training. As training progresses, the cross-entropy loss is utilized to minimize the difference between the predicted labels by the classifier model and the ground truth labels. The label classification loss L_{LC} is defined as,

$$L_{LC} = H\left(y, y^{predict}\right) = -\sum_{i=1}^{N} y_i \log y_i^{predict} \tag{10}$$

where H represents the cross-entropy loss function, y denotes the ground truth label, $y^{predict}$ represents the predicted label, and N denotes the total number of samples classified.

The domain discriminator is used to distinguish whether these learned features come from the source domain or the target domain. We use adversarial training to implement the game between the feature extractor and domain discriminator on the source domain and target domain. Within the framework of adversarial training, the goal of the feature extractor is to learn features that confuse the domain discriminator and exhibit certain generalizations and invariance. Specifically, we aim for the domain discriminator to be unable to accurately distinguish whether the learned features originate from the source or target domain.

To achieve this goal, we introduce a Gradient Reversal Layer (GRL) between the feature extraction and the domain discriminator. This approach enables adversarial learning through gradient reversal, thereby promoting more robust feature learning across different domains. In the adversarial learning module, the source domain refers to the labeled data from training subjects, while the target domain corresponds to the unlabeled data from unseen subjects. The GRL-based adversarial learning strategy aims to minimize the feature distribution discrepancy between the source and target domains [48]. The feature extractor is optimized to generate domain-invariant representations that confuse the domain discriminator, thereby promoting cross-subject generalization. Thus, to reduce the difference between the source domain and the target domain, we utilize cross-entropy loss to minimize the difference between the domain labels predicted by the domain discriminator model and the true domain labels.

$$L_{DC} = -\sum_{i=1}^{N} d_i \log(d_i^{predict}) + (1 - d_i)\log(1 - d_i^{predict}) \tag{11}$$

where d_i denotes the domain label for the i-th sample. An input sample is considered to originate from the source domain when its domain label is $d_i = 0$. Conversely, if $d_i = 1$, it indicates that the input sample is derived from the target domain. $d_i^{predict}$ represents the predicted label of the domain discriminator. The

GRL layer reverses the direction of gradients during backpropagation. This means that in the domain discriminator, the gradient direction for the feature extractor of the target domain is opposite to that of minimizing domain loss, thereby achieving the goal of adversarial training.

C. Sample Level Uncertainty Learning

Sample level uncertainty arises from individual differences or inconsistent annotations. To mitigate this uncertainty, we employ a self-paced learning (SPL) strategy inspired by the human cognitive learning process that dynamically adjusts the difficulty of training samples based on model confidence, thereby improving learning efficiency and model generalization. Specifically, SPL assigns a dynamic weight to each training sample according to its loss value. Samples with smaller losses (i.e., higher confidence) are learned first, while uncertain samples are incorporated later as the model becomes more robust. By progressively controlling the learning pace, SPL helps the network stabilize training, alleviates the impact of noisy samples, and enhances the overall generalization ability of the model. The introduction of SPL enables the above model to autonomously select the most easily classified training samples. By progressively introducing samples with increasing difficulty, SPL allows the model to start learning from simpler samples and gradually transition to more complex ones. This helps the model establish a stronger foundational knowledge and progressively improve its proficiency during the learning process. The model automatically adjusts its learning pace based on its performance on the training set. Initially, the model may focus more on easier-to-learn samples, and then gradually introduce more challenging samples as its performance improves. Specifically, we suppose $v = [v_1, \ldots, v_n]^T \in [0,1]^n$ denotes the vector of sample weight variables based on the training loss. The sample uncertainty is calculated by minimizing the following function:

$$\min_{\omega, v \in [0,1]} E(\omega, v, \lambda) = \sum_{i=1}^{n} (v_i L(x_i, y_i, \omega) + f(v_i, \lambda)) \tag{12}$$

where $f(v_i, \lambda)$ denotes the self-paced regularizer, which controls the pace of sample selection. It is typically designed such that, as the self-paced parameter λ increases, more samples are gradually incorporated into the training process. n represents sample number, ω represents the parameters of the model to be optimized, given the training dataset $\{(x_1, y_1), \ldots, (x_i, y_i)\}$, where x_i, y_i represent the i-th observed sample and its label, respectively, $L(x_i, y_i, \omega)$ denotes the loss function, and λ is the self-learning parameter for controlling the learning pace, used to determine which samples to include for learning in the current iteration. Repeat the process until all samples have been added to the model. When ω is fixed, the optimal v^* can be obtained.

$$v^* = \begin{cases} 1, & L(x_i, y_i, \omega) < \lambda \\ 0, & otherwise \end{cases} \tag{13}$$

Samples with losses less than a certain threshold λ are considered as "easy" samples and are selected for training $(v^* = 1)$, otherwise unselected $(v^* = 0)$. When v is fixed, the optimal ω^* can be obtained, the training on selected "easy" samples. When λ is small, only "easy" samples with small losses are considered. As λ grows, more and more hard-learning samples are added to train more mature models.

The overall loss function is defined as:

$$L = L_{LC} + \alpha L_{CA} + \beta L_{DC} \tag{14}$$

where L_{LC} represents the label classification loss, and is utilized to minimize the difference between the predicted labels by the classifier model and the ground truth labels. L_{CA} denotes the contrastive alignment loss, which is used to preserve the similarity of different modalities, L_{DC} represents domain discriminator loss, and is used to minimize the difference between the domain labels predicted by the domain discriminator model and the true domain labels. α represents the weight parameter for the contrastive learning loss, and β denotes the weight parameter for the domain discriminator loss.

TABLE I
INTRODUCTION TO THE DATASETS

Datasets	Channel	Subjects	W	N1	N2	N3	REM	Samples
Sleep-20	Fpz-Cz Pz-Oz EOG	20	8285 19.6%	2804 6.6%	17799 42.1%	5703 13.5%	7717 18.2%	42308
Sleep-78	Fpz-Cz Pz-Oz EOG	78	65951 33.7%	21522 11.0%	69132 35.4%	13039 6.7%	25835 13.2%	195479
SHHS	C4-A1 C3-A2 EOG	100	14671 14.8%	2926 3.0%	42433 42.7%	19039 19.2%	20198 20.3%	99267

IV. EXPERIMENTS

A. Datasets

We evaluated the performance of our HMDT-Net method on the three multi-modal sleep stage classification datasets, including Sleep-EDF-20, Sleep-EDF-78, and SHHS. All these datasets include horizontal EOG and EEG from two channels Fpz-Cz and Pz-Oz. Specifically, the Sleep-EDF-20 comprises data from 20 healthy subjects, the Sleep-EDF-78 includes data from 78 subjects, and the SHHS dataset consists of data from 100 subjects [11]. Sleep-EDF-78 is an extension of Sleep-EDF-20, comprising 197 full-night polysomnographic (PSG) recordings, i.e., 153 polysomnography (PSG) recordings and 44 sleep telemetry (ST) recordings. These data are obtained from both male and female individuals aged between 25 and 101 years old, who were not under any medication. The sampling rates for Sleep-EDF-78 and Sleep-EDF-20 are 100 Hz, while for SHHS, it is 125 Hz. Specific details of the datasets are provided in Table I. Sleep experts observe different time sequences of nocturnal sleep records and name each 30-second time segment as an epoch according to reference naming conventions such as AASM [5] rules or R & K rules [49], as experts use epoch labels for marking. Additionally, we only consider data from the 30 minutes after sleep onset and the 30 minutes before awakening.

B. Experimental Setup

Each dataset is randomly divided into an 80% training set and a 20% test set. For example, in the Sleep-EDF-20 dataset, 16 subjects are used for training and 4 subjects for testing. This operation is repeated 5 times. In our experiment, the EEG data are divided into successive time windows of equal length (t=10 s) to capture the dynamic characteristics of EEG signals. It is important to note that each electrode signal is a non-overlapping subsequence. Our model is trained using cross-entropy loss, with different weights assigned to losses for different classes to address data imbalance issues. The learning rate for the Adam optimizer during model training is set to 0.0001, with a batch size of 128, and the number of epochs is set to 50. Our code is available at: https://github.com/EEGBrainNet/HMDT-Net.

In the experiments, we employ the accuracy (ACC), Cohen's Kappa (Kappa), precision, recall, F1-score (F1), and macro-averaged F1-score (MF1) to evaluate the performance of the proposed HMDT-Net for sleep stage classification on three datasets. Specifically, the precision, recall, Cohen's Kappa (Kappa), accuracy (ACC), F1-score (F1), and macro-averaged F1-score (MF1) are defined as,

$$ACC = \frac{TP + FN}{TP + FP + TN + FN} \tag{15}$$

$$Kappa = \frac{ACC - P_e}{1 - P_e} \tag{16}$$

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{19}$$

$$MF1 = \frac{1}{N} \sum_{i=1}^{N} \frac{2 * Precision_i * Recall_i}{Precision_i + Recall_i} \tag{20}$$

where TP denotes true positives, FP represents false positives, TN denotes true negatives, FN represents false negatives, N denotes the type of sleep stage, and P_e represents the hypothetical probability of chance agreement.

C. Comparison Methods

To validate the effectiveness of our HMDT-Net method, we compare it with several sleep stage classification methods, including DeepSleepNet [50], Attention-Based Deep Learning Approach for Sleep Stage Classification (AttnSleep) [8], Temporal-Spectral fused and Attention-based deep neural Network model (TSA-Net) [10], multi-resolution convolutional neural network (MRCNN) [8], MultiChannelSleepNet [21], Adversarial Domain Adaptation framework based on preserving attention mechanism and iterative self-training strategy (ADAST) [27], CNN-LSTM, Minimum Discrepancy Estimation for Deep Domain Adaptation (MDDA) [51], Adversarial Discriminative Domain Adaptation (ADDA) [52], and Domain Adaptation Neural Network (DANN) [41]. Below is a brief overview of these methods.

- *DeepSleepNet:* The method utilizes convolutional neural networks to extract invariant features, and employs bidirectional LSTM to automatically learn the transition rules between sleep stages from EEG signals.
- *AttnSleep:* The method uses a temporal context encoder module to capture temporal dependencies among the extracted features using a multi-head attention mechanism.
- *TSA-Net:* This method consists of a dual-stream feature extractor, feature context learning, and Conditional Random Field (CRF). The dual-stream feature extraction module considers both time and frequency features to provide rich discriminative information for sleep stage classification from EEG signals.
- *MRCNN:* The method uses multi-resolution convolutional neural networks to extract low-frequency and high-frequency features.
- *MultiChannelSleepNet:* Utilizes a Transformer-based feature extraction block to extract features from time-frequency images of each channel separately.
- *ADAST:* A method based on a non-shared attention mechanism to preserve features from two specific domains and improve classification performance by using pseudo-labels from the target domain selected by dual classifiers.
- *CNN-LSTM:* Employs Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) for feature extraction.
- *MDDA:* Adjusts features using CORAL and MMD metrics to minimize the differences between the source and target domains.
- *DANN:* Trains on a large-scale annotated dataset in the source domain and a large-scale unannotated dataset in the target domain, aligning them through standard backpropagation training.
- *ADDA:* Similar to DANN, it performs operations by flipping labels instead of using GRL. ADDA typically adopts a two-stage training strategy, separately training the feature extractor and the discriminator.

D. Results and Analysis

The sleep stage classification performance of our proposed method and the comparison methods on Sleep-EDF-20, Sleep-EDF-78, and SHHS are reported in Tables II, III, and IV. From these tables, it can be observed that our proposed method achieves the best classification performance in all metrics on all three datasets. For example, our proposed method achieves 88.77%, 82.96%, and 84.28% on the three datasets than the best result (i.e., the average accuracies are 85.36%, 81.33%, and 82.68% achieved by TSA-Net) in terms of average accuracies, respectively. In addition, not only does our method achieve the highest average accuracy, but it also achieves the highest average Kappa compared to other algorithms. Furthermore, we also present the classification accuracies of the five categories in Fig. 4 on the Sleep-EDF-20 dataset. As can be seen from Fig. 4, our method is also higher than those of the comparison methods in all class classification metrics. These results demonstrate that the effectiveness of our method in sleep stage classification. The main reason for the superiority of our method is that, compared with other methods, our proposed method not only mitigates the impact of temporal and sample uncertainty on the recognition model, but also alleviates the domain adaptation issue by inter-subject variability. For example, methods such as TSA-Net

TABLE II
EXPERIMENT RESULTS OF DIFFERENT APPROACHES ON SLEEP-EDF-20 DATASET (MEAN± STD%)

Methods	ACC (↑)	Kappa (↑)	MF1 (↑)	Precision (↑)	Recall (↑)
DeepSleepNet	82.76 ± 1.94	73.00 ± 2.48	74.80 ± 1.37	81.47 ± 1.66	68.87 ± 1.98
AttnSleep	81.60 ± 4.65	74.59 ± 6.45	74.59 ± 5.98	76.17 ± 6.16	73.57 ± 6.26
TSA-Net	85.36 ± 3.23	79.70 ± 4.64	79.09 ± 3.69	81.06 ± 4.50	77.91 ± 4.20
MRCNN	77.61 ± 2.13	68.32 ± 1.78	68.23 ± 5.43	72.17 ± 3.28	66.62 ± 6.76
MultiChannelSleepNet	84.98 ± 1.74	78.63 ± 2.68	77.21 ± 3.66	79.24 ± 6.35	75.67 ± 3.06
ADAST	84.31 ± 3.76	78.35 ± 5.30	78.33 ± 4.23	79.03 ± 3.98	77.74 ± 4.73
CNN-LSTM	69.11 ± 3.84	55.47 ± 7.25	54.81 ± 4.12	60.76 ± 2.86	53.86 ± 2.99
MDDA	76.39 ± 2.19	66.36 ± 4.11	65.60 ± 5.87	74.47 ± 1.25	67.45 ± 7.63
ADDA	81.50 ± 2.15	74.64 ± 3.23	74.71 ± 2.32	74.96 ± 1.99	74.79 ± 3.02
DANN	84.45 ± 3.56	78.47 ± 4.97	78.91 ± 3.63	79.81 ± 4.65	78.12 ± 4.47
HMDT-Net (Ours)	**88.77 ± 3.23**	**84.43 ± 4.64**	**82.90 ± 3.69**	**85.17 ± 4.50**	**81.58 ± 4.20**

TABLE III
EXPERIMENT RESULTS OF DIFFERENT APPROACHES ON SLEEP-EDF-78 DATASET (MEAN± STD%)

Methods	ACC (↑)	Kappa (↑)	MF1 (↑)	Precision (↑)	Recall (↑)
DeepSleepNet	79.76 ± 9.86	71.77 ± 14.41	74.28 ± 12.49	76.12 ± 9.92	72.97 ± 10.78
AttnSleep	79.99 ± 4.17	72.13 ± 6.13	71.80 ± 5.30	74.62 ± 3.94	70.65 ± 6.01
TSA-Net	81.33 ± 1.35	74.18 ± 2.24	72.24 ± 4.68	75.11 ± 5.16	70.93 ± 4.40
MRCNN	73.73 ± 6.27	62.66 ± 8.55	61.83 ± 8.31	76.08 ± 7.21	63.09 ± 8.59
MultiChannelSleepNet	80.61 ± 8.63	73.41 ± 11.61	73.11 ± 10.38	77.29 ± 4.83	72.72 ± 9.33
ADAST	79.98 ± 4.14	72.33 ± 4.85	72.29 ± 1.79	73.52 ± 3.48	71.57 ± 1.87
CNN-LSTM	61.53 ± 0.41	46.28 ± 0.78	52.01 ± 0.42	52.69 ± 0.96	52.83 ± 0.49
MDDA	76.15 ± 1.04	66.64 ± 0.56	66.42 ± 0.48	69.23 ± 0.38	66.37 ± 0.83
ADDA	77.49 ± 2.20	68.53 ± 2.18	68.35 ± 1.83	69.42 ± 5.32	68.62 ± 0.89
DANN	75.82 ± 2.14	64.86 ± 3.64	62.71 ± 4.50	73.99 ± 1.78	61.87 ± 5.00
HMDT-Net (Ours)	**82.96 ± 3.25**	**76.12 ± 4.80**	**75.43 ± 4.69**	**77.78 ± 4.62**	**75.31 ± 4.54**

TABLE IV
EXPERIMENT RESULTS OF DIFFERENT APPROACHES ON SHHS DATASET (MEAN± STD%)

Methods	ACC (↑)	Kappa (↑)	MF1 (↑)	Precision (↑)	Recall (↑)
DeepSleepNet	80.48 ± 0.33	72.33 ± 0.86	72.10 ± 2.12	76.46 ± 3.38	70.71 ± 0.10
AttnSleep	80.67 ± 5.19	72.81 ± 7.31	71.20 ± 7.73	74.47 ± 6.30	71.77 ± 8.47
TSA-Net	82.68 ± 3.13	76.26 ± 3.57	70.48 ± 4.60	74.07 ± 6.50	71.33 ± 3.67
MRCNN	75.06 ± 1.51	66.25 ± 2.33	61.36 ± 1.54	63.46 ± 1.07	65.31 ± 1.29
MultiChannelSleepNet	81.71 ± 2.85	74.73 ± 4.10	72.97 ± 4.64	75.23 ± 4.02	71.75 ± 3.83
ADAST	81.30 ± 2.17	73.71 ± 3.09	69.71 ± 1.37	77.26 ± 2.18	68.23 ± 0.85
CNN-LSTM	63.53 ± 1.76	46.63 ± 2.71	48.02 ± 2.38	57.33 ± 3.02	46.87 ± 1.89
MDDA	77.17 ± 1.63	66.99 ± 2.26	64.60 ± 1.60	75.68 ± 0.69	64.43 ± 1.42
ADDA	78.24 ± 4.27	68.93 ± 5.67	67.71 ± 4.36	74.72 ± 4.96	67.08 ± 3.79
DANN	81.10 ± 2.48	73.44 ± 3.73	71.38 ± 3.30	72.40 ± 2.49	71.62 ± 3.63
HMDT-Net (Ours)	**84.28 ± 0.98**	**77.39 ± 0.91**	**75.10 ± 2.28**	**79.77 ± 3.57**	**72.46 ± 1.71**

and MRCNN only consider the improvement of classification accuracy and ignore the impact across subjects. Liu et al. [13] proposed a deep neural network combining an MSE-based U-structure with CBAM to extract multi-scale salient features from single-channel EEG signals, achieving strong performance in single-modal EEG classification. However, their experimental setup does not account for cross-subject data shifts. In addition, ADAST, CNN-LSTM, ADDA, and DANN introduce domain adaptation to resolve the differences between different subjects to improve the performance and generalization ability of the model, but ignore the impact of uncertainty in sleep data.

E. Feature Representation Visualization

To intuitively analyze the performance of our method and other methods in sleep stage classification tasks, we plot a visualization analysis of the data distribution in the Sleep-EDF-20 dataset using T-distributed stochastic neighbor embedding (T-SNE) [53]. The T-SNE is a nonlinear dimensionality reduction method that can map high-dimensional data to low-dimensional space to observe the characteristics of the data. In essence, samples that are close in the original feature space remain close in the reduced space, allowing intuitive visualization of clustering behavior. We employ T-SNE to compare the separability of sleep-stage features before and after the application of the proposed fusion and alignment modules. Clearer boundaries and denser clusters in the T-SNE plots indicate that our method effectively enhances feature discrimination and distribution consistency across modalities and subjects. This visualization thus provides qualitative evidence supporting the superiority of our feature representation learning strategy. The result is reported in Fig. 5. Specifically,

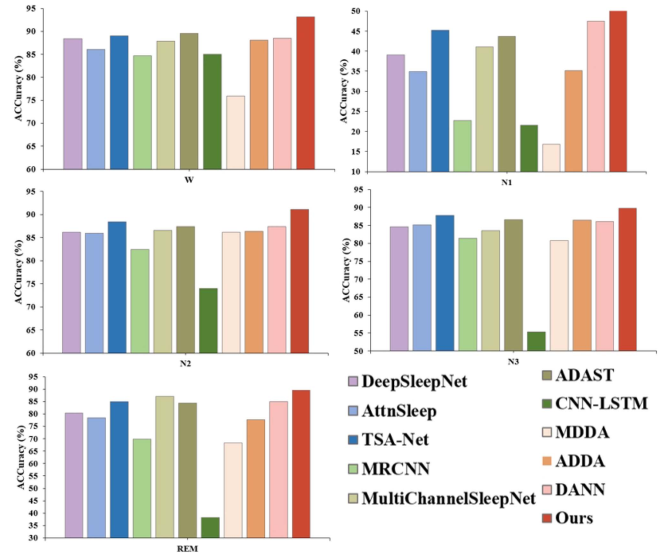

Fig. 4. Experimental results of our method and the competing methods in recognizing five sleep stages on Sleep-EDF-20 dataset.

Fig. 5(a) to (d) depict the distribution of single-modal EEG data, while Fig. 5(e) to (h) depict the distribution of multi-modal data. From Fig. 5(a) and (e), it can be observed that the separability of the original data, whether single-modal or multi-modal, is very poor, making it almost impossible to distinguish between the five types of sleep stages. As is evident from Fig. 5, when compared to the other two feature extraction algorithms, our approach has demonstrated the most pronounced enhancement in class separability across both single-modal and multi-modal tasks. Furthermore, by comparing Fig. 5(d) with (h), it can be seen that the incorporation of multi-modal data can further enhance the classification performance of our method, demonstrating that our method is capable of effectively capturing the complementary information between EEG and EOG in sleep stage classification tasks.

F. Effectiveness of the Multi-Modal Fusion in HMDT-Net

To further validate the effectiveness of the model in multi-modal fusion, we conducted experiments on single-modal EEG data, single-modal EOG data, and multi-modal data from Sleep-EDF-20 using four different methods. From Fig. 6, it is evident that the performance of the multi-modal approach surpasses that of EEG and EOG individually. Due to the weaker signal of EOG data, the classification performance of EEG is higher than that of EOG. Fig. 7 illustrates a comparative example of a Hypnogram from the same subject under different modalities, highlighting the superiority of our research method in sleep stage classification tasks. By observing the charts, we can discern that single-modal EOG leads to a higher classification error rate, particularly during transitions between sleep stages, where its classification accuracy is significantly affected. Although single-modal EEG performs better than single-modal EOG, its performance is still unsatisfactory during subtle transitions across the N1, N2, and N3 sleep stages. Furthermore, these results indicate that methods combining multi-modal information can better address the challenges in sleep stage classification tasks, especially in identifying transitions between different sleep stages. Therefore, the proposed method has significant advantages in sleep stage classification tasks, offering a more reliable staging scheme for the field of sleep research.

TABLE V
ABLATION EXPERIMENTS ON CROSS-SUBJECT MULTI-MODAL SLEEP STAGE CLASSIFICATION PERFORMANCE (MEAN± STD%). (HMDTF: HETEROGENEOUS MODALITY DYNAMIC TRUSTWORTHY FUSION. CA: CROSS-SUBJECT ALIGNMENT. SUL: SAMPLE LEVEL UNCERTAINTY LEARNING.)

HMDTF	CA	SUL	ACC (↑)	Kappa (↑)	MF1 (↑)
			76.65 ± 1.72	66.93 ± 1.55	63.93 ± 1.65
✓			86.45 ± 5.58	81.20 ± 3.73	79.01 ± 3.11
	✓		84.85 ± 1.69	78.63 ± 2.48	78.90 ± 1.81
		✓	80.37 ± 3.15	72.17 ± 4.85	68.84 ± 3.69
✓	✓		87.29 ± 0.70	82.49 ± 0.86	81.82 ± 1.43
	✓	✓	85.03 ± 2.61	79.24 ± 3.70	79.19 ± 2.43
✓		✓	86.76 ± 2.42	81.63 ± 3.43	80.78 ± 4.81
✓	✓	✓	**88.77 ± 3.23**	**84.43 ± 4.64**	**82.90 ± 3.69**

V. DISCUSSION

A. Ablation Study

In this section, we conduct extensive ablation experiments on the Sleep-EDF-20 dataset to investigate several key modules of the proposed HMDT-Net, including heterogeneous modality dynamic trustworthy fusion (HMDTF), sample level uncertainty learning (SUL), and cross-subject alignment (CA). The experimental results are shown in Table V, HMDTF achieves better classification performance, which integrates heterogeneous multi-modal data and adaptively estimates uncertainty across different time windows. Meanwhile, CA can solve variations between subjects and enhance the generalization ability of the models. Additionally, SUL-based self-paced learning further improves the performance for sleep stage classification, which reveals the effect of sample uncertainty on model performance. Removing any module from the proposed method would lead to a decrease in classification performance, which indicates that they promote each other for the diagnosis task.

B. Uncertainty Analysis of Sample and Temporal

Sample level uncertainty learning (SUL) is a strategy that enhances both learning efficiency and the model's generalization ability by dynamically adjusting the difficulty of training samples by the model's confidence [54]. We performed a visualization and analysis of the training process. The experimental setup involved 50 training epochs, with visualizations captured every 10 epochs, as shown in Fig. 8. The results demonstrate that at the beginning of the training process, the model focuses on simpler samples and gradually incorporates more difficult ones as it gains confidence. Sample level uncertainty learning mitigates the risk of overfitting and achieves a significant improvement in classification accuracy.

In addition, we always explore the use of self-attention mechanisms to capture long-range dependencies within sequential

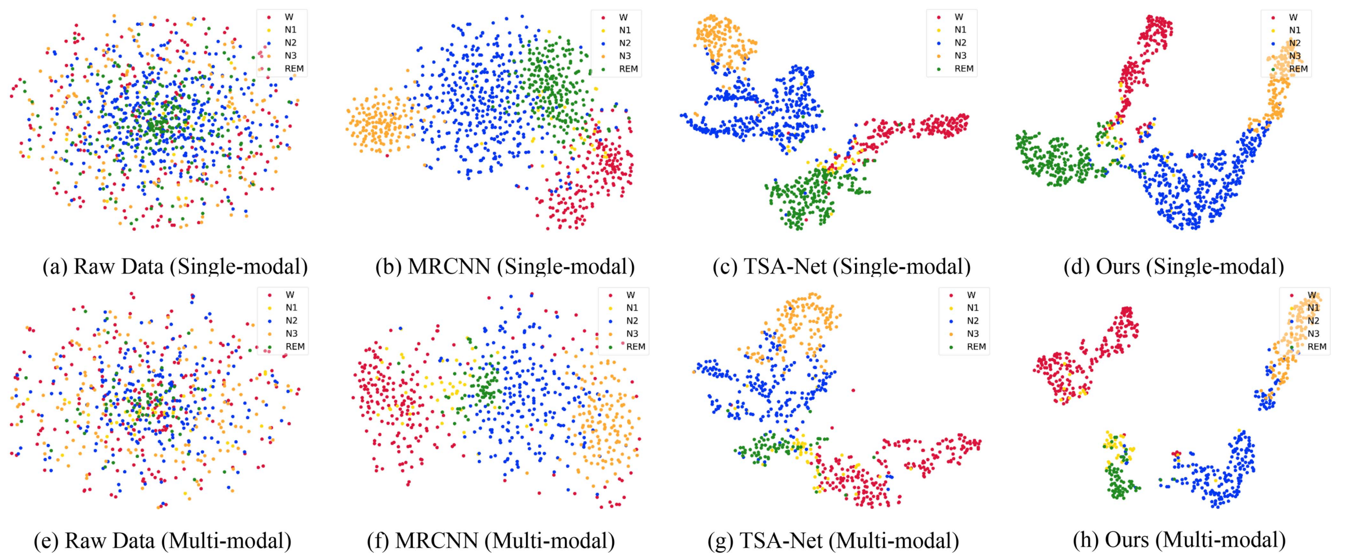

Fig. 5. The T-SNE of raw data and processed data through various feature extraction methods. First row: single-modal EEG Data. Second row: Bimodal EEG and EOG Data.

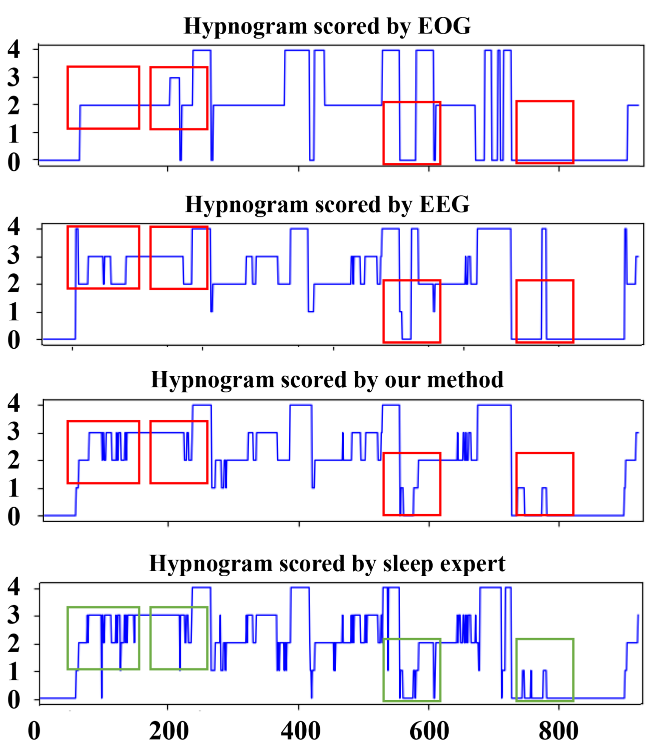

Fig. 6. Experimental results of different approaches on Sleep-EDF-20 dataset in different modalities.

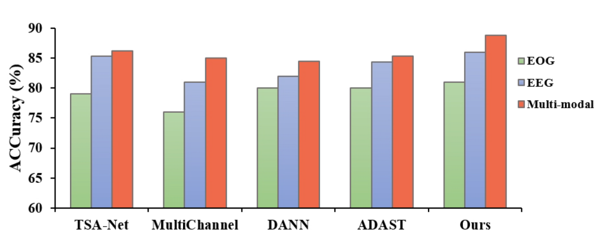

Fig. 7. The comparison of the hypnogram scored by different classification methods.

TABLE VI
COMPUTATIONAL EFFICIENCY OF THE PROPOSED METHOD AND OTHER METHODS

Method	#Param (M)	FLOPs (M)	Test Time / sample (ms)
DeepSleepNet	24.24	17.88	1.4054
AttnSleep	2.53	143.46	0.4946
TSA-Net	1.10	12.34	0.2688
MRCNN	0.96	121.67	0.6385
MultiChannelSleepNet	32.98	194.70	1.0340
ADAST	26.11	52.61	1.1064
HMDT-Net (Ours)	**0.52**	**8.38**	**0.1977**

data [55]. Attention mechanisms allow the model to focus on and select the most relevant features throughout the entire sequence for the final prediction [56]. Self-attention allows the model to focus on the relationships between different time windows, enabling it to learn the importance of each time window. The distribution of self-attention weights reflects the model's level of attention to different time windows. We present heat maps of uncertainty across different time windows in Fig. 9, and the experimental results show that the self-attention mechanism can dynamically adjust the weight of each time window based on its relevance. This dynamic adjustment significantly enhances the model's performance by reducing the impact of uncertainty, leading to more accurate sleep staging predictions.

C. Model Complexity Analysis

We also explore the computational resources required by the proposed method. We present the model parameters, floating-point operations (FLOPs) for our proposed method and other comparison methods. In addition, test time is a crucial metric for real-world deployment, especially in clinical settings, and we also report the test time for each sample. The results are reported in Table VI.

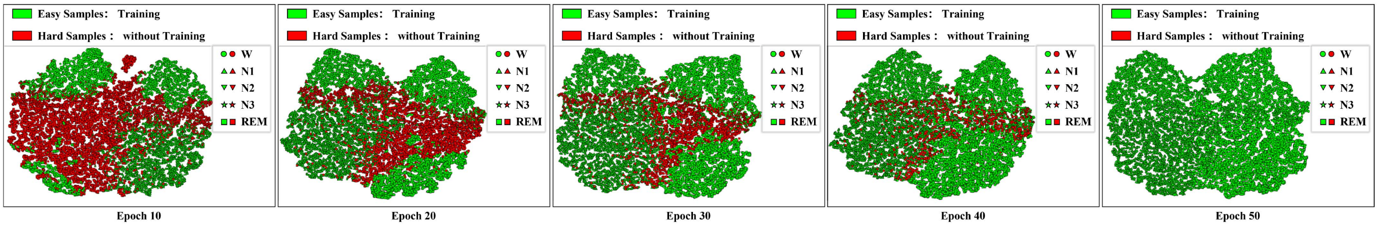

Fig. 8. Visualization of the dynamic learning process with sample uncertainty.

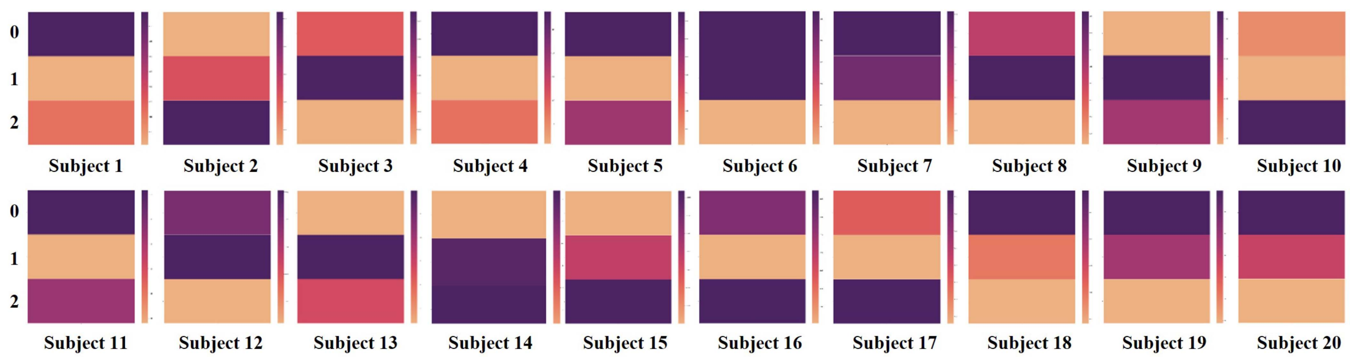

Fig. 9. Visualization of the attention weights for twenty subjects on the Sleep-EDF-20 dataset across all time windows.

As shown in Table VI, the proposed method requires only 0.52 M model parameters and 8.38 M floating-point operations, which is better than other competing methods. Additionally, it takes 0.1838 ms for each sample. Therefore, our method can achieve superior performance with fewer computational resources.

D. Limitations and Future Work

Although the proposed HMDT-Net method demonstrates superior performance in sleep stage classification across three datasets, there are summarized limitations and solutions methods as follows. Firstly, our method for sleep stage classification is currently based on EEG and EOG data. In the future, we plan to incorporate additional modalities, to enhance the model's capacity for sleep stage classification and improve its robustness. Secondly, the spatial topology of sleep provides crucial prior information about the brain's functional organization and the coordination of cross-regional activities during different sleep stages. Therefore, we will focus on leveraging spatial topological information between brain regions to improve the accuracy and reliability of both classification and analysis in further work.

VI. Conclusion

In this paper, we propose a heterogeneous modality dynamic trustworthy fusion network for cross-subject sleep stage classification, which integrates heterogeneous EEG and EOG sleep data from different subjects and mitigates uncertainty at multiple levels. Firstly, we devise a multi-modal fusion strategy grounded in uncertainty estimation. This strategy employs a self-attention mechanism to dynamically assess the importance of time windows within each modality and integrates data across modalities through the cross-attention mechanism. Secondly, our method incorporates adversarial learning to capture a common feature representation across subjects, which enhances the features' ability to discriminate between classes and reduces the impact of individual differences on the model. Furthermore, to suppress uncertainty at the sample level, we adopt self-paced learning to adaptively evaluate the difficulty of learning samples, gradually incorporating hard-learning samples into the model. Finally, we employ an MLP to classify the features extracted by our proposed method. Experimental results on several multi-modal sleep datasets show that our method outperforms the state-of-the-art methods in cross-subject classification performance, thereby facilitating more effective and personalized interventions for sleep-related disorders.

References

[1] A. Coutrot et al., "Reported sleep duration reveals segmentation of the adult life-course into three phases," *Nature Commun.*, vol. 13, no. 1, 2022, Art. no. 7697.

[2] H. Wang, H. Guo, K. Zhang, L. Gao, and J. Zheng, "Automatic sleep staging method of EEG signal based on transfer learning and fusion network," *Neurocomputing*, vol. 488, pp. 183–193, 2022.

[3] M. Tashakori et al., "Interhemispheric differences of electroencephalography signal characteristics in different sleep stages," *Sleep Med.*, vol. 117, pp. 201–208, 2024.

[4] B. Yang, W. Wu, Y. Liu, and H. Liu, "A novel sleep stage contextual refinement algorithm leveraging conditional random fields," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 2505313.

[5] C. Iber, "The AASM manual for the scoring of sleep and associated events: Rules, terminology, and technical specification," *Amer. Academy Sleep Medicine*, Westchester, 2007.

[6] L. W. Hermans et al., "Representations of temporal sleep dynamics: Review and synthesis of the literature," *Sleep Med. Rev.*, vol. 63, 2022, Art. no. 101611.

[7] J. Han et al., "IoT-V2E: An uncertainty-aware cross-modal hashing retrieval between infrared-videos and EEGs for automated sleep state analysis," *IEEE Internet Things J.*, vol. 11, no. 3, pp. 4551–4569, Feb. 2024.

[8] E. Eldele et al., “An attention-based deep learning approach for sleep stage classification with single-channel EEG,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 809–818, 2021.

[9] A. Supratak, H. Dong, C. Wu, and Y. Guo, “DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, Nov. 2017.

[10] G. Fu, Y. Zhou, P. Gong, P. Wang, W. Shao, and D. Zhang, “A temporal-spectral fused and attention-based deep model for automatic sleep staging,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1008–1018, 2023.

[11] J. Fan, C. Sun, M. Long, C. Chen, and W. Chen, “EOGNET: A novel deep learning model for sleep stage classification based on single-channel EOG signal,” *Front. Neurosci.*, vol. 15, 2021, Art. no. 573194.

[12] S. Maiti, S. K. Sharma, and R. S. Bapi, “Enhancing healthcare with EOG: A novel approach to sleep stage classification,” in *Proc. 2024 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2024, pp. 2305–2309.

[13] Z. Liu, S. Luo, Y. Lu, Y. Zhang, L. Jiang, and H. Xiao, “Extracting multi-scale and salient features by MSE based U-structure and CBAM for sleep staging,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 31–38, 2023.

[14] Y. Zhou et al., “Simplifying multimodal with single EOG modality for automatic sleep staging,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 1668–1678, 2024.

[15] M. M. Rahman, M. I. H. Bhuiyan, and A. R. Hassan, “Sleep stage classification using single-channel EoG,” *Comput. Biol. Med.*, vol. 102, pp. 211–220, 2018.

[16] R. Yan, F. Li, D. D. Zhou, T. Ristaniemi, and F. Cong, “Automatic sleep scoring: A deep learning architecture for multi-modality time series,” *J. Neurosci. Methods*, vol. 348, 2021, Art. no. 108971.

[17] Y. Lin, M. Wang, F. Hu, X. Cheng, and J. Xu, “Multimodal polysomnography based automatic sleep stage classification via multiview fusion network,” *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 2504112.

[18] A. Habib, M. A. Motin, T. Penzel, M. Palaniswami, J. Yearwood, and C. Karmakar, “Performance of a convolutional neural network derived from PPG signal in classifying sleep stages,” *IEEE Trans. Biomed. Eng.*, vol. 70, no. 6, pp. 1717–1728, Jun. 2023.

[19] Y. Li, J. Chen, W. Ma, G. Zhao, and X. Fan, “MVF-SleepNet: Multi-view fusion network for sleep stage classification,” *IEEE J. Biomed. Health Inform.*, vol. 28, no. 5, pp. 2485–2495, May 2024.

[20] J. Lu, C. Yan, J. Li, and C. Liu, “Sleep staging based on single-channel EEG and EOG with tiny U-Net,” *Comput. Biol. Med.*, vol. 163, 2023, Art. no. 107127.

[21] Y. Dai et al., “MultiChannelSleepNet: A transformer-based model for automatic sleep stage classification with PSG,” *IEEE J. Biomed. Health Inform.*, vol. 27, no. 9, pp. 4204–4215, Sep. 2023.

[22] J. Wang, S. Zhao, H. Jiang, S. Li, T. Li, and G. Pan, “Generalizable sleep staging via multi-level domain alignment,” in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 265–273.

[23] C. Bellos, K. Stefanou, A. Tzallas, G. Stergios, and M. Tsipouras, “Methods and approaches for user engagement and user experience analysis based on electroencephalography recordings: A systematic review,” *Electronics*, vol. 14, no. 2, 2025, Art. no. 251.

[24] J. Wang, X. Wang, X. Ning, Y. Lin, H. Phan, and Z. Jia, “Subject-adaptation salient wave detection network for multimodal sleep stage classification,” *IEEE J. Biomed. Health Inform.*, vol. 29, no. 3, pp. 2172–2184, Mar. 2025.

[25] S. Ma, Y. Zhang, Y. Chen, T. Xie, S. Song, and Z. Jia, “Exploring structure incentive domain adversarial learning for generalizable sleep stage classification,” *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 1, pp. 1–30, 2024.

[26] P. Oza, V. A. Sindagi, V. V. Sharmini, and V. M. Patel, “Unsupervised domain adaptation of object detectors: A survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 4018–4040, Jun. 2024.

[27] E. Eldele et al., “ADAST: Attentive cross-domain EEG-based sleep staging framework with iterative self-training,” *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 1, pp. 210–221, Feb. 2023.

[28] P. Franken and D.-J. Dijk, “Sleep and circadian rhythmicity as entangled processes serving homeostasis,” *Nature Rev. Neurosci.*, vol. 25, no. 1, pp. 43–59, 2024.

[29] Y. Vanrobaeys et al., “Spatial transcriptomics reveals unique gene expression changes in different brain regions after sleep deprivation,” *Nature Commun.*, vol. 14, no. 1, 2023, Art. no. 7095.

[30] Q. Shen et al., “LGSleepNet: An automatic sleep staging model based on local and global representation learning,” *IEEE Trans. Instrum. Meas.*, vol. 72, 2023, Art. no. 2521814.

[31] S. Supriya, S. Siuly, H. Wang, and Y. Zhang, “EEG sleep stages analysis and classification based on weighed complex network features,” *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 2, pp. 236–246, Apr. 2021.

[32] Z. Chen et al., “A novel ensemble deep learning approach for sleep-wake detection using heart rate variability and acceleration,” *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 5, pp. 803–812, Oct. 2021.

[33] B. Kompa, J. Snoek, and A. L. Beam, “Second opinion needed: Communicating uncertainty in medical machine learning,” *NPJ Digit. Med.*, vol. 4, no. 1, 2021, Art. no. 4.

[34] K. Zheng, C. Lan, W. Zeng, Z. Zhang, and Z.-J. Zha, “Exploiting sample uncertainty for domain adaptive person re-identification,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3538–3546.

[35] Y. Fang, P.-T. Yap, W. Lin, H. Zhu, and M. Liu, “Source-free unsupervised domain adaptation: A survey,” *Neural Netw.*, vol. 174, 2024, Art. no. 106230.

[36] E. Hatefi, H. Karshenas, and P. Adibi, “Distribution shift alignment in visual domain adaptation,” *Expert Syst. Appl.*, vol. 235, 2024, Art. no. 121210.

[37] D. Goswami, Y. Liu, B. Twardowski, and J. van de Weijer, “FeCAM: Exploiting the heterogeneity of class distributions in exemplar-free continual learning,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2024, pp. 6582–6595.

[38] Y. Fan, J. Liu, J. Tang, P. Liu, Y. Lin, and Y. Du, “Learning correlation information for multi-label feature selection,” *Pattern Recognit.*, vol. 145, 2024, Art. no. 109899.

[39] X. Liu, Z. Liu, J. Li, and X. Zhang, “Semi-supervised contrastive learning for time series classification in healthcare,” *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 9, no. 1, pp. 318–331, Feb. 2025.

[40] C.-L. Li, W.-C. Chang, Y. Cheng, Y. Yang, and B. Póczos, “MMD GAN: Towards deeper understanding of moment matching network,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 2200–2210.

[41] Y. Ganin et al., “Domain-adversarial training of neural networks,” *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.

[42] M. Long, Z. Cao, J. Wang, and M. I. Jordan, “Conditional adversarial domain adaptation,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1647–1657.

[43] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. Hauptmann, “Self-paced curriculum learning,” in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2694–2700.

[44] T. Takase, R. Karakida, and H. Asoh, “Self-paced data augmentation for training neural networks,” *Neurocomputing*, vol. 442, pp. 296–306, 2021.

[45] Q. Zhu, C. Zheng, Z. Zhang, W. Shao, and D. Zhang, “Dynamic confidence-aware multi-modal emotion recognition,” *IEEE Trans. Affect. Comput.*, vol. 15, no. 3, pp. 1358–1370, Jul./Sep. 2024.

[46] Y. Dou, T. Xing, X. Zhao, X. Chen, J. Zhou, and S. Peng, “Spatio-temporal representation learning with selective state space models for EEG-based depression detection,” *Biomed. Signal Process. Control*, vol. 112, 2026, Art. no. 108707.

[47] X. Xu, F. Cong, Y. Chen, and J. Chen, “Sleep stage classification with multi-modal fusion and denoising diffusion model,” *IEEE J. Biomed. Health Inform.*, vol. 29, no. 12, pp. 9233–9244, Dec. 2025.

[48] Y. Liu, L. Qin, X. Chen, R. L. B. Jeannes, J. L. Coatrieux, and H. Shu, “Advancing cross-subject domain generalization in brain-computer interfaces with multi-adversarial strategies,” *IEEE Trans. Instrum. Meas.*, vol. 74, 2025, Art. no. 4010012.

[49] T. Hori et al., “Proposed supplements and amendments to ‘A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects’, The Rechtschaffen & Kales (1968) standard,” *Psychiatry Clin. Neurosciences*, vol. 55, no. 3, pp. 305–310, 2001.

[50] L. Fiorillo, P. Favaro, and F. D. Faraci, “DeepSleepNet-Lite: A simplified automatic sleep stage scoring model with uncertainty estimates,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2076–2085, 2021.

[51] S. Zhao et al., “Multi-source distilling domain adaptation,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 12975–12983.

[52] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, “Adversarial discriminative domain adaptation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7167–7176.

[53] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[54] S. Zhang, D. Han, J. Dezert, and Y. Yang, “Weighted self-paced learning with belief functions,” *Expert Syst. Appl.*, vol. 255, 2024, Art. no. 124535.

[55] H. Wu, Y. Liang, X.-Z. Gao, J.-N. Heng, and Z. Chen, “Sleep-induced network with reducing information loss for short-term load forecasting,” *IEEE Trans. Power Syst.*, vol. 40, no. 2, pp. 1492–1503, Mar. 2025.

[56] X. Zheng et al., “Enhanced self-attention mechanism for long and short term sequential recommendation models,” *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 8, no. 3, pp. 2457–2466, Jun. 2024.

Kun Wang is currently working toward the Ph.D. degree in software engineering with the College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing, China. Her research interests include machine learning, brain-computer interface, and EEG analysis.

Qi Zhu (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 2007, 2010, and 2014, respectively. He is currently a Professor with the College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing, China. He has authored or coauthored more than 100 scientific articles in refereed international journal and conference proceedings. His research interests include pattern recognition, feature extraction, and medical image analysis.

Junyong Zhao received the M.S. degree from Shandong Normal University, Jinan, China, in 2020, and the Ph.D. degree in computer science and technology from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2025. His research interests include deep learning, EEG analysis, and medical image segmentation.

Chuhang Zheng received the M.S. degree in computer science from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2025. He is currently working toward the Ph.D. degree with Tianjin University. His research interests include computational imaging and brain decoding.

Wei Shao received the B.S. and M.S. degrees from the Nanjing University of Technology, Jiangsu, China, in 2009 and 2012, respectively, and the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2018. He is currently an Associate Professor with the College of Artificial Intelligence, Nanjing University of Aeronautics and Astronautics. His research interests include machine learning and bioinformatics.

Daoqiang Zhang (Senior Member, IEEE) received the B.S. and Ph.D. degrees in computer science from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 1999 and 2004, respectively. He joined the Department of Computer Science and Engineering, NUAA, as a Lecturer in 2004, where he is currently a Professor. His research interests include machine learning, pattern recognition, data mining, and medical image analysis. In these areas, he has authored or coauthored more than 200 scientific articles in refereed international journals such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON MEDICAL IMAGING, IEEE TRANSACTIONS ON IMAGE PROCESSING, *NeuroImage*, *Medical Image Analysis*, *Pattern Recognition*, *Artificial Intelligence in Medicine*, IEEE Transactions on Neural Networks, and conference proceedings such as NIPS, IJCAI, AAAI, SDM, and ICDM. He is a member of the Machine Learning Society of the Chinese Association of Artificial Intelligence, the Artificial Intelligence and Pattern Recognition Society of the China Computer Federation, and International Association for Pattern Recognition Fellow.