Simplifying Multimodal With Single EOG Modality for Automatic Sleep Staging

Yangxuan Zhou, Sha Zhao, *Member, IEEE*, Jiquan Wang, Haiteng Jiang, Zhenghe Yu, Shijian Li, Tao Li, and Gang Pan, *Senior Member, IEEE*

***Abstract*— Polysomnography (PSG) recordings have been widely used for sleep staging in clinics, containing multiple modality signals (i.e., EEG and EOG). Recently, many studies have combined EEG and EOG modalities for sleep staging, since they are the most and the second most powerful modality for sleep staging among PSG recordings, respectively. However, EEG is complex to collect and sensitive to environment noise or other body activities, imbedding its use in clinical practice. Comparatively, EOG is much more easily to be obtained. In order to make full use of the powerful ability of EEG and the easy collection of EOG, we propose a novel framework to simplify multimodal sleep staging with a single EOG modality. It still performs well with only EOG modality in the absence of the EEG. Specifically, we first model the correlation between EEG and EOG, and then based on the correlation we generate multimodal features with time and frequency guided generators by adopting the idea of generative adversarial learning. We collected a real-world sleep dataset containing 67 recordings and used other four public datasets for evaluation. Compared with other existing sleep staging methods, our framework performs the best when solely using the EOG modality. Moreover, under our framework, EOG provides a comparable performance to EEG.**

***Index Terms*— Multi modalities, PSG recordings, sleep staging.**

I. Introduction

SLEEP quality is vital for everyone's wellbeing, since an individual spends almost one-third of her life either sleeping or trying to do so [1], [2]. Sleep staging is important for both monitoring sleep quality and diagnosing sleep disorders [3], which categorizes sleep into different stages, such as Wake, REM (Rapid Eye Movement) and non-REM sleep. In clinical practice, polysomnography (PSG) has been widely used for sleep staging, recording various physiological signals of the human body, such as EEG (electroencephalogram), EOG (electrooculogram), EMG (electromyogram), and ECG (electrocardiogram). PSG recordings are stored as consecutive epochs, each of which is 30-seconds. Traditionally, experts categorize each epoch into five different stages, namely, W, N1, N2, N3, and REM, following the sleep staging standards established by the American Academy of Sleep Medicine (AASM) [4]. It usually takes several hours for one expert to scoring the overnight PSG recordings of one person. Obviously, such manual process is time-consuming. Meanwhile, the sleep staging results are relatively subjective, since the manual staging heavily depends on experts' experiences.

With the rapid advancement of deep learning techniques, there is a growing interest in the development of automatic sleep staging methods using PSG recordings [5], [6], [7], [8]. Given that there are various types of signals in PSG, many studies have tried different types of single modality for sleep staging, and relatively popularly used modalities are EEG, EOG, EMG, and ECG. For instance, some studies [10], [11], and [12] have solely used EEG for sleep staging and achieved good performance across multiple publicly available datasets. Instead of employing EEG, Eognet [13] proved that using single EOG modality also effectively discriminates different sleep stages, but the predictive ability of EOG is not so strong as that of EEG. As for EMG, Andreotti et al. [14] demonstrated that only using EMG modality is not feasible for

Manuscript received 24 November 2023; revised 1 February 2024 and 8 March 2024; accepted 5 April 2024. Date of publication 18 April 2024; date of current version 25 April 2024. This work was supported in part by the Science Technology Innovation (STI) 2030 Major Projects under Grant 2021ZD0200400; in part by the Key Program of the Natural Science Foundation of Zhejiang Province, China, under Grant Z24F020009; and in part by the Natural Science Foundation of China under Grant 61925603. *(Corresponding author: Sha Zhao.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board of Zhejiang University.

Yangxuan Zhou, Sha Zhao, Jiquan Wang, and Shijian Li are with the College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310013, China, and also with the State Key Laboratory of Brain–Machine Intelligence, Zhejiang University, Hangzhou, Zhejiang 311121, China (e-mail: zyangxuan@zju.edu.cn; szhao@zju.edu.cn; wangjiquan@zju.edu.cn; shijianli@zju.edu.cn).

Haiteng Jiang and Tao Li are with the Department of Neurobiology, Affiliated Mental Health Center, Hangzhou Seventh People's Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310063, China, also with the MOE Frontier Science Center for Brain Science and Brain–Machine Integration, Hangzhou, Zhejiang 310063, China, and also with the State Key Laboratory of Brain–Machine Intelligence, Zhejiang University, Hangzhou, Zhejiang 311121, China (e-mail: h.jiang@zju.edu.cn; litaozjusc@zju.edu.cn).

Zhenghe Yu is with the Department of Sleep Medicine, Affiliated Mental Health Center, Hangzhou Seventh People's Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310063, China (e-mail: yuzhcoo@sina.com).

Gang Pan is with the College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310013, China, also with the State Key Laboratory of Brain–Machine Intelligence, Zhejiang University, Hangzhou, Zhejiang 311121, China, and also with the MOE Frontier Science Center for Brain Science and Brain-Machine Integration, Zhejiang University, Hangzhou, Zhejiang 310063, China (e-mail: gpan@zju.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2024.3389077

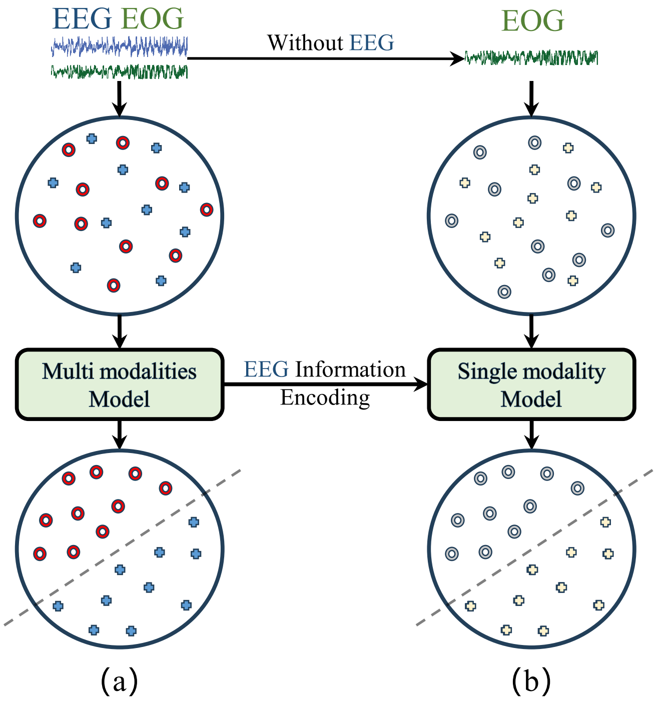


Fig. 1. (a) Sleep staging using multimodalities and (b) Simplifying multimodal with single EOG modality for sleep staging.

sleep staging, and its predictive ability is much weaker than that of EEG and EOG. Similarly, solely using other types of signals (i.e., ECG [15]) in PSG recordings cannot work for the sleep staging task. Taken together, the EEG modality stands out as the most powerful for sleep staging among all PSG signals, and EOG is the second most powerful.

However, the collection of EEG signals is quite complex and expensive. Typically, subjects are required to do several inevitable preparations, such as preparing head skin, wearing a cap with dozens of electrodes, and injecting conductive gel. Moreover, EEG signals are very sensitive and subtle to environment noise or disturbance of other body activities (i.e., eye movement, leg movement). It is hard to guarantee the high quality of EEG signals, especially for one person lying for so long time of approximately 8 hours during sleep. These limitations severely restrict the usability of EEG in real-world practical sleep-related applications. Comparatively, EOG, another important type of cues for sleep staging, is relatively easily to collect by simply placing sensors near the eyes during sleep. Meanwhile, the EOG signals are not so sensitive to environment and other body activities. Due to the powerful ability of EEG in sleep staging and the easy collection nature of EOG, it is necessary to figure out how to solely use EOG modality for sleep staging but take advantage of EEG information.

In order to make full use of the advantages of EEG and EOG for sleep staging, some studies have employed both EEG and EOG modalities to address sleep staging [16], [17], [18], [19]. As expected, the combination of EEG and EOG improves the sleep staging performance, compared with the performance when solely using EEG or EOG [14]. The success of multimodal studies indicates the potential correlation between EEG and EOG modality. It motivates us to first learn the multimodal representation of EEG and EOG and capture their correlation. Then, based on such correlation, we try to generate multimodal representation from the single EOG modality for sleep staging. As shown in Fig. 1, we can simplify multimodal with single EOG modality to classify sleep stages when EEG is not available, making full use of the easy collection nature of EOG and powerful ability of EEG and avoiding the complex collection of EEG.

However, simplifying multimodal with EOG for sleep staging is a nontrivial task, and there are some difficulties. The first one is how to generate multimodal representations containing both EEG and EOG information when EEG modality is not available. Undoubtedly, the generated multimodal representations are the key factor for simplifying multimodal, since it decides the performance when we only use EOG for sleep staging. We train generators to generate multimodal representations with single EOG based on the correlation between EEG and EOG, by adopting the idea of generative adversarial learning. The second difficulty is how to align the characteristics in time and frequency of EOG with that of EEG. Many studies have proved the significance of temporal and spectral features of EEG for sleep staging [20]. Thus, we conditionally guide the generators from the perspectives of time and frequency, respectively, to generate multimodal representations for sleep staging when EEG is not available.

In this paper, in order to overcome the limitations of EEG in clinical practice, we propose a novel framework to simplify multimodal sleep staging using a single EOG modality. It makes full use of the powerful predictive ability of EEG and easy collection nature of EOG. We collected one real-world dataset consisting of 67 subjects and used other four public datasets to evaluate our framework. Our contributions can be summarized as follows:

- We propose a novel framework to simplify multimodal with single EOG modality for sleep staging, which can perform well with only inputting EOG instead of inputting EEG and EOG together.
- We first model the correlation between EEG and EOG. Then, we generate multimodal representations based on the correlation by adopting the idea of generative adversarial learning. In particular, we consider the temporal and spectral features into the generated multimodal representations.
- Our framework is evaluated on our collected dataset and four public datasets, and the results demonstrate its effectiveness. **Compared with existing methods, when only using EOG as input, our framework performs the best. Moreover, by our framework, the EOG provides comparable performance to EEG**.

II. RELATED WORK

A. Sleep Staging With Single Modality

In previous studies, considering the multiple types of signals present in PSG, many studies have employed different modalities for BCI tasks [21], [22], [23], such as EEG, EOG and ECG. For instance, Supratak et al. [9] proposed DeepSleepNet which is a CNN-BiLSTM based network using EEG, aiming to extract invariant features across different shifts and learn the transition rules among different sleep stages. U-time [10], [24] is a fully CNN network based on the

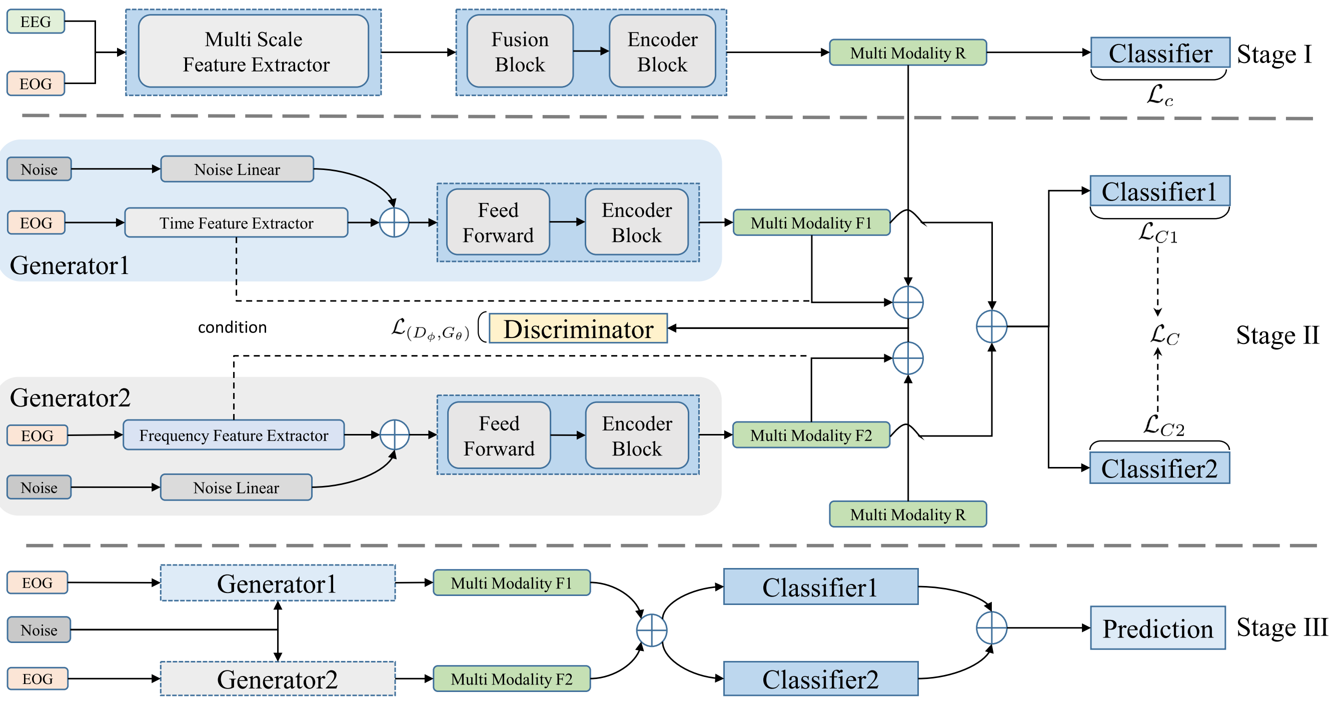


Fig. 2. Overview of the multimodal simplification framework. The entire process can be divided into three phases. The multimodal correlation will be modeled in Stage I. In Stage II, synthetic multimodal representations will be generated, and the dashed lines represent the guiding conditions. In Stage III, only the EOG data will be input for inference. Here, $\oplus$ denotes the concatenation operation.

U-net architecture that can excellently model sleep-related features from EEG. RecSleepNet [11] improved the modeling accuracy of EEG by reconstructing the intermediate-level feature representations. Eognet [13] introduced a two-step training strategy, focusing on modeling the EOG feature representations for sleep staging. Sun et al. [15] employed CNN-RNN based network to extract the ECG features for distinguishing different sleep stages, but the performance is much worse compared to that of models based on EEG or EOG signals. Taking together, the EEG modality stands out as the most powerful for sleep staging among all the PSG signals and the EOG modality ranks the second.

B. Sleep Staging With Multiple Modalities

Considering the complementarity between EEG and other modalities, some studies constructed sleep staging models based on multimodal feature representations, achieving better performance than single-modal based approaches. Based on EEG and EOG, Jia et al. [17] proposed SalientSleepNet, which includes a Multimodal Attention Module designed to extract multimodal features for specific sleep stage. Compared with SalientSleepNet, MMASleepNet [25] and XSleepNet [16] additionally introduced EMG modality, learning sleep information from three different modalities. These multimodal based models demonstrate better performance than single-modal based methods on public sleep datasets. Whether employing single modal or multimodal based approaches, the EEG modality proves the irreplaceability for sleep staging. Some studies [26] and [27] have explored cross-modal knowledge distillation. Zhang et al. [28] proposed a visual-to-EEG cross-modal knowledge distillation for emotion recognition. Zhang et al. [29] proposed a knowledge distillation algorithm based on Multi-Channel Multi-Domain to enhance single EEG channel based sleep staging. Liang et al. [30] proposed SleepKD, a Teacher Assistant-Based model using knowledge distillation, which can capture multi-level sleep features using a single EEG channel. These existing knowledge distillation methods can effectively improve the performance of single EEG-based task. However, the complexity and the high-cost of EEG modality acquisition limit its practical application in real-world scenarios. Based on GANs (Generative adversarial nets) [31], Yan et al. [32] tackled the challenges of one-to-many cross-modal transfer in the domain of emotion recognition. Inspired by the previous work, we propose a novel multimodal representation generation based framework to simplify the multimodal sleep staging with single EOG modality, alleviating the limitations of EEG utilization.

III. METHODS

A. Problem Formulation

Here, we simplify multimodal with single EOG modality, and introduce a novel task of generating multimodal representations for sleep staging. Formally, given a sleep sequence $S{=}(x_1, x_2, x_3, \ldots, x_L)$ of length L, where x_i denotes the i-th epoch in the sequence S. Our goal is to compute the sequence of outputs $Y{=}(y_1, y_2, y_3, \ldots, y_L)$ that maximizes the conditional probability $p(x_1, x_2, x_3, \ldots x_L|y_1, y_2, y_3, \ldots, y_L)$. Here $y_i \in \{0, 1\}^N$ is corresponding one-hot encoding of real sleep stage of x_i and $N = 5$ denotes the number of sleep stage.

B. Overview

Figure. 2 illustrates the architecture of the proposed framework, which is composed of three stages. In the first stage, we model the correlation between EEG and EOG to obtain real multimodal features. In the second stage, based on the modeled correlation, we generate synthetic multimodal features by training dual generators using a generative adversarial learning method. In the third stage, with the trained generators, we input single EOG modality for sleep staging. **Stage I**: Inputting the EEG and EOG modality, we first capture the correlation between EEG and EOG by pretraining a model, and obtain a real multimodal representations containing EEG and EOG information. **Stage II**: We generate multimodal representations from Gaussian noises based on the learned correlation in the Stage I, and employ adversarial learning with real multimodal features obtained in the Stage I to enhance the reliability of generated multimodal representations. The temporal and spectral features of EOG are considered into the generated multimodal representations. **Stage III**: In test stage, we only use the EOG modality as input to generate the reliable multimodal features, and classify sleep stages. In the subsequent subsections, we will introduce more details for each stage.

C. *Stage I: Modeling Multimodal Correlation*

In this stage, we first model the correlation between EEG and EOG by pretraining a sleep staging model, which is fundamental for simplifying multimodal sleep staging. After pretraining, we can obtain real multimodal features, and the quality of the obtained multimodal features is crucial for generating synthetic multimodal representations in Stage II. Here, the pretrained model consists of three components: the Feature Extractor, Fusion Block and Temporal Encoder. Taking EEG and EOG sequences as inputs, we employ multi-scale convolutional networks to extract invariant features across different shifts from each modality. Given that small-scale convolutions are good at capturing temporal information, and large-scale convolutions are usually used to capture frequency information [9], we perform the small-scale and large-scale convolutions separately for each channel of EEG and EOG modalities. For the spatial features of different channels, we apply the Style-based Recalibration Module (SRM) [33] to weight each channel, and focus on the more important channels. The equation is as follows:

$$X = F_{srm}(X, \theta) \cdot X \tag{1}$$

We utilize several fully connected layers to fuse the EEG and EOG channels. Subsequently, a feedforward neural network is employed to integrate the EEG and EOG modalities and extract multimodal features. After fusion, the multimodal features will be input into the Transformer Encoder [34] to obtain real multimodal representations containing EEG and EOG information. We apply the feed-forward network as the classifier and train the model by minimizing the cross-entropy loss:

$$\mathcal{L}_c = -\frac{1}{L}\sum_{i}^{L} y_i \log \hat{y}_i \tag{2}$$

where y_i is the real sleep stage label for input epoch x_i.

D. *Stage II: Generating Multimodal Representations*

Based on the real multimodal representation obtained in Stage I, our objective is to generate synthetic multimodal features using single EOG modality in this stage. Similarly to other generation methods, we generate multimodal features from gaussian noise. Considering the sequential nature of sleep signals, it remains challenging to guarantee the reliability of the generation without guidance. Hence, we utilize the EOG modality as a guided condition to generate synthetic multimodal features from gaussian noise. Given the real multimodal features obtained in stage I and the generated synthetic multimodal features, we first map these two multimodal representations into a high-dimensional feature space and then employ adversarial training to align them. Specifically, in our framework, a discriminator D is trained to classify the real and synthetic multimodal features, while two generators $G1$ and $G2$ try to generate indistinguishable representations between the real and synthetic features. By doing so, the generated multimodal features can be trained to approximate the distribution of the real ones, promoting better performance when using single EOG modality. Adopting a standard GAN loss, the discriminator can be optimized using a cross-entropy loss. The objective of this operation $\mathcal{L}_D$ can be defined as:

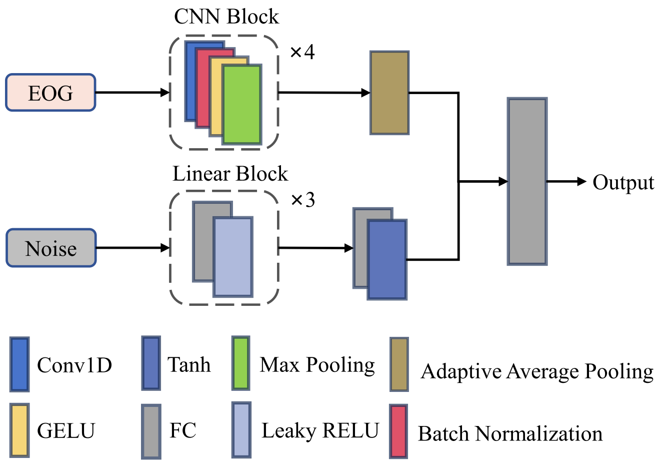


Fig. 3. The backbone of Time/Frequency dual generators.

$$\begin{aligned}\min_D \mathcal{L}_D = &-\mathbb{E}_{(f_r,X_{eog})\sim P_r}[log D(f_r, X_{eog})]\\ &-\mathbb{E}_{(f_g,X_{eog})\sim P_g}[log(1-D(f_g, X_{eog}))]\end{aligned} \tag{3}$$

where f_r and f_g denote the real and generated multimodal features, respectively. X_{eog} denotes the EOG conditional constraint. $\mathcal{L}_D$ is used to optimize the discriminator separately so that it discriminates the real and synthetic features. Here we adopt the inverted labels to address the gradient vanishing [35].

Simultaneously, the generators are trained to confuse the discriminator by generating synthetic features that closely resemble the real ones. The objective function can be described as:

$$\begin{aligned}\min_G \mathcal{L}_{adv} = &-\mathbb{E}_{(X_{eeg},X_{eog})\sim P_r}[log(1-D(f_r, X_{eog}))]\\ &-\mathbb{E}_{(X_n,X_{eog})\sim P_g}[log D(G(X_n, X_{eog}), X_{eog})]\end{aligned} \tag{4}$$

where X_{eog} and X_n denote the EOG constraint and the gaussian noise. f_r denotes the real multimodal features. Notably, we have applied conditional constraints to both the generator and the discriminator. Due to the freezing of pretrained network parameters, the first step of the Eq. 4 does not participate in network optimization. Therefore, the equation can be reformulated as:

$$\min_{G} \mathcal{L}_{adv} = -\mathbb{E}_{(X_n, X_{eog})\sim P_g}[log D(G(X_n, X_{eog}), X_{eog})] \quad (5)$$

The above is about the process of how to generate synthetic multimodal features. We will introduce our generators, discriminators and classifiers in detail.

1) Time-Frequency Dual Generators: Considering the importance of time and frequency characteristics of sleep signals for sleep staging [36], [37], we employ time-frequency dual generators to generate different multimodal features: in the time domain generator $generator1$, we apply the time-domain EOG as the input and we perform Fourier transform on the EOG signal and take its modulus as the input in the frequency domain generator $generator2$. Then we adopt randomly generated Gaussian noise as one of the inputs for generators. It is worth noting that the variance of the standard normal distribution significantly differs from that of the original sleep signals. So we set its mean and variance to match those of the original EOG signals, making the distribution of Gaussian noise closer to that of the EOG signal. This step facilitates the subsequent adversarial generation training. The noise will be input into multiple fully connected layers to improve the fitting ability. Then, we input time-frequency domain EOG signals into different feature extractors, each of which contains multiple CNN layers and shares the same structure, respectively. After feature extraction, the time-frequency domain EOG signals will be combined with the noise in order to add conditional constraints. Notably, the Time Feature Extractor and the Frequency Feature Extractor share the same structure. Further details are illustrated in Fig. 3. Subsequently, we apply transformer encoder to learn temporal information within a sequence of synthetic features. The generated multimodal features F_1 and F_2 will be input into the discriminator to optimize the generators by minimizing Eq. 4. The total adversarial loss can be described as follows:

$$\mathcal{L}_G = \lambda_1 \mathcal{L}_{adv1} + \lambda_2 \mathcal{L}_{adv2} \quad (6)$$

where $\mathcal{L}_{adv1}$ and $\mathcal{L}_{adv2}$ denote the adversarial loss of generated time domain multimodal features F_1 and frequency domain multimodal F_2, respectively.

2) Discriminator: The discriminator is a binary classifier used to distinguish the real multimodal features f_r or synthetic multimodal features f_g. We first concatenate these two multimodal features f_r and f_g. Then in the discriminator, the input vector will first be concatenated with the guiding conditions: the single EOG modality. Then, the concatenated vector undergoes a linear fusion layer, mapping it to a dimension of 512. We have constructed multiple layers of linear units, enabling the discriminator to output a binary classification probability indicating whether the multi-modal features are from real or generated distributions. The discriminator can be optimized by minimizing Eq. 3. The total loss of discriminator can be described as follows:

$$\mathcal{L}_D = \lambda_1 \mathcal{L}_{dis1} + \lambda_2 \mathcal{L}_{dis2} \quad (7)$$

where $\mathcal{L}_{dis1}$ and $\mathcal{L}_{dis2}$ denote the discriminator loss of generated multimodal features F_1 and F_2, respectively.

Algorithm 1 Multimodal Simplification Algorithm

```
Input: $X_{EEG}$, $X_{EOG}$
Output: Evaluation indicators of test data
The Stage I:
Initialize parameters $\theta$ in the pretrained model.
for $i = 1$ to $n$ do
    Optimize $\theta$ by minimizing Eq. 2
end
return multimodal feature $R$.
The Stage II:
Initialize generator $G1$ and $G2$, classifier $C1$ and $C2$,
  discriminator $D$.
for $i = 1$ to $n$ do
    Generate synthetic multimodal features $F_1$ and $F_2$.
    Concentrate $F_1$ and $F_2$ with $R$, respectively.
    Optimize $D$ by minimizing Eq. 7.
    Optimize $G1$, $G2$, $C1$ and $C2$ by minimizing Eq.
      10.
end
return trained $G1$, $G2$, $C1$ and $C2$.
The Stage III:
Using random noise and $X_{EOG}$ to generate
  multimodal features through trained $G1$ and $G2$.
Use the trained classifier $C1$ and $C2$ for sleep staging.
return evaluation indicators.
```

3) Dual Classifiers: After generation, we do the addition operation for the multimodal features F_1 and F_2 and control the operation by a hyperparameter α as follows:

$$F_{out} = \alpha F_1 + (1-\alpha) F_2 \quad (8)$$

Some existing studies [38], [39] have demonstrated that dual classifiers can assist the model in reducing variance during the training process and decreasing the probability of low-confidence predictions by utilizing the average prediction vector. Given the real multimodal feature R and the synthetic multimodal feature F_{out} generated from dual generators, the discrepancies are aligned through discriminative cross-modality alignment. Then, the dual classifiers, which share the same architecture, will further enhance the sleep staging decision boundaries and improve the robustness of predictions. We choose two sets of initialization methods: He [40] and Glorot and Bengio [41] to ensure the diversity of predictions, making sure that the dual classifier does not converge to become the same one throughout training. We use cross-entropy loss to optimize dual classifiers as follows:

$$\mathcal{L}_C = \lambda_1 \mathcal{L}_{C1} + \lambda_2 \mathcal{L}_{C2} \quad (9)$$

To sum up, we integrate the adversarial loss with the classification loss in one objective loss function as follows:

$$\mathcal{L}_{overall} = \gamma \mathcal{L}_C + (1-\gamma) \mathcal{L}_G \quad (10)$$

Notably, the loss of discriminator $\mathcal{L}_D$ will be updated and optimized separately. In our study, we are more concerned with the parameters updated on the classifiers. We set γ to 0.7. For $generator1$ and $generator2$, we attach equal importance to them, setting $\lambda_1 = \lambda_2 = 0.5$. But for the multimodal features F_1 and F_2 with different generation conditions, we pay more

TABLE I
AN OVERVIEW OF THE DATASETS

Dataset	Size	K-CV	Sample Rate	Scoring	Channels
SSND	67	10	100	AASM	F3-M2, C3-M2, O1-M2, F4-M1, C4-M1, O2-M1, E1-M2, E2-M1
ISRUC	98	10	100	AASM	F3-M2, C3-M2, O1-M2, F4-M1, C4-M1, O2-M1, E1-M2, E2-M1
SleepEDF-153	153	10	100	R&K	Fpz-Cz, Pz-Oz, EOG horizontal
HMC	147	10	100	AASM	F4-M1, C4-M1, O2-M1, C3-M2, E1-M2, E2-M2
MASS SS2	20	5	150	AASM	F3-M2, C3-M2, O1-M2, F4-M1, C4-M1, O2-M1, EOG LH, RH, UV, LV

K-CV denotes the K-fold number. On the MASS SS2 dataset, EOG LH, RH, UV, LV denote EOG Left Horiz, EOG Right Horiz, EOG Upper Vertic, EOG Lower Vertic, respectively.

attention to the features guided by time-domain information after we conduct the paramenter experiment of α. Finally, we set $\alpha = 0.7$.

E. *Stage III*: Sleep Staging With Single EOG Modality

During the inference phase, we solely use single EOG modality as input. The trained $generator1$ and $generator2$ can generate corresponding synthetic multimodal features F_1 and F_2. After addition, F_{out} will be passed into trained dual classifiers. We calculate the average probability from the two classifiers and get the final prediction $\hat{y}$ as follows:

$$P = \frac{1}{2}[C_1(F_{out}) + C_2(F_{out})] \tag{11}$$

$$\hat{y} = argmax(P) \tag{12}$$

The algorithm of our multimodal simplification framework is illustrated in Algorithm 1.

IV. EXPERIMENTS

A. Datasets

In order to evaluate our framework, we collected a sleep dataset, namely, **SSND**. For the fairness of the evaluation, we further conducted evaluation experiments on four publicly available sleep datasets, including **ISRUC**, **SleepEDF-153**, **HMC**, and **MASS SS2**. The PSG recordings in each dataset differ from others in many aspects, such as EEG channels, collection instruments, and populations. A brief summary of all the datasets is given in Tab. I.

SSND is collected by ourselves and contains 67 PSG recordings from 17 healthy subjects and 50 subjects with narcolepsy. The dataset consists of 42 females and 35 males, and the PSG recordings were collected at the Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, Zhejiang University School of Medicine. The research was conducted at Zhejiang University with Institutional Review Board approval, and written consent was acquired from all the subjects or their caregivers. For each subject, we collected PSG recordings for an entire night, starting at approximately 21:00 and ending at 5:00 of the following morning, totaling approximately 8 hours. All signals were stored in the standard EDF+ data format with a.edf extension. The recordings were divided into 30-second epochs, with each epoch manually labeled as a sleep stage by sleep experts or technicians according to AASM [4] guidelines. In our SSND dataset, there are a total of 84,546 epochs, consisting of 56,895 sleep epochs from 50 patients aged 11 to 49, and 27,651 sleep epochs from 17 healthy people aged 22 to 32. All the 67 PSG sleep recordings were used for evaluation.

ISRUC is a public dataset [42] composed of 3 sub-groups. We choose sub-group1 which includes overnight PSG recordings of 100 adults. We excluded subject 8 and 40 due to some missing channels.

SleepEDF-153 is a public PhysioNet dataset [43] consisting of 78 healthy subjects aged 25-101. Each subject contains two day-night PSG recordings except subjects 13, 36, and 52 whose one recording is missed due to device failure. All the 153 PSG sleep recordings were used for evaluation.

HMC is a public dataset [44] including a total of 154 PSG recordings gathered retrospectively from the sleep center dataset of the Haaglanden Medisch Centrum (The Netherlands). We excluded subject 14,32,33,64,112 and 135 due to some missing channels.

MASS SS2 is a subset of public MASS dataset [45] composed 20 PSG recordings which were segmented into 20s epochs. All the 20 PSG sleep recordings were used for evaluation.

B. Settings

1) Implementation and Metrics: We employ K-fold cross-validation (CV) to assess the performance of our model across the 5 different datasets and the values of K number are listed in Tab. I. In each fold, we apply a subject-independent policy where the samples of the same one subject cannot appear in the test data and the training data simultaneously. We use the Adam optimizer to train the model. The β is set to [0.5,0.99], the weight decay is set to 3e-4 and the learning rate is set to 1e-4. The length of sequence L is set to 20 and the batch size is set to 16. We employ Accuracy (ACC) and Macro-F1 score (MF1) as the evaluation metrics. The model is trained on a single machine equipped with an Intel Core i9 10900K CPU and eight NVIDIA RTX 3080 GPUs using PyTorch.

2) Compared Methods of Sleep Staging: Based on single EOG modality as input, we compared with several deep learning methods and one traditional machine learning method for automatic sleep staging. All the selected methods are designed for single modality. Some methods designed for multiple modalities are not compared here, such as SalientSleepNet [17], MMASleepNet [25], and XSleepNet [16], since we focus on the performance of using single EOG modality. **RF** [46] is a classical ensemble learning method, which has been widely used in classification tasks. Here, we calculate the average power spectral density of different frequency bands to construct features. **DeepSleepNet**

TABLE II
THE OVERVIEW PERFORMANCE OF DIFFERENT MODALITIES ON 5 DATASETS

	ISRUC		SleepEDF		HMC		MASS		SSND	
	ACC	MF1	ACC	MF1	ACC	MF1	ACC	MF1	ACC	MF1
EEG	78.2	74.7	82.4	75.9	80.7	76.7	87.4	80.0	83.7	78.8
EEG+EOG	80.6	77.9	84.1	78.2	81.5	77.9	86.1	78.7	85.4	81.2
EOG	74.6	70.8	79.5	71.6	75.9	71.2	82.2	74.4	82.4	77.1
EOG(Ours)	**77.3(+2.7)**	**74.3(+3.5)**	**80.6(+1.1)**	**73.3(+1.7)**	**77.7(+1.8)**	**74.0(+2.8)**	**85.0(+2.8)**	**76.4(+2.0)**	**83.8(+1.4)**	**79.3(+2.1)**

TABLE III
PERFORMANCE COMPARISON WITH EXISTING METHODS FOR SLEEP STAGING BASED ON EOG MODALITY

Dataset	System	ACC	MF1	Per-class Macro F1				
				W	N1	N2	N3	REM
ISRUC	RF	56.4	48.8	68.9	21.0	62.2	59.3	32.8
	DeepSleepNet	70.9	67.9	80.3	40.9	69.7	79.6	66.8
	TinySleepNet	71.7	67.0	81.7	36.0	69.9	78.4	65.1
	RecSleepNet	72.1	67.8	83.1	38.1	70.3	79.6	68.0
	CNN+Transformer	74.6	70.8	81.2	39.7	74.4	83.7	75.0
	Eognet	76.6	73.7	85.1	**44.5**	74.2	83.3	81.2
	sDREAMER	76.1	72.1	**86.0**	38.1	75.1	82.4	79.4
	Ours	**77.3**	**74.3**	85.9	43.9	**76.3**	**84.0**	**81.3**
SleepEDF	RF	56.6	39.3	63.6	10.9	67.9	21.7	32.5
	DeepSleepNet	75.6	67.4	86.9	34.9	79.2	67.3	68.5
	TinySleepNet	75.7	66.9	87.3	29.8	78.7	68.5	70.0
	RecSleepNet	76.2	67.4	87.7	31.1	79.5	68.1	69.4
	CNN+Transformer	79.5	71.6	89.5	35.6	81.6	69.4	82.2
	Eognet	76.3	69.3	81.6	34.9	81.6	**72.5**	75.7
	sDREAMER	79.9	70.5	89.8	30.4	82.2	66.9	83.6
	Ours	**80.6**	**73.2**	**90.6**	**40.5**	**82.5**	68.6	**83.9**
HMC	RF	53.7	44.8	61.4	13.9	63.2	56.4	28.7
	DeepSleepNet	74.0	70.2	80.1	39.8	76.4	81.1	73.9
	TinySleepNet	74.2	69.6	80.7	36.4	75.6	80.8	74.1
	RecSleepNet	74.8	70.8	80.7	39.9	75.8	81.5	75.4
	CNN+Transformer	75.9	71.2	81.4	35.5	77.5	82.0	79.9
	Eognet	76.1	71.9	80.5	38.3	77.1	81.3	**82.4**
	sDREAMER	76.0	68.8	81.3	23.0	77.9	82.0	79.7
	Ours	**77.7**	**74.0**	**83.2**	**42.7**	**78.6**	**82.8**	82.2
MASS	RF	65.2	49.4	49.6	2.6	78.0	64.0	52.8
	DeepSleepNet	76.1	65.1	74.8	24.0	81.0	76.0	69.4
	TinySleepNet	76.1	63.7	72.6	15.8	80.8	78.2	71.0
	RecSleepNet	76.8	65.0	77.4	16.8	81.8	77.8	71.0
	CNN+Transformer	82.2	74.4	83.0	39.4	85.4	78.6	85.0
	Eognet	79.8	71.8	78.8	**41.2**	83.6	74.6	80.4
	sDREAMER	84.0	75.4	**83.6**	38.0	87.2	82.6	85.4
	Ours	**85.0**	**76.4**	82.0	39.8	**88.8**	**84.2**	**87.8**
SSND	RF	57.1	46.8	61.8	10.5	66.6	63.4	31.6
	DeepSleepNet	81.3	76.0	89.9	44.9	82.4	86.0	76.6
	TinySleepNet	82.0	76.8	89.5	47.4	83.2	**86.9**	76.9
	RecSleepNet	81.6	76.2	89.5	45.5	82.0	86.6	77.3
	CNN+Transformer	82.4	77.1	89.8	45.5	83.2	85.9	81.0
	Eognet	81.2	75.5	88.6	40.9	81.7	85.5	79.2
	sDREAMER	81.3	73.3	88.4	30.3	81.8	86.2	80.1
	Ours	**83.8**	**79.3**	**91.9**	**51.0**	**83.7**	86.4	**83.6**

[9] is a classical CNN-BiLSTM network using multi-scale convolution to capture features. **TinySleepNet** [47] is a more lightweight model based on DeepSleepNet. **RecSleepNet** [11] uses a Convolutional Reconstruction Block to enhance modeling capabilities. **Eognet** [13] is a novel sequential hierarchical neural network that utilizes a single EOG modality for sleep staging. **sDREAMER** [48] is a MoME module-based network that employs a self-distilled strategy to handle single or multiple modalities for sleep staging. In addition, we build a **CNN+Transformer** model for comparison, considering our original single EOG prediction network is based on this architecture. The above models except CNN+Transformer, Eognet and sDREAMER, in the original papers all used single EEG modality as input, we implement all of them according to the core code provided in their papers, and input single EOG modality instead of EEG in our comparison study. Notably, based on our settings, we implement the sequence sDREAMER instead of the epoch sDREAMER for comparision.

C. Result Analysis

1) Comparison With Other Methods Based on EOG: As shown in Tab. III, when solely using single EOG modality as input, **our method achieves the state-of-the-art performance across all the datasets**, with 80.9% on average ACC and 75.4% on average MF1. It proves that our multimodal simplification framework can effectively improve the performance of sleep staging task based on single EOG modality. Among deep learning-based methods using a standard training strategy (employing single EOG data for training and testing), DeepSleepNet performs the worst, specifically with only the EOG modality, about 75.6% on average ACC and 69.3% on average MF1. The performance of TinySleepNet is close to that of DeepSleepNet, with ACC 75.9% and MF1 68.8% on average. RecSleepNet performs better than these two models, with 76.3% on average ACC and 69.4% on average MF1. Eognet performs a little better than those three models, with 78.0% on average ACC and 72.4% on average MF1. CNN+Transformer achieves the best performance compared to the above CNN-LSTM based models, with 78.9% on average ACC and 73.0% on average MF1. It demonstrates that transformer-based models have an advantage over LSTM-based models in modeling temporal sequences. For the sDREAMER which employs a self-distillation strategy to transfer multimodal information to the single-modal-based network, it performs a little worse than our approach, especially on the average ACC with a difference of 1.4% (79.5% vs 80.9%). It even performs slightly better than our model on the easily distinguishable sleep stage of W (wake) on the MASS and ISRUC datasets. However, its average MF1 is worse than ours, with a gap of 3.4% (72.0% vs 75.4%). Moreover, it performs much worse than our model for N1 stage that is difficult to distinguish on several datasets, about 20.7% MF1 lower on SSND, 19.7% MF1 lower on HMC, and 10.1% MF1 lower on SleepEDF.

2) Comparison Between Single Modality and Multiple Modalities: We further investigate how the single EOG modality performs under our framework, especially compared with EEG or the combination of EEG and EOG. Here, we test the performance of single EEG modality, single EOG modality,

TABLE IV
ABLATION EXPERIMENT OVERVIEW

Dataset	ISURC		SleepEDF		HMC		MASS		SSND		Average	
Model	ACC	MF1	ACC	MF1	ACC	MF1	ACC	MF1	ACC	MF1	ACC	MF1
G1-Time	76.4	72.9	80.3	72.6	77.4	73.8	84.6	75.4	83.4	78.5	80.4	74.6
G2-Frequency	76.3	73.4	79.6	71.8	74.2	70.4	79.2	70.5	82.1	77.1	78.3	72.6
Ours	**77.3**	**74.3**	**80.6**	**73.3**	**77.7**	**74.0**	**85.0**	**76.4**	**83.8**	**79.3**	**80.9**	**75.5**

TABLE V
PERFORMANCE COMPARISON WITH EXISTING METHODS FOR SLEEP STAGING BASED ON EEG MODALITY

		ISRUC		SleepEDF	
		ACC	MF1	ACC	MF1
Ours	EOG	77.3	74.3	80.6	73.3
	EEG	78.2	74.7	82.4	75.9
	Δ	0.9	0.4	1.8	2.6
TinySleepNet	EOG	-	-	75.7	66.9
	EEG	-	-	83.1	78.1
	Δ	-	-	7.4	11.2
RecSleepNet	EOG	72.1	67.8	76.2	67.4
	EEG	79.7	77.9	83.0	77.9
	Δ	7.6	10.1	6.8	10.5
U-time	EOG	-	71.0	-	70.0
	EEG	-	77.0	-	76.0
	Δ	-	6.0	-	6.0

Δ denotes the gap between singly using EEG and singly using EOG.

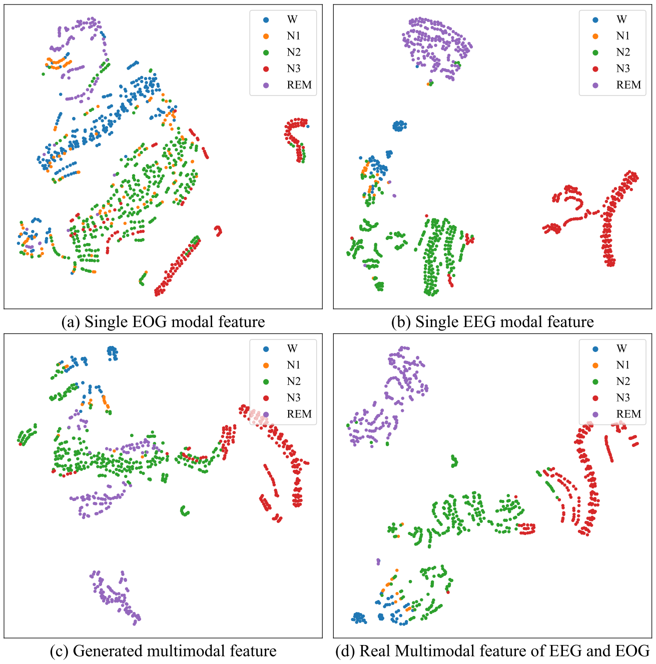
and multiple modalities of EEG and EOG, using the pretrained model (baseline model) in the Stage I (as shown in Fig. 2 Stage I). Notably, when we input the single EEG or single EOG modality, we remove the Fusion Block, designed to fuse multimodal features, from the model in Stage I. And then, we compare the results of the above three experiments with the performance of a single EOG modality using our framework, shown in Tab. II. As we can see, as expected in the most cases, inputting multimodal of EEG and EOG performs the best, which is reasonable as we mentioned above (On the MASS dataset, solely using EEG signals performs slightly better than using multimodal EEG and EOG). What is worth paying more attention is that, when using single EOG modality, our framework significantly improves the performance compared to the baseline model (1.96% average improvement in ACC and 2.4% in MF1), especially on the ISRUC (2.7% improvement in ACC and 3.5% in MF1) and MASS (2.8% improvement in ACC and 2.0% in MF1). It proves the effectiveness of our multimodal simplification framework in generating synthetic multimodal features even without EEG. Moreover, solely using EOG modality under our framework provides a comparable performance to using single EEG modality under the baseline model (80.9% vs. 82.5% on average ACC, 75.5% vs. 77.2% on average MF1). This significantly closes the gap between single EOG and single EEG under the model (an average difference of 3.6% in ACC and 4.2% in MF1 before and now with an average difference of 1.6% in ACC and 1.7% in MF1). Particularly, on the SSND dataset, using single EOG under our framework outperforms using single EEG modality under the baseline model (83.8% vs. 83.7% on average ACC and 79.3% vs. 78.8% on average MF1). This demonstrates the potential of EOG modality, which is convenient to collect using wearable devices in daliy life, making it possible to monitor the sleep quality in a home-based setting.

3) Ablation Study: In this experiment, we investigate the effectiveness of time-frequency domain generators of our framework. In our work, we design two generators using time and frequency domain EOG signal as guilding conditions, respectively, to generate synthetic multimodal features. Here, we conduct ablation experiments to validate the effectiveness of each generator in our model. The model variants are defined as follows:

- **G1**: the $generator2$ in frequency is removed from our framework.
- **G2**: the $generator1$ in time is removed from our framework.
- **G1+G2**: we apply both of $generator1$ and $generator2$ in time and frequency.

As shown in Tab. IV, both of time and frequency features contribute to generating synthetic multimodal features for sleep staging, and combining of them performs the best. As we can see, when we employ the time-domain EOG signal as the guiding condition, the model provides superior overall performance compared to using the frequency-domain EOG signal as the guiding condition (80.4% vs. 78.3% on average ACC and 74.6% vs. 72.6% on average MF1). In particular, on both HMC and MASS datasets, the frequency-domain EOG guided models perform much worse compared to the time-domain EOG guided models, about 3% to 5% lower in ACC and MF1, respectively. And using solely time-domain or frequency-domain as the guiding conditions perform closely on the datasets of ISRUC and Sleep-EDF. It may be caused by the differences in collection devices and environmental conditions during data gathering. In this ablation study, we set the hyperparameter α to 0.7 when fusing the dual synthetic features from different generators (α denotes the proportion of time-frequency domain feature fusion). The more details about the choice of α will be explained in subsequent experiment.

4) Comparison With Other EEG-Based Methods: In this section, we list three existing EEG-based methods [10], [11], [47] for comparision shown in Tab. V. We referred to the performance of the existing methods using EEG reported in the corresponding papers, and we obtained the performance of these methods using EOG by implementing them by



(a) Single EOG modal feature

(b) Single EEG modal feature

(c) Generated multimodal feature

(d) Real Multimodal feature of EEG and EOG

Fig. 4. Feature visualization based on different modalities, where different colors represent different sleep stages.

ourselves. Due to the page limit, here, we conducted this experiment on the two datasets of ISRUC and Sleep-EDF. The results demonstrate that our proposed framework provides a competitive performance when only using EOG, compared with only using EEG, and the gap is only less than 1% (ACC: 0.9%, MF1:0.4%) on the ISRUC dataset, and the gap is 1.8% in ACC and 2.6% in MF1 on the SleepEDF dataset. When compared to other existing models, although their performance based on EEG surpasses ours, the gap under the distinct input of EEG and EOG is much larger than that of ours. For example, there is an average ACC gap of 7.4% and an average MF1 gap of 11.2% on the SleepEDF dataset for TinySleepNet. Among the compared methods, U-time has the smallest gap between EEG and EOG, with an average MF1 difference of 6.0% on the both ISRUC and SleepEDF datasets.

5) Single and Multiple Modal Features Visualization: To demonstrate the effectiveness of our method, we chose a subject from the ISRUC dataset to visualize the intermediate features based on single or multiple modalities. The visualization is based on the t-SNE method [49]. Fig. 4 (a) and Fig. 4 (b) illustrate the feature distributions by employing single EOG and single EEG as input. Fig. 4 (c) and Fig. 4 (d) depict the generated and real multimodal feature distributions. As we can see, the samples from EEG modality belonging to the same sleep stage are nicely clustered within the same cluster in Fig. 4 (b), compared with the samples from the EOG modality in Fig. 4 (a). It demonstrates that the EEG modality has a more powerful predictive ability than that of the EOG modality. Notably, the samples represented by the generated multimodal features from the same stage also form a cluster, which looks quite similar to those by EEG modality features, as shown in Fig. 4 (b) and Fig. 4 (c). Compared with the single EOG modal features in Fig. 2 (a), where different stages lie in a chaotic, the nice clusters by the generated features in Fig. 2 (c) further prove the effectiveness of our method. As shown in Fig. 4 (d), based on the real multimodal feature of EEG and EOG, the different stages are clustered separately. The visualization comparison shows that, our method is capable of learning the correlation between EEG and EOG and generating reliable multimodal feature representations based on single EOG modality.

6) Analysis of Time-Frequency Generators Ratio: In our framework, the multimodal generators consist of two parts: a time-domain generator and a frequency-domain generator. The two generators share the same structure but use time and frequency domain EOG modality as a guiding condition, respectively. As mentioned above, we use a hyperparameter α to control the ratio of multimodal features fusion by dual generators. In this section, we explored the impact of the ratio α of time and frequency features for the fusion on the experimental results. We conduct the hyperparameter study on five datasets and set the α from 0 to 1 in increments of 0.1. As shown in Fig. 5, in most cases, when α is equal to 0.7, which means that the time domain synthetic feature has a higher proportion compared to the frequency domain feature (0.7 v.s. 0.3), the model performs the best. In other words, the synthetic feature generated using the time domain EOG signal as the guiding condition includes much more important information in this study. Particularly, on the Sleep-EDF dataset, the model performs the best when α is set to 0.9 (with an average accuracy of 80.6% compared to 80.7%). On the MASS dataset, when α is set to 0.8, the average MF1 score is slightly better (76.4% compared to 76.8% with α set to 0.7 and 0.8, respectively). Notably, on the Sleep-EDF and SSND datasets, the curves of their evaluation metrics vary more smoothly when compared to the other datasets, implying the networks are not very sensitive to the changes in the feature fusion ratio in these two datasets. To summarize, the variation trends of the metrics on the five datasets remain consistent: as α increases, the model's performance first improves and then declines, reaching its optimal performance within the range where time domain features have a larger proportion. Some subtle variation differences may be because of the different environmental conditions during data gathering.

V. Discussion

In this work, our proposed method can effectively simplify the multimodal sleep staging task, making the performance based on single EOG modality to closely approximate that based on EEG modality. This simplification framework allows us to make full use of the easy collection nature of EOG and the powerful capabilities of EEG in sleep staging. This makes it possible to use only single EOG for sleep staging. In current clinical practice, patients are required to wear a cap with dozens of electrodes to collect EEG and EOG data from several dozen channels for monitoring sleep quality in the hospital, which is complex and expensive. Hence, we are interested in exploring the possibility of home-based sleep monitoring. When in a home-based setting, it is hard to guarantee the high quality of the collected EEG signals due to its sensitivity to environment. The existing mainstream end-to-end sleep staging models require both EEG and EOG signals with high qualities to build sleep staging task to achieve good performance [50]. Fortunately,

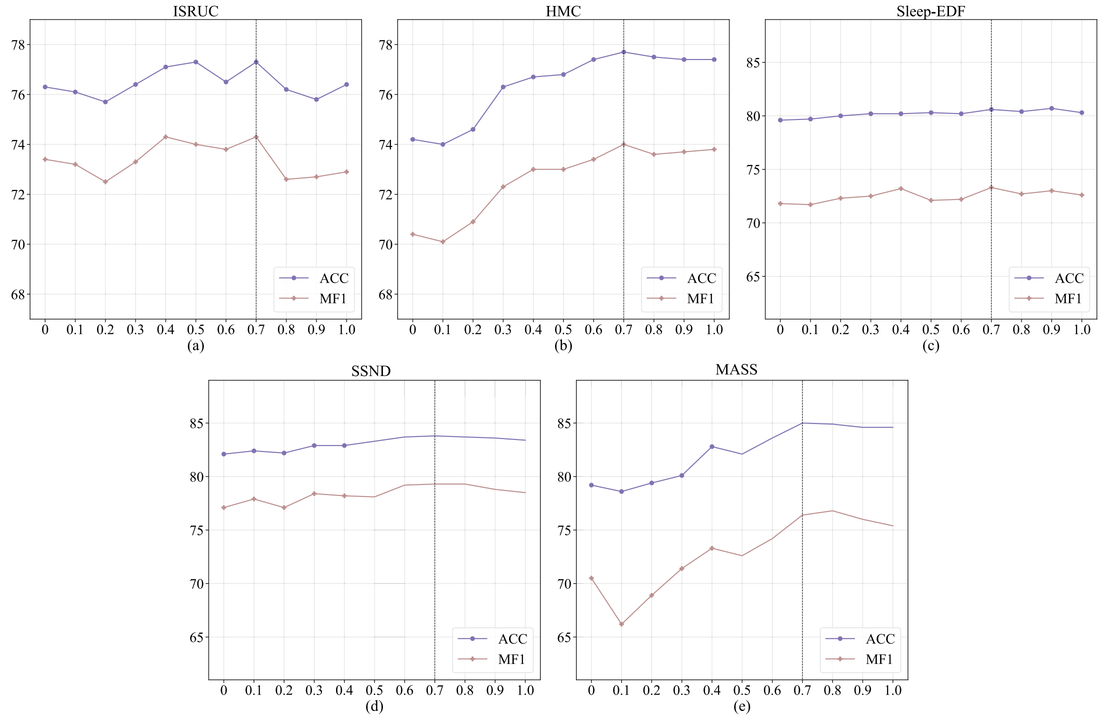


Fig. 5. Analysis of the Time-Frequency Generator Ratio, where we vary the hyperparameter α from 0 to 1 in increments of 0.1.

we can collect high-quality EOG signals at home due to their low environmental requirements. As mentioned above, our framework can simplify the multimodal sleep staging task and solely require EOG data as input. Therefore, our proposed method can be integrated in wearable devices to gather individual EOG data for sleep staging. Once after inputting the single EOG modality into the pre-trained network, the highly reliable predictions could be obtained. This process can be conducted at home by wearing a lightweight EOG data collection device, eliminating the need to visit a hospital and use professional EEG acquisition equipment to collect EEG data for sleep staging.

On the downside, there are still some limitations of this work. Firstly, the total number of subjects in our utilized datasets is very small. As shown in Fig. 2, our framework has two pretraining stages before we can use it for inference: one for obtaining real multimodal features and another for training the time-frequency generators and dual classifiers. The quality of the real multimodal features obtained from the pre-trained network can directly impact the ultimate performance in the test stage to some extent. This necessitates our utilization of a large-scale sleep staging dataset for pretraining. However, the largest dataset in our experiments is HMC, which contains only 153 PSG recordings. The size of the datasets limits the generalization of the two-step pre-trained network. Secondly, the individual discrepancies among different subjects are significant. Our method is fundamentally based on the pre-trained generators that can generate synthetic multimodal features from subjects using a single EOG modality. The input EOG modality combined with Gaussian noise, should align with the real multimodal features obtained from the training set. If there are significant individual differences between the target domain and the source domain, which means that the actual general multimodal feature distributions obtained from the training set may differ from those of the unseen subjects, potentially resulting in poor performance for them.

VI. Conclusion

In this paper, we propose a novel multimodal simplification framework for sleep staging that allows us to generate multimodal feature representations based on single EOG modality. Specifically, we first model the multimodal correlations between the EEG and EOG modalities. Leveraging this correlation, we adopt a conditional generative framework guided by the time-frequency EOG signals to generate multimodal feature representations in the absence of EEG modality. Then, we input single EOG modality in the test stage for sleep staging, reducing the dependence on EEG modality. The framework was evaluated on our collected dataset and four public datasets. Compared with existing methods, when only using EOG as input, our framework performs the best. Moreover, by our framework, the single EOG modality provides comparable performance to single EEG modality. The results demonstrate the potential of single EOG modality for sleep staging in clinics, overcoming the collection limitations of EEG. Motivated by the success of simplification multimodal with single EOG, in the near future, we plan to generalize the proposed framework to other more easily collected signals, such as ECG signal, for monitoring sleep quality, making sleep monitoring more easily accessible.

References

[1] H. Phan and K. Mikkelsen, “Automatic sleep staging of EEG signals: Recent development, challenges, and future directions,” *Physiological Meas.*, vol. 43, no. 4, Apr. 2022, Art. no. 04TR01.

[2] Z. Zhang, R. Yin, and H. Ning, “Internet of brain, thought, thinking, and creation,” *Chin. J. Electron.*, vol. 31, no. 6, pp. 1025–1042, 2022.

[3] F. He et al., "Effects of 20 Hz repetitive transcranial magnetic stimulation on disorders of consciousness: A resting-state electroencephalography study," *Neural Plasticity*, vol. 2018, pp. 1–8, 2018.

[4] R. B. Berry et al., "Rules for scoring respiratory events in sleep: Update of the 2007 AASM manual for the scoring of sleep and associated events: Deliberations of the sleep apnea definitions task force of the American academy of sleep medicine," *J. Clin. Sleep Med.*, vol. 8, no. 5, pp. 597–619, Oct. 2012.

[5] O. Tsinalis, P. M. Matthews, Y. Guo, and S. Zafeiriou, "Automatic sleep stage scoring with single-channel EEG using convolutional neural networks," 2016, *arXiv:1610.01683*.

[6] S. Mousavi, F. Afghah, and U. R. Acharya, "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach," *PLoS ONE*, vol. 14, no. 5, May 2019, Art. no. e0216456.

[7] 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, Jun. 2022.

[8] M. Melek, N. Manshouri, and T. Kayikcioglu, "An automatic EEG-based sleep staging system with introducing NAoSP and NAoGP as new metrics for sleep staging systems," *Cognit. Neurodynamics*, vol. 15, no. 3, pp. 405–423, Jun. 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] M. Perslev, M. Jensen, S. Darkner, P. J. Jennum, and C. Igel, "U-time: A fully convolutional network for time series segmentation applied to sleep staging," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 4415–4426.

[11] H. Nie, S. Tu, and L. Xu, "RecSleepNet: An automatic sleep staging model based on feature reconstruction," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2021, pp. 1458–1461.

[12] M. Diykh, Y. Li, and P. Wen, "EEG sleep stages classification based on time domain features and structural graph similarity," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 11, pp. 1159–1168, Nov. 2016.

[13] 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," *Frontiers Neurosci.*, vol. 15, Jul. 2021, Art. no. 573194.

[14] F. Andreotti, H. Phan, N. Cooray, C. Lo, M. T. M. Hu, and M. De Vos, "Multichannel sleep stage classification and transfer learning using convolutional neural networks," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 171–174.

[15] H. Sun et al., "Sleep staging from electrocardiography and respiration with deep learning," *Sleep*, vol. 43, no. 7, Jul. 2020, Art. no. zsz306.

[16] H. Phan, O. Y. Chen, M. C. Tran, P. Koch, A. Mertins, and M. De Vos, "XSleepNet: Multi-view sequential model for automatic sleep staging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5903–5915, Sep. 2021.

[17] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, "SalientSleepNet: Multimodal salient wave detection network for sleep staging," 2021, *arXiv:2105.13864*.

[18] A. Guillot and V. Thorey, "RobustSleepNet: Transfer learning for automated sleep staging at scale," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1441–1451, 2021.

[19] J. Wang, S. Zhao, H. Jiang, S. Li, T. Li, and G. Pan, "Generalizable sleep staging via multi-level domain alignment," 2023, *arXiv:2401.05363*.

[20] V. Bajaj and R. B. Pachori, "Automatic classification of sleep stages based on the time-frequency image of EEG signals," *Comput. Methods Programs Biomed.*, vol. 112, no. 3, pp. 320–328, Dec. 2013.

[21] Y. Wang et al., "DiffMDD: A diffusion-based deep learning framework for MDD diagnosis using EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 32, pp. 728–738, 2024.

[22] G. Pan et al., "Rapid decoding of hand gestures in electrocorticography using recurrent neural networks," *Frontiers Neurosci.*, vol. 12, p. 555, Aug. 2018.

[23] X. Sun, C. Qian, Z. Chen, Z. Wu, B. Luo, and G. Pan, "Remembered or forgotten?—An EEG-based computational prediction approach," *PLoS ONE*, vol. 11, no. 12, Dec. 2016, Art. no. e0167497.

[24] M. Perslev et al., "U-Sleep: Resilient high-frequency sleep staging," *NPJ Digit. Med.*, vol. 4, no. 1, p. 72, 2021.

[25] Z. Yubo, L. Yingying, Z. Bing, Z. Lin, and L. Lei, "MMASleepNet: A multimodal attention network based on electrophysiological signals for automatic sleep staging," *Frontiers Neurosci.*, vol. 16, Aug. 2022, Art. no. 973761.

[26] H. Hu, L. Xie, R. Hong, and Q. Tian, "Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3123–3132.

[27] S. Ren, Y. Du, J. Lv, G. Han, and S. He, "Learning from the master: Distilling cross-modal advanced knowledge for lip reading," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13320–13328.

[28] S. Zhang, C. Tang, and C. Guan, "Visual-to-EEG cross-modal knowledge distillation for continuous emotion recognition," *Pattern Recognit.*, vol. 130, Oct. 2022, Art. no. 108833.

[29] C. Zhang et al., "Multichannel multidomain-based knowledge distillation algorithm for sleep staging with single-channel EEG," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 11, pp. 4608–4612, Nov. 2022.

[30] H. Liang, Y. Liu, H. Wang, and Z. Jia, "Teacher assistant-based knowledge distillation extracting multi-level features on single channel sleep EEG," in *Proc. Thirty-Second Int. Joint Conf. Artif. Intell.*, Aug. 2023, pp. 3948–3956.

[31] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–12.

[32] X. Yan, L.-M. Zhao, and B.-L. Lu, "Simplifying multimodal emotion recognition with single eye movement modality," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1057–1063.

[33] H. Lee, H.-E. Kim, and H. Nam, "SRM: A style-based recalibration module for convolutional neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1854–1862.

[34] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[35] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. CVPR*, 2017, pp. 7167–7176.

[36] Q. Li, Q. Li, C. Liu, S. P. Shashikumar, S. Nemati, and G. D. Clifford, "Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram," *Physiological Meas.*, vol. 39, no. 12, Dec. 2018, Art. no. 124005.

[37] S. Mahvash Mohammadi, S. Kouchaki, M. Ghavami, and S. Sanei, "Improving time–frequency domain sleep EEG classification via singular spectrum analysis," *J. Neurosci. Methods*, vol. 273, pp. 96–106, Nov. 2016.

[38] 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.

[39] R. Wang, L. Qi, Y. Shi, and Y. Gao, "Better pseudo-label: Joint domain-aware label and dual-classifier for semi-supervised domain generalization," *Pattern Recognit.*, vol. 133, Jan. 2023, Art. no. 108987.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[41] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[42] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "ISRUC-sleep: A comprehensive public dataset for sleep researchers," *Comput. Methods Programs Biomed.*, vol. 124, pp. 180–192, Feb. 2016.

[43] B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberye, "Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 9, pp. 1185–1194, Sep. 2000.

[44] D. Alvarez-Estevez and R. M. Rijsman, "Inter-database validation of a deep learning approach for automatic sleep scoring," *PLoS ONE*, vol. 16, no. 8, Aug. 2021, Art. no. e0256111.

[45] C. O'Reilly, N. Gosselin, J. Carrier, and T. Nielsen, "Montreal archive of sleep studies: An open-access resource for instrument benchmarking and exploratory research," *J. Sleep Res.*, vol. 23, no. 6, pp. 628–635, Dec. 2014.

[46] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[47] A. Supratak and Y. Guo, "TinySleepNet: An efficient deep learning model for sleep stage scoring based on raw single-channel EEG," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 641–644.

[48] J. Chen et al., "SDREAMER: Self-distilled mixture-of-modality-experts transformer for automatic sleep staging," in *Proc. IEEE Int. Conf. Digit. Health (ICDH)*, Jul. 2023, pp. 131–142.

[49] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[50] J. Wang et al., "Narcolepsy diagnosis with sleep stage features using PSG recordings," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 3619–3629, 2023.