Seeing your sleep stage: cross-modal distillation from EEG to infrared video

Jianan Han, Shaoxing Zhang, Aidong Men, Yang Liu, Ziming Yao, Yan Yan, Qingchao Chen✉

***Abstract*—It is inevitably crucial to classify sleep stage for the diagnosis of various diseases. However, existing automated diagnosis methods mostly adopt the "gold-standard" Electroencephalogram (EEG) or other uni-modal sensing signal of the PolySomnoGraphy (PSG) machine *in hospital*, that are expensive, importable and therefore unsuitable for point-of-care monitoring *at home*. To enable the sleep stage monitoring *at home*, in this paper, we analyze the relationship between infrared videos and the EEG signal and propose a new task: to classify the sleep stage using infrared videos by distilling useful knowledge from EEG signals to the visual ones. It is different from previous video classification and multi-modal analysis tasks, mainly in that (i) the temporal duration of the infrared video is relatively long (10 hours per night); (ii) and the semantic gap between the EEG and infrared video is disparate and much larger than conventional cross-modal data in multimedia analysis such as video and audio. To establish a solid cross-modal benchmark for this application, we develop a new dataset termed as Seeing your Sleep Stage via Infrared Video and EEG (S^3VE). S^3VE is a large-scale dataset including synchronized infrared video and EEG signal for sleep stage classification, including 105 subjects and 154,573 video clips that is more than 1100 hours long. Our contributions are not limited to datasets but also about a novel cross-modal distillation baseline model namely the structure-aware contrastive distillation (SACD) to distill the EEG knowledge to infrared video features. The SACD achieved the state-of-the-art performances on both our S^3VE and the existing cross-modal distillation benchmark. Both the benchmark and the baseline methods will be released to the community. We expect to raise more attentions and promote more developments in the sleep stage classification and more importantly the cross-modal distillation from clinical signal/media to the conventional media. Code and open datasets are available at https://github.com/SPIResearch/SACD.**

***Index Terms*—sleep stage classification, dataset, EEG, infrared video, cross-modal distillation**

I. Introduction

It is of significance to estimate the sleep quality and stage accurately, as it is directly related to the phenomenon (phenotype) of chronic disease and mental disease. According to the recent scientific research results, millions of chronic and mental disease patients have sleep related problems that are highly correlated to daily life dis-functioning and even traffic accidents etc. It is essential to tackle this global healthcare problems by measuring the sleep quality accurately and in-time especially at home.

Existing analysis methods and sleep quality or stage classification approaches adopt the usage of EEG signal as the "gold standard" sensing modality; however, it is time-consuming and costly to estimate the quality in the hospital via analyzing and diagnosing the EEG signal from the PSG machine. In addition, the annotation efforts and training of clinical workers to use and diagnosis are expensive. Moreover, it is nearly impossible to setup PSG machine at home as it is expensive and difficult to operate on. Besides, the PSG operation mode needs to attach tens of sensors on the patients' head and they sometimes fell off, which generates inaccurate results. ***Therefore, it is of huge demand and extremely essential to estimate the sleep quality/stage using the portable and point-of-care sensors and solutions at home.***

In this paper, we propose a novel cross-modal methodology to solve the previous barriers, enabling point-of-care sleep stage monitoring at home. We propose to sense the human body visually via an infra-red camera video synchronized with the PSG EEG signal. As shown in the Fig. 1 (a), with the help of EEG signal features and distilling EEG knowledge to the visual features, we investigate the possibility, capability and limitations of distilled infra-red visual features to classify the sleep stage.

To enable the developments of point-of-care healthcare research and distillation methods from clinical to visual modality, to our best knowledge, we are the first to collect a large-scale cross-modal distillation dataset, namely S^3VE, including in total 1,100 hours synchronized infra-red videos and EEG signals, 105 subejcts from the real-world hospital and 154,573 multi-modal clips to investigate the problem of sleep stage classification. Besides the datasets, we also raise and analyze the following challenges of cross-modal distillation between the EEG and the infra-red data.

Challenge 1: the large cross-modal semantic gap between EEG and IR signals. EEG signal represents the intrinsic features of the sleep stage and is regarded as the gold-standard clinical diagnosis modality. However, the IR videos are commonly used for monitoring external characteristics of the subjects, e.g. motion, events and abnormal behaviour of the subjects. The two synchronized data in our S^3VE are less correlated and with a large semantic gap than those in the conventional audio-video classification benchmark, e.g. UCF51. (please see Fig. 1 (b) for more details). Therefore, we argue that directly aligning the cross-modal features of the same *instance* may lead to inferior distillation performance and the collapsed joint embedding space.

Challenge 2: the appearances of the infra-red sleep videos are similar globally and the differences among

✉ Corresponding author.

This work is supported by Peking University Medicine Seed Fund for Interdisciplinary Research (BMU2022MX011), Hygiene and Health Development Scientific Research Fostering Plan of Haidian District Beijing(HP2021-11-50102), the Fundamental Research Funds for the Central Universities, and PKU-OPPO Innovation Fund BO202103.

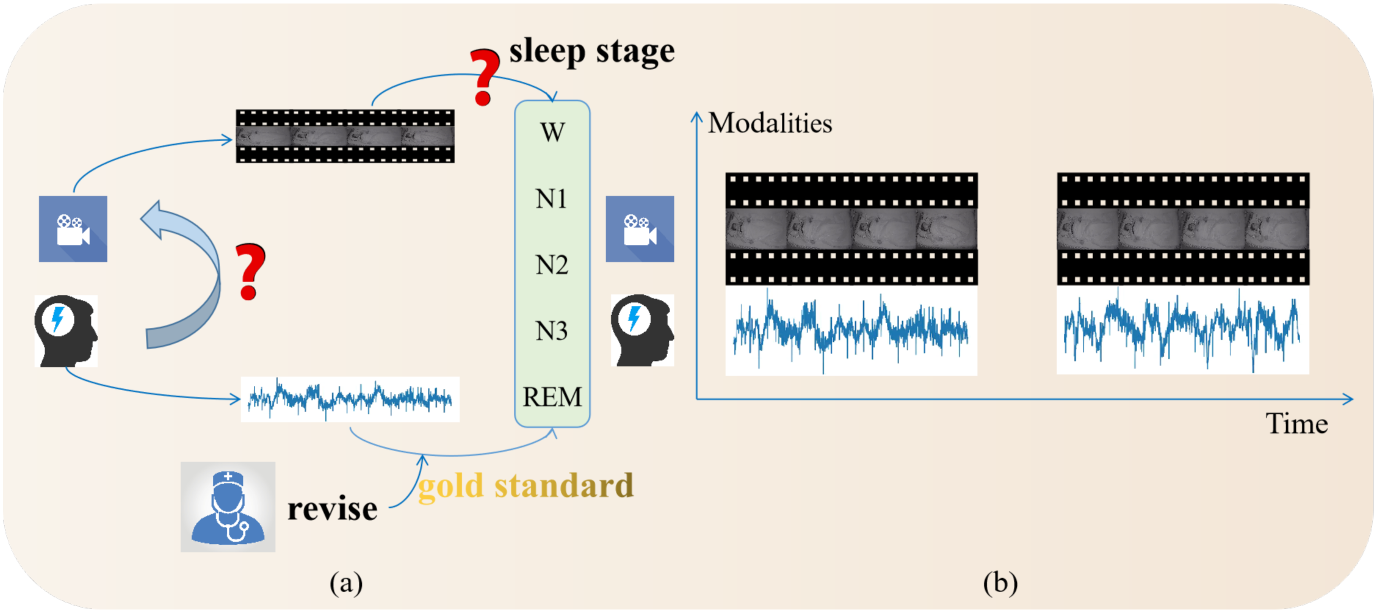

Fig. 1. (a). diagram of the task: cross-modal distillation from EEG to infrared video (b). semantic gap between video and EEG modality. The example in the figure is selected from clips with both labels $N1$; the EEG signal can be successfully judged, but the second term is judged as W through the video signal; In this $N1$ sleep state , some unknowing short-term movements affect the judgment of the IR video. For this example, the video cannot "see" the similarity of the two EEG signals, and the small movements of the legs and hands are equivalent to the confounding factor of the IR video modality.

different classes' videos are subtle locally. As shown in Fig. 1 (b), the global similarity between inter-class visual features are caused by the confounded similar background and scene. How to design a loss function to reflect and reveal the fine-grained *contrast* between visual pairs of the same or different classes remains a challenge in our dataset.

To tackle the first challenge, ***we propose to align the relationship structure formed by multiple cross-modal clip features instead of aligning the individual instance***. Due to the large semantic gap in our scenario, directly pulling the cross-modal features of the same *instance* may lead to inferior distillation performance and the collapsed joint embedding space. Therefore, we alternates to align the structural information. The intuition is that in spite of the instance-level semantic gap, the uni-modal structure relationships should be similar among multiple instances as a group. In addition, the cross-modal data exhibit unique temporal characteristics and a strong temporal correlation among the features in a mini-batch is observed. Formulating both points as a constraint, we propose to build a graph among multiple instances in each unimodal data and regularize the consistency of two unimodal relationship structures.

To tackle the second challenge, ***we propose to use the contrastive learning framework to learn the fine-grained "contrast" from the subtle differences in the IF video.*** However, conventional cross-modal contrastive learning is limited in our scenario, because the negative pairs are normally selected from the cross-modal features of different instances, which have large semantic gaps. Therefore, the normal negative pairs are not "hard" enough which cannot reflect and reveal the fine-grained contrast required in our setup. Therefore, we proposed to design two K-hard negative memory bank for the EEG and visual modality respectively, selecting the hard negative sample set in the online manner. In addition, we propose to use the symmetric contrastive cross-modal distillation to reduce the cross-modal semantic gap.

To sum up, we made the following contributions in this paper:

- To our best knowledge, we proposed a new and the first dataset and benchmark to investigate cross-modal distillation between the clinical EEG signal and the IF videos. We provided extensive experiments and comparisons with conventional distillation methods in the dataset.
- We also propose a novel cross-distillation method termed as Structure-aware contrastive distillation (SACD), including the structure-aware cross-modal alignment module and the dual memory banks for the contrastive learning.
- Our method achieves SOTA results on both our benchmark and the conventional distillation benchmark, e.g. UCF51.

II. Related works

A. Various Modalities of Sleep Stage Classification

Accurate sleep stage classification has been of great interest in analysing sleep quality and determining the effectiveness of treatment. As the EEG signal is considered as the "gold-standard" for sleep stage classification, The most mainstream approach is classifying and analyzing sleep stage by employing physiological electrical signals. In addition to PSG-based sleep grading and apnea-related studies, there are many other approaches. For example, Goederen *et al.* [1] used broadband radar to analyze children's sleep stages and sleep status. Deng *et al.* [2] designed adaptive vertical box (AV-Box) based breathing/snoring detection using a decision tree classifier for sleep stage classification. Korkalainen *et al.* [3] identified the sleep stages from the photoplethysmogram (PPG) signal obtained with a simple finger pulse. Yi *et al.* [4] extracted a total of 74 features, including heart rate variability (HRV), features, respiratory rate variability (RV) features, and linear frequency cepstral coefficients (LFCC) from bed sensor data and performs sleep stage classification. ***Unlike the above methods, we are the first to systematically investigate the use of infrared sleep video for sleep stage classification and have developed a new dataset.***

B. PSG-based Sleep Stage Classification Datasets and Methods

Sleep Stage Classification Datasets: Various physio-electrical signal datasets have been collected for sleep research. *The Sleep Heart Health Study (SHSS)* was a multi-center cohort sleep study [5] [6], whose two dataset versions, SHHS-1 and SHHS-2, contained the polysomnograms (PSG) data of 6441 and 3295 subjects respectively. The PSG data consist of multi-channel physio-electrical signals, including the Electroencephalogram(EEG) (C3-A2 and C4-A1), Electrooculogram (EOG), Electromyogram (EMG), Thoracic excursions (THOR) and abdominal excursions (ABDO), etc. *The SleepEDF-20 and SleepEDF-78* were obtained from the PhysioBank [7], including 20 and 78 subjects respectively. The data contains 2 EEG channels (Fpz-Cz and Pz-Oz). *The Montreal Archive of Sleep Studies (MASS) dataset* [8] was collected including the whole-night sleep data from 200 subjects (103 females and 97 males), aged from 18-76. It mainly consists of about 20 EEG channels, plus EOG, EMG, ECG, and respiration signals. ***Different from the previous datasets, to our best knowledge, we proposed a novel cross-modal dataset to promote multi-modal learning for sleep stage classification, including the synchronous IR videos and the EEG signals.***

PSG-based Automated Sleep Stage Classification Methods: Multiple sleep classification methods have been proposed using the previously mentioned datasets. Conventional machine learning methods extracted time-frequency analysis features and applied the Support Vector Machine (SVM), random forest, wavelet transform and information entropy algorithms [9] [10] [11]. However, these methods incorporated strong prior knowledge and hand-crafted features, therefore the classification accuracy relies on the feature qualities. Recently, deep learning based methods have been the mainstream for sleep stage classification. For single-channel EEG classification, the AttnSleep [12] used the multi-resolution convolutional neural network (MRCNN). The dilated convolution and synthetic minority oversampling technique (SMOTE) [13] also achieved competitive results. As for methods using multi-channel EEG signals, the BrainSleepNet [14] captured the comprehensive features of multi-channel EEG signals. The MSTGCN [15] and GraphSleepNet [16] used the structure-aware encoders for automatic sleep staging. In addition, a joint CNN framework [17] adopted temporal information as a context to predict sleep stages. ***However, our dataset consists of sleep data of 105 subjects and most differently, we leveraged the synchrounous EEG and infra-red videos to analyze and trian the classification model, which is not available in the above datasets. To our best knowledge, we are the first to investigate the usage of EEG signals and video for sleep stages classification.***

C. Cross-modal distillation methods

Knowledge distillation (KD): The KD [18] [19] [20] [21] methods transfer knowledge from the teacher to the student network, by supervising the student network using the pseudo labels. Komodakis *et al.* [22] used an attention-based distillation method to match the activation-based and gradient-based spatial attention maps. The flow of solution procedure (FSP) [23], generated by computing the Gram matrix of features across layers, was used to transfer knowledge. The RKD [24] captures cross-instance relations by designing a loss function to penalize the structure variations. Li *et al.* [25] developed a new framework to correct noisy labels by using knowledge learned from small clean datasets and semantic knowledge graphs. ***Different from them, we not only pull the features between the student and the teacher directly, but also allows the student network to learn the relative positional relationships between instances in the teacher's network during the distillation process. This relative positional relationship, specifically (which instances are close, far, and how they are distributed), within a mini-batch, is what we call "structure." Hence the name of our distillation method is "structure-aware."***

Structure—aware distillation methods: Structure-aware distillation adopts the idea of distilling knowledge from structural data. The CRCD [26] estimates the mutual relation and transfers structured knowledge from anchor teacher to anchor student in a contrastive learning framework. Pairwise distillation methods distilled pairwise and holistic similarities [27]. The similarity-preserving KD [28] constrains the similarity between teacher network features and the student one, which complements the conventional distillation methods. ***However, different from their works, we propose to a new method use graph neural network to model the similarity relationship between different modalities, by updating the "node" and "edge" inside one modality, forming an entire graph-level representation to describe this modality and then take them closer.***

Cross-modal distillation using contrastive learning: Most recently, KD methods using the contrastive learning pulls the representations of positive pairs but push the negative pairs. The Contrastive Multiview Coding (CMC) [29] is a cross-view learning method to align different views of the same instances by contrastive learning. The Contrastive Representation Distillation (CRD) [30] transfers knowledge by instance-level contrastive learning and uses a large memory bank to store negative samples. Chen et al. [31] proposed to use contrastive learning to distill information from the image and audio to video analysis. ***Different from other contrastive learning KD method, we proposed a novel dual-modality K-hard negative queues, storing the negative samples of the sleep stages. In***

TABLE I
CHARACTERISTICS OF FIVE SLEEP STAGES.

Sleep stages	Characteristics
W	Awake. An EEG contains β waves when the eyes are closed and α waves when the eyes are open.
$N1$	Transitions from W to other stages. Cranial apex waves are present in later stages.
$N2$	Spindles or unawakening associated K-complex waves are present.
$N3$	High amplitude low frequency σ wave appears.
REM	There is rapid eye movement and typical sawtooth wave

addition, we designed the symmetric contrastive distillation losses, leveraging negative pairs from two modalities.

III. DATASETS AND BENCHMARKS

A. Problem Formulation

Sleep disordering has become an important problem and it is of great significance to establish a sleep stage classification standard for sleep medicine. The American Academy of Sleep Medicine (AASM) standard [32] sets out the rules, terminology and techniques for sleep and related events. AASM divides human sleep into five stages, including the Rapid Eye Movements(REM), Wake(W), Non REM1 ($N1$), Non REM2 ($N2$) and Non REM3 ($N3$). The characteristics of each stage are shown in the following Table I [33].

B. Datasets and benchmarks construction

Data Collection. We collected the synchronized EEG and the infra-red video signals from the Peking University Third Hospital. The dataset collection time spans more than two yearsIn order to ensure the diversity of subjects in the dataset, we selected subjects of different ages (the youngest is 7 years old; the oldest is 70 years old), genders, and sleep apnea indices. Through consulting with a clinical expert, we formulated the training, validation, and test sets accordingly. The data collection time is approximately from 9:30pm to 7:30am the next day.

Annotations. In the process of constructing the dataset annotations, the coarse-grained annotations are firstly generated from the PSG machine, and then they are inspected and examined by five well-trained sleep apnea physicians. In this process, some unreasonable and mis-classified labels are modified through discussions to achieve agreements. To eliminate the effects of subjective diagnosis, weekly cross-checks will be performed to ensure the agreed consensus. Conflicted annotations and opinions are collectively discussed and voted so that the annotation agreements within each subject should be more than 95%. At present, the above annotation protocol is widely recognized by various medical institutions and hospitals, and is considered to be the “gold-standard” for sleep stage classification.

C. Dataset Statistics and Properties

The time duration statistics. As shown in the upper left subfigure in Fig. 2, we collected PSG signals and the synchronous infrared videos in our dataset, including the monitoring of 105 subjects’ (82 Males and 23 Females) whole night data, consisting of 154,573 data clips. Each clip is 30 seconds long and the whole dataset contains 1124 hours’ data. The number of clips per subject ranges from 1080 to 1360.

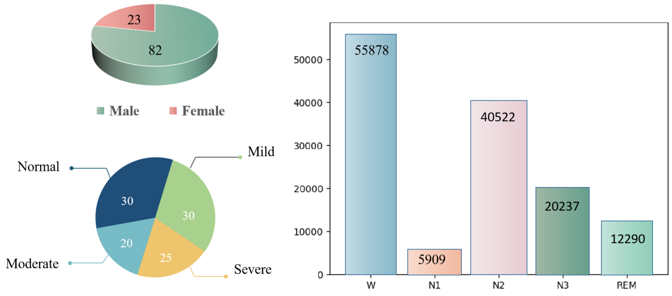

Fig. 2. Statistical information on our dataset. The top left subfigure shows the distribution of male and female cases in the dataset; The bottom left subfigure shows the distribution of each AHI-related group case in the dataset; The right-hand subfigure shows the distribution of each sleep stage in the dataset.

The Apnea-Hypopnea Index (AHI) distribution. The medical community usually classifies AHI meaningful into four clinically significant groups(<5, 5-15, 16-30, >30). The above four groups correspond to normal, mild obstructive sleep apnea (OSA), moderate obstructive sleep apnea (OSA), severe obstructive sleep apnea (OSA), respectively. As shown in the bottom left subfigure of Figure 2, out of the 105 patients in our dataset, 30 are normal, 30 are mild, 20 are moderate, and 25 are severe.

Sleep stage distribution. Adult’s sleep cycle lasts about 90 to 100 minutes, alternating about four to five times in a night. As shown in the right subfigure of the Fig. 2, there are five stages, where W represents wake period, $N1$ denotes the sleepy phase, which lasts about five minutes that describes the period between awake and falling asleep. $N2$ represents the period of light sleep, and with feeling of falling or weightlessness during sleep, as well as sudden body twitching. $N3$ is a deep sleep phase, in which brain wave activity drops to 1-2 seconds, and the respiration and heart rate reach the lowest. Deep sleep period accounts for about 20% of the sleep time per night. REM is a period of rapid eye movement, in which the eyes begin to move rapidly and the blood pressure, heart rate and respiration rate are more active than in the Non-REM stage. As shown in the right subfigure in Figure 2, in our dataset, there are 55,878, 5909, 40,522, 20,237 and 12,290 clips in W, $N1$, $N2$, $N3$ and REM stages respectively.

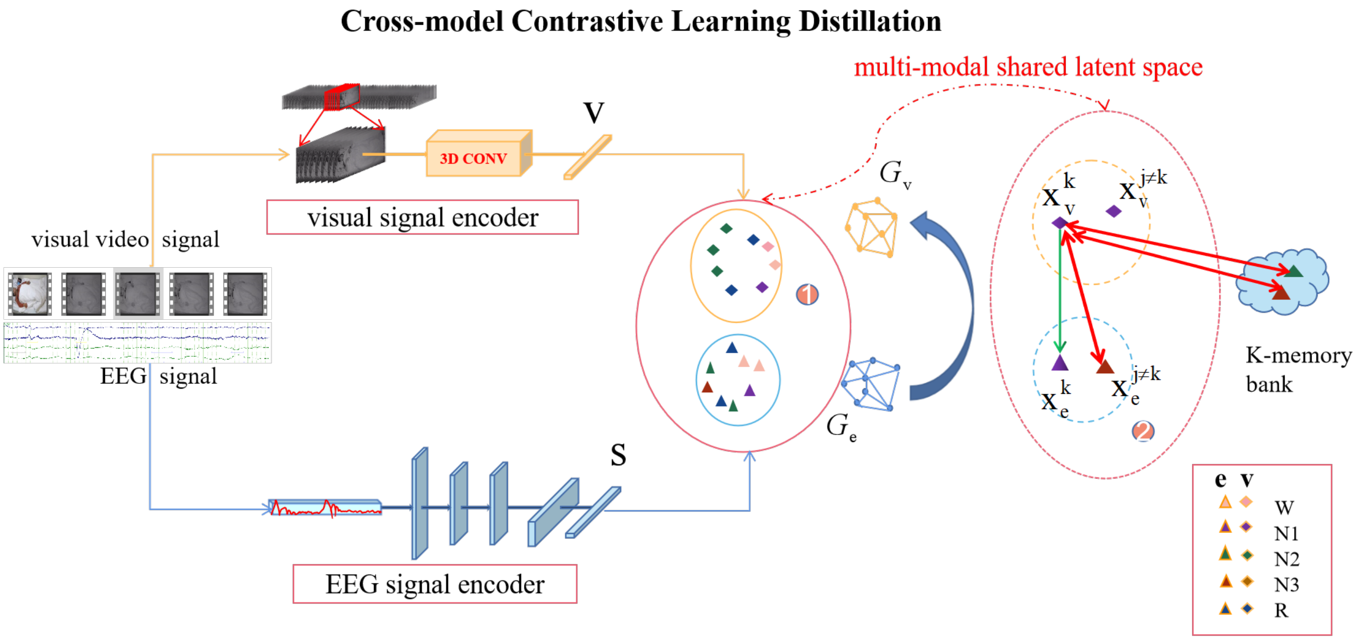

Fig. 3. Overall network architecture of cross-modal distillation.

IV. Method

A. Overall Framework

The overall architecture is shown in Fig. 3, where we are given the i^{th} synchronized EEG signal s_i, the ground-truth sleep stage annotation y_i and the Infra-red video v_i collected from real-world subjects in the hospital (see more analysis and statistics of the dataset in section III-B). Given the video encoder E_v and EEG signal encoder E_s, the conventional cross-entropy loss L_{CE} is utilized to train E_v, E_s and the EEG and Infra-red classifiers C_s and C_v respectively as the following Eq.(1):

$$\min_{E_v,E_s,C_v,C_s} \sum_i L_{CE}(C_v(E_v(v_i)),y_i)+L_{CE}(C_s(E_s(s_i)),y_i). \tag{1}$$

The aim of our proposed method is not only to train an Infra-red video classification network, but to distill the cross-modal knowledge and discriminative structural information from the clinically-recognized "gold-standard" EEG feature $E_s(s)$ to the portable and point-of-care visual infra-red video feature $E_v(v)$. If it is achieved, we are able to classify the sleep stage using the infra-red video features only at home. Our intuition is that the infra-red video can capture the visual appearance patterns of the sleeping subjects but EEG signal cannot. However, the EEG signal can capture the intrinsic features of the brain electric signal that are clinical relevant but the videos cannot. Therefore, we argue that there exists a huge semantic gap between infra-red video feature distribution $P(E_v(v_i))$ and the EEG signal one $P(E_s(s_i))$ and their features complement each other in the sleep stage classification.

To distill the discriminative knowledge from $P(E_s(s_i))$ ***to*** $P(E_v(v_i))$***, the first challenge is to tackle the cross-modal semantic gaps.*** We propose a structure-aware cross-modal distillation module, consisted of two graphical neural networks G_v and G_s to model the unimodal inter-sample relationships $G_v(E_v(v))$ and $G_s(E_s(s))$ respectively. Then we propose to reduce the cross-modal gaps of inter-sample relationships by reducing their structural distance loss L_D based on a metric D. The optimization is shown in the following Eq.(2) and more details are described in Section IV-B.

$$\min_{E_v,E_s,G_v,G_s} \sum_{v,s} L_D(G_v(E_v(v)),G_s(E_s(s))). \tag{2}$$

Reducing the structural cross-modal gap does not align the fine-grained semantic concepts between EEG and visual embedding space. ***To distill the discriminative and fine-grained knowledge from*** $P(E_s(s_i))$ ***to*** $P(E_v(v_i))$, we propose a cross-modal contrastive distillation framework utilizing two hard-negative memory selectors that stores the K-hardest negative samples based on the EEG and the video anchor samples. Specifically, our cross-modal contrastive distillation framework trains the encoders based on the contrastive learning loss $L_C(v_i,s_i,v_j,s_h)$ in Eq.(3), where v_i,s_i are the i^{th} cross-modal sample as the anchor points, while the v_j,s_h are the negative pairs optimally selected based on our novel $K-$hardest negative sample selection module. The selection module consists of two memory queues designed for two modalities respectively. In our cross-modal contrastive distillation, the positive pairs are the synchronized features $< E_v(v_i), E_s(s_i)) >$ but our selection module leveraged two negative pairs v_j and s_h, that are selected optimally from EEG and the visual memory queue Q_s and Q_v respectively. The selection criterias are as follows: for a cross-modal target anchor pair v_i,s_i, the hard negative pair are selected as v_j and s_h respectively, where $j = argmax(E_s(s_i)^T E_v(v_j)), for\ \forall v_j \in Q_v$ and $h = argmax(E_v(v_i)^T E_s(s_h)), for\ \forall s_h \in Q_s$. This procedure maintains and aligns the fine-grained cross-modal semantic embedding. More details are shown in Section IV-C.

$$\min_{E_v,E_s} \sum_{v,s} L_C(E_v(v_i), E_v(v_j), E_s(s_i), E_s(s_h)). \tag{3}$$

Besides the feature space semantic alignment, we also adopt the conventional class prediction space distillation as shown in the following Eq.(4). We utilized and reduced the well-known Jenson-Shannon-Divergence (JSD) between the visual prediction distribution $P(C_v(E_v(v)))$ and the EEG one $P(C_s(E_s(s)))$. The overall optimization is shown in Eq.(5) where λ_1,λ_2 and λ_3 are hyper-parameters.

$$\min_{E_v,E_s,C_v,C_s} JSD(P(C_v(E_v(v))), P(C_s(E_s(s)))). \tag{4}$$

$$\min L_{CE} + \lambda_1 L_D + \lambda_2 L_C + \lambda_3 JSD. \tag{5}$$

B. Structure-aware coarse-grained semantic alignment

To distill the knowledge from clinical signal to IR videos, the two synchronized data in our S^3VE are less correlated and with a large semantic gap than those in the conventional audio-video classification benchmark, e.g. UCF51 [34]. Therefore, we argue that directly pulling the cross-modal features of the same *instance* and pushing the features of different *instances* may lead to inferior distillation performance and the collapsed joint embedding space.

To tackle the previous challenge, *we propose to align the relationship structure formed by multiple cross-modal clip features instead of from the individual instance*. Our intuition is that in spite of the instance-level semantic gap, the uni-modal structure relationships should be similar among multiple instances as a group. In addition, the cross-modal data exhibit unique temporal characteristics and a strong temporal correlation among the features in a mini-batch is observed. Formulating both points as a constraint, we propose to build a graph among multiple instances in each unimodal data and regularize the consistency of two unimodal relationship structures.

More specifically, given the i^{th} clip of an input mini-batch, $E_s(x_i)$ and $E_v(x_i)$ denotes the output features of two modalities' graph level encoders. Below, we take the video and EEG signal v_i, e_i as an example, Each graph $G = (V, E)$ is represented as sets of nodes V and edges E,

$$h_{v_i}^{(0)} = \mathrm{MLP}_{\mathrm{node}}\left(x_{v_i}\right) \tag{6}$$

$$e_{ij} = \mathrm{MLP}_{\mathrm{edge}}\left(x_{v_{ij}}\right) \tag{7}$$

The output of $E_v(x_i) : h_{v_i}^{(0)}$ is a 64-dimensional vector,through an MLP, the hidden node vector $h_{v_i}^{(0)}$'s dimension is 128, and the edges vector e_{ij}'s dimension is 64 . Nodes are transfer information in the propagation layers. After the $t-th$ pass, the propagation layer maps a node representations $h_i^{(t)}$ to new node representations $h_i^{(t+1)}$, here node's dimension have not change. F_m is an MLP work on the concatenated inputs of $h_i^{(t)}$, $h_j^{(t)}$ and e_{ij}, here the output vector dimension here is 512-dim. then we utilize weighted summation based on graph attention mechanism A [35] to update node features. Through multiple layers of propagation, the representation for each node will accumulate information in its local neighborhood.

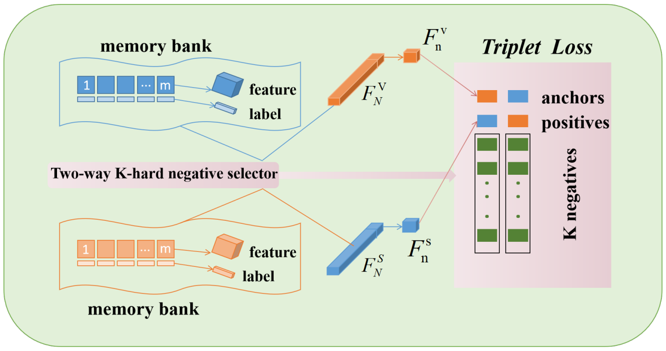

Fig. 4. Two K-hard negative memory bank structure diagram, and a momentum updated Encoder to generate momentum updated embeddings stored in a large memory bank.

$$m_{j\rightarrow i} = F_m\left(h_i^{(t)}, h_j^{(t)}, e_{ij}\right) \tag{8}$$

Here F_m is an MLP on the concatenated inputs including neighbor node features and edge features.

$$h_i^{(t+1)} = A\left(h_i^{(t)}, \sum_{j\in\delta_i} m_{j\rightarrow i}\right) \tag{9}$$

After the $t-th$ layer propagation, a graph level representation $O_{(G_v)}$ can form through a $READOUT$ function and a graph level MLP_G , and here we choose a fully connected layer as the READOUT function. $O_{(G_v)}$ is a 1024-dim vector.

$$O_{(G_v)} = MLP_G(READOUT\{h_i^{(t)} : X_{v_i}\} \in V) \tag{10}$$

we use Euclidean similarity to compute the structural distance loss L_D between G_v and G_a :

$$\begin{aligned} L_D &= E_{(G_e,G_v)}(max, r - d(G_v, G_a)) \\ d(G_V, G_a) &= ||O_{(G_v)} - O_{(G_v)}||^2 \end{aligned} \tag{11}$$

where $\gamma > 0$ is a margin parameter.

C. Hard-negative fine-grained cross-modal contrastive learning

As mentioned in the introduction section, the IR video appearance features are less discriminative than the EEG signal, especially since the inter-class differences are locally subtle in the video. To distil the discriminative EEG feature to the ambiguous visual ones, usage of the contrastive learning strategy–pulling positive pairs and pushing away the negative ones seems a straightforward choice. However, conventional cross-modal contrastive learning is limited in our scenario because the negative pairs are normally selected from the cross-modal features of different instances, which have large semantic gaps. Therefore, the normal negative pairs are not "hard" enough, which cannot reflect and reveal the fine-grained contrast required in our setup. Furthermore, the current contrastive learning method usually obtains negative samples in the same batch, and the number of negatives directly affects the contrastive loss. For our task, since the distribution of clips within a batch is usually flat, the number of negative samples is insufficient, and it is possible to take negative samples within one class to indirectly affect the classification performance.

To tackle the challenges, *we propose two K-hard negative memory bank Figure 4 for the EEG and visual modality respectively, selecting the negative sample set in the online manner. In addition, we propose to use the symmetric contrastive cross-modal distillation to reduce the cross-modal semantic gap.*

Specifically, each K-hard negative memory bank stores the first K hardest negative pairs for each class categories, preserving the global hard negative sample features in the dataset level. This helps ensure that the contrast in the optimization are fine-grained. In addition, to shrinkage the cross-modal semantic gap, we proposed the two-way symmetrical contrastive learning loss as shown in the following equation(12) :

$$\begin{aligned} L_{D_N} = \sum_{n=1}^{N} \{ & [\alpha - S(F_n^v, F_n^s) + S(F_n^v, F_i^s)]_+ \\ & + [\alpha - S(F_n^v, F_n^s) + S(F_j^v, F_n^s)]_+ \} \end{aligned} \tag{12}$$

where α is the margin we set 0.2 here and can be tuned using the validation sets. $S(\cdot)$ is the similarity function in the feature space. F_n^v, E_n^s is the output feature from encoder E_v and E_s , which are consistent with the description in IV-A. i and j are the index for the hard negatives , n is the anchor index for distillation and N denotes the batch size.

Specifically, we used two K-hard negative memory bank to enlarge the hard negative sample set, the memory bank M_B stores the feature representation of historical samples with a size of m. We rewrite the formula as:

$$\begin{aligned} L_{D_N}^K = \sum_{k=1}^{K} \sum_{n=1}^{N} \{ & [\alpha - S(F_n^v, F_n^s) + S(F_n^v, F_k^s)]_+ \\ & + [\alpha - S(F_n^v, F_n^s) + S(F_k^v, F_n^s)]_+ \} \end{aligned} \tag{13}$$

During the network training, we dynamically update the memory bank by discarding the oldest items and feeding the new batch of embedded features, where the memory bank acts as a queue. And we also keep the class label along with representation in the memory bank to filter the negatives.The update method of memory bank we use is Momentum updated Encoder (MuEncoder) [36]. The memory bank structure is shown in the figure 4.

V. Experiment

A. Dataset Splits

Because some samples have individual reasons (short sleep time, difficulty falling asleep, strange posture, etc.), we reduced the samples by 3 to 102 subjects.

According to the sleep apnea indexes (no apnea, mild apnea, moderate apnea, severe apnea), samples at each apnea stage are subdivided, and 80% of each apnea stage is taken as the training set (Among them, 20% were chosen as the validation set), the last 20% of each apnea stage is taken as the testing set. Twenty people are used for the test to ensure that the training and testing are equally distributed. Therefore, the clip number of the train set is 102,519, the clip number of the validation set is 20,503, and the clip number of the test set is 31,551.

B. Implementation Details and Metrics

To be fair and comprehensive, we use R(2+1)D-18 [37] and R3D-18 [38] as the video student network, which is pretrained in Kinetics-700 [39] and Moments in Time [40] and fine-tuned in our dataset (S^3VE). We use AttnSleep [12] as the EEG teacher network, which is only trained on our dataset, and the model weights of the teacher network are kept frozen during training. The video network baseline is trained by SGD with a learning rate of 0.001, and a weight decay of 0.0005. When training, our batch size is set to 16; we trade off computational efficiency and set the size of the memory bank as 256, the number of negatives K as 64. The hyperparameters λ_1, λ_2 and λ_3 are set to 0.5, 1 and 1. The dimension of the latent feature space is 512. We do not add projections on the network, but linear projections can be added to map all embeddings to the same dimension. The video clips are cropped to 320×240 and each clip contains about 750 frames. During the testing phase, just the IR sleep video was used for classification.

We adopted the following three evaluation metrics to measure the performance of the sleep classification: the accuracy (ACC), macro-averaged F1-score (MF1) and Cohen Kappa (κ) [41].They are calculated as follows:

$$ACC = \frac{\sum_{c \in C^S} TP_c}{N} \tag{14}$$

$$MF1 = \frac{\sum_{c \in C^S} F1_c}{5} \tag{15}$$

where TP_c and $F1_c$ are the true positive and per-class F1 score of class $c \in C^S$, respectively, and N is the total number of test samples. κ is a statistical measure of the interrater agreement (IRA) level calculated as:

$$\kappa = \frac{\sum_{c \in C^S} p_{cc} - \sum_{c \in C^S} p_{c+} p_{+c}}{1 - \sum_{c \in C^S} p_{c+} p_{+c}} = \frac{p_a - p_e}{1 - p_e} \tag{16}$$

where $p_c c$ represents the percentage of epochs classified as category c by the network and the annotated label simultaneously,and p_c+ and p_+c represent the percentages of epochs classified as category c by the network and annotated label, respectively.

C. Results and Comparisons

We compare our SACD with the only infrared video baseline models without any distillation, and six state-of-the-art other distillation methods **(Fitnet [42], PKT [43], CRD [30], IFD [44], CMC [29], CCL [31])**. For the fairness and comprehensiveness of the experiments, we train each model on the same experimental setup but replace the distillation objective based on their open-source implementation.

As shown in Table II, our method achieves state-of-art in the accuracy (ACC), macro-averaged F1-score (MF1), Cohen Kappa (κ) and per-class accuracy. By comparison, we found that CRD [30] and CCL [31] are the two most powerful opponents. Both methods apply contrastive learning, which

TABLE II
COMPARISON RESULTS (%) AMONG SACD AND STATE-OF-ART MODELS. THE BEST VALUES ON S^3VE DATASET ARE HIGHLIGHTED IN BOLD. THE NUMBER FOLLOWING ± REPRESENTS THE STANDARD DEVIATION OF MULTIPLE EXPERIMENTS.

		Overall Metrics			Per-class Accuracy				
EEG Baseline	Distillation Method	Accuracy	MF1	κ	W	N1	N2	N3	REM
-	Baseline R(2+1)D	52.5	46.6	0.47	64.7	41.3	48.4	47.8	57.3
-	Baseline R3D	52.8	46.9	0.37	64.7	41.4	49.0	48.0	58.0
Attnsleep	Fitnet	56.1	49.9	0.49	68.5	47.2	55.9	53.8	61.8
	PKT	58.1	51.2	0.53	71.2	47.8	56.3	54.0	62.9
	CRD	62.3	55.8	0.56	73.8	49.8	58.8	57.1	66.4
	IFD	61.1	54.0	0.57	73.0	49.9	58.8	57.1	65.1
	CMC	58.9	53.1	0.53	71.5	46.2	56.9	55.0	63.2
	CCL	62.3	57.0	0.56	74.4	50.1	58.7	53.8	64.7
	SACD(ours)	**64.4±0.45**	**58.9±0.40**	**0.60±0.38**	**75.6±0.56**	**51.0 ±0.36**	**60.2±0.42**	**59.3±0.45**	**67.8+0.45**
DeepSleepNet	Fitnet	54.6	47.8	0.46	67.0	45.2	51.9	49.8	59.5
	PKT	55.7	48.6	0.49	69.6	46.2	52.3	50.1	59.8
	CRD	58.3	51.9	0.50	70.2	47.8	55.0	53.2	62.4
	IFD	57.2	51.3	0.49	69.8	46.8	54.0	53.4	61.2
	CMC	55.5	50.8	0.48	68.5	44.8	53.0	51.0	59.5
	CCL	58.7	52.3	0.52	71.0	47.7	54.7	50.8	61.0
	SACD(ours)	**60.7±0.52**	**54.6±0.39**	**0.54±0.45**	**72.2+0.60**	**48.4±0.36**	**56.2±0.38**	**55.3±0.42**	**63.8±0.45**
Modified-SEN	Fitnet	54.9	48.9	0.48	67.7	45.5	52.2	50.1	60.0
	PKT	57.5	51.0	0.50	51.4	48.0	54.1	51.7	61.6
	CRD	60.5	53.9	0.53	72.6	48.7	57.0	55.1	64.6
	IFD	59.0	53.0	0.51	71.7	48.6	55.6	54.8	63.0
	CMC	57.0	51.1	0.51	70.0	46.6	54.5	52.4	61.2
	CCL	60.6	54.2	0.53	72.8	**49.6**	56.5	52.7	62.8
	SACD(ours)	**62.6±0.45**	**55.6±0.48**	**0.56±0.38**	**73.8±0.52**	49.5±0.45	**58.5±0.55**	**57.3±0.48**	**66.1±0.36**

TABLE III
VIDEO CLASSIFICATION ON THE PUBLIC DATASET UCF51. METRIC:TOP1 ACCURACY (%). KNOWLEDGE IS TRANSFERRED FROM AUDIO MODALITY TO IMPROVE THE VIDEO RECOGNITION MODEL.

	UCF51							
Methods	Baseline R(2+1)D	Fitnet	PKT	RKD	CRD	CMC	CCL	SACD(ours)
Audio to Video	57.5	48.4	53.2	53.0	60.3	59.2	64.9	66.0

indirectly shows that contrastive learning can significantly improve the cross-modal tasks. However, we are using a new contrastive learning method which has enabled us to achieve an overall lead in all evaluation metrics. Specifically, our method SACD is 11.9% (52.5%-64.4%) higher than the baseline in terms of accuracy, and there is also a 12.3% (46.6%-58.9%) and 0.13 (0.47-0.60) increase in both MF1 and κ, respectively. Compared to our next closest rival CCL, we are also 2.1% (62.3%-64.4%), 1.9% (57.0%-58.9%) and 0.04 (0.56-0.60) higher in ACC, MF1 and κ, respectively; and our method is also higher than the CCL in all five separate sleep stages, as shown in the TABLE II. The above results show that our method SACD has superior performance on our dataset S^3VE and outperforms other SOTA methods under various evaluation metrics.

To demonstrate that our cross-modal distillation method works not only under a single EEG modality baseline (AttnSleep), we also performed experiments similar to the previous section under more SOTA EEG baselines. Therefore, we chose DeepSleepNet [45] and a single-channel version of SEN-DAL [46] (no EOG signal is input, only a single-channel EEG signal is input, and the two output heads of Label Prediction and Domain Classification are still retained) as the EEG modality baseline, and the experimental results are shown in TABLE II. When DeepSleepNet is selected as the EEG modality's baseline, our SACD exceeds other cross-modal distillation methods, including CRD and CCL, in the S^3VE dataset. Specifically, compared to the IR video modality baseline, SACD outperforms ACC, MF1 and κ by 6.1% (54.6%-60.7%), 6.8% (47.8%-54.6%) and 0.08 (0.46-0.54), respectively. In addition to this, SACD is superior to existing SOTA methods in per-class accuracy. When a single-channel version of SEN-DAL (Modified-SEN) is chosen as the baseline for our EEG modality, our SACD outperforms the baseline by 7.7% (54.9%-62.6%), 6.7% (48.9%-55.6%) and 0.08 (0.48-0.56) for ACC, MF1 and kappa, respectively. Furthermore, compared to our most competitive rival CCL, our SACD improves by 2% (60.6%-62.6%), 1.4% (54.2%-55.6%) and 0.03 (0.53-0.56) for ACC, MF1 and κ, respectively. In terms of per-class accuracy, our methodology is significantly better than the current one, except for the $N1$ stage, where our SACD is slightly lower than the CCL.

To illustrate that our proposed distillation method is not only applicable to our S^3VE dataset, we also conduct experiments on the UCF51 [34] public dataset. UCF51 is a subset of UCF101 that contains audio in videos, including 6,845 videos from 51 action classes. We use the public split 1 for evaluation. We choose the pre-trained audio encoder 1D-CNN14 [47] trained on AudioSet [48] as the teacher and freeze its parameters. The experimental results are shown in TABLE III and our SACD also obtains the state-of-the-art consistently and improves over the prior methods. Specifically, in terms of ACC, our SACD improved by 8.5% (57.5%-66.0%) over the baseline method and by 1.1% (64.9%-66.0%) over CCL. The results show that our method can also achieve superior distillation results on other datasets.

In general, the above-mentioned mainstream distillation methods we compare include single-modal distillation and cross-modal distillation, and the latter three methods also apply contrastive learning. Compared with them, we first introduce

TABLE IV
SLEEP STAGE CLASSIFICATION FOR IR+EEG. METRIC:TOP1 ACCURACY (%). *(In Experiment A, We use a single fully connected layer as the prediction head; In Experiment B, We use three fully connected layers as the prediction head. Experiment C is a version of the released gradients from Experiment A; Experiment D is a version of the released gradients from Experiment B. In experiment E, we use our SACD to distill IR video information (frozen gradients) onto the EEG baseline. In experiment F, we use our SACD to distill IR video information (released gradients) onto the EEG baseline. In experiment G, we use CCL to distill IR video information (released gradients) onto the EEG baseline)*

		Per-class Accuracy				
Methods	Accuracy	W	N1	N2	N3	REM
Baseline R3D (only IR video)	52.8	64.7	41.3	48.4	47.8	57.3
Baseline Attnsleep (only EEG)	78.8	92.8	54.1	74.2	78.7	70.4
Experiment A	79.3	93.4	54.8	74.5	78.7	70.8
Experiment B	80.4	94.0	55.5	75.3	79.4	71.8
Experiment C	84.4	96.1	59.2	80.2	84.3	76.1
Experiment D	85.5	96.9	60.5	81.8	85.6	77.4
Experiment E	80.6	94.2	55.5	75.6	79.8	72.2
Experiment F	85.3	96.7	60.1	81.5	85.3	77.4
Experiment G	83.2	95.0	58.7	79.6	82.0	74.3

structural similarity across modalities, aiming to improve the ability to deal with weak inter-class gaps while narrowing the semantic gap. Secondly, we apply two K-hard negative samples to train highly transferable sample visual representations. Experimental results demonstrate that our method significantly improves cross-modal distillation on the S^3VE dataset. It also indicates that it can do sleep stage classification with reasonable accuracy from IR sleep videos alone.

D. Evaluation of sleep stage using both the EEG and IR modality

Our starting point is to distill the knowledge of the EEG modality to the IR video, and we also verified our conjecture using experiments In section V-C. This section will analyze the complementary information in EEG and IR modalities. In other words, is it beneficial to provide better predictions if we combine information from these two modalities? As EEG and IR data are naturally synchronized in the data collection, the simplest and most effective fusion method is the late-fusion on the features that can fully consider the interactions and correlations between each modality. We use AttnSleep [12] as the EEG teacher network trained on the $C3$ (Channel 3) data in the S^3VE training set for 58 epochs. The five-class sleep stage classification using EEG data achieves the accuracy of 78.8% on the validation set. For the IR baseline, we chose R3D-18, trained for 80 epochs on our dataset S^3VE. As shown in the second row of TABLE II, the sleep stage classification accuracy for IR video is 52.8%. We first save the weights of the two networks, then perform the inference operation on the 80 subjects of the training set and save the encoded features of the EEG encoder and IR video encoder of each clip. The two model output feature dimensions must be consistent, and we perform two experiments using 512 and 3000 dimension features as the output, respectively. The experimental results show little difference between the results of the two choices.

As shown in TABLE IV, high accuracy (78.8% for overall ACC, 92.83% for stage W, 54.18% for stage $N1$, 74.16% for stage $N3$, 78.74 for stage $N3$, and 70.48% for stage REM) can already be achieved using only the EEG baseline, because EEG is the "gold-standard" in sleep stage classification. In experiment A, we concatenate the two 512-dimensional tensors saved by the EEG and IR video encoder and use a fully-connected layer with an input dimension of 1024 and an output dimension of 5 as the prediction module. We train the network for 100 epochs, and the results are in TABLE IV. It is easy to observe that the accuracy improvement of various classes is not obvious. Specifically, 0.5% (78.8%-79.3%) for overall accuracy; 0.6% (92.8%-93.4%) for stage W; 0.6% (54.2%-54.8%) for stage N1; 0.3% (74.2%-74.5%) for stage N2; 0% (78.7%-78.7%) for stage N3; 0.4% (70.4%-70.8%) for stage REM. We argue that the slight improvement in classification accuracy might be because the parameters of a single MLP classification module are not big enough to cover the dataset. Then we select more fully-connected layers for feature-level fusion. As shown in TABLE IV, in experiment B, we use three fully-connected layers as the prediction module. Their input and output dimensions are:

- Input 1024 dimensions, output 256 dimensions.
- Input 256 dimensions, output 64 dimensions.
- Input 64 dimensions, output 5 dimensions.

We use the SGD to train this prediction module for 100 epochs and the results are significantly improved compared to Experiment A. Specifically, experiment B obtains 1.3% (78.8%-80.1%) improvement in overall accuracy; 1.3% (92.8%-94.1%) for stage W; 1.4% (54.1%-55.5%) for stage N1; 1.1% (74.2%-75.3%) for stage $N2$; 1.1% (78.7%-79.8%) for stage N3; 1.3% (70.5%-71.8%) for stage REM. In both Experiments A and B, we freeze the gradients of the encoders and train only the later linear projection. In contrast, in Experiments C and D, we release the gradients of the encoders. The experimental results show that the accuracy of the sleep grading has a significant improvement of 5.6% (78.8%-84.4%) and 6.7% (78.8%-85.5%), respectively, over using only the EEG baseline AttnSleep.

To observe the performance of reverse distillation from the IR video to the EEG modality, we again design Experiment E and Experiment F, which are the frozen gradients version and the released gradients version, respectively. It can be observed that the results of reverse distillation reached 80.6% and 85.3%, which is already extremely close to the fusion version of the experiment. In addition, we also perform a complementary experiment G (the released gradients version of the CCL distillation method based on the reverse distillation from IR video to EEG modality), whose overall accuracy is 2.1% (83.2%-85.2%) lower than ours for the same setup.

From these results, we can draw the following conclusions:

- **Sleep stage classification accuracy improves by using the feature-level fusion of EEG and IR videos.**
- **Most importantly, we show that complementary information exists between the two modalities. In other words, studying the cross-modal distillation of IR and EEG modalities makes sense.**

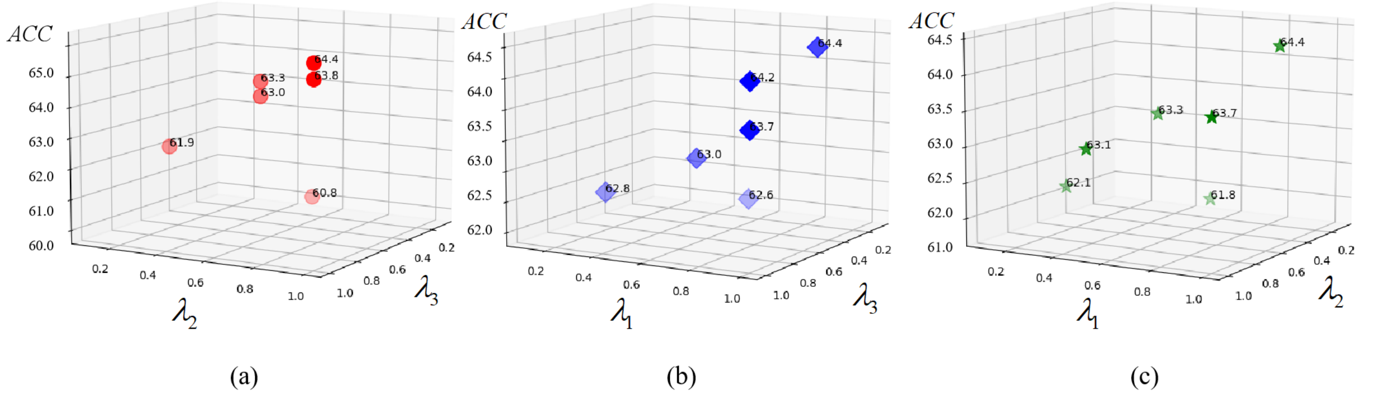

Fig. 5. Visualization of hyperparametric analysis. (a) shows the relationship between hyperparameters λ_2 and λ_3 and the overall classification accuracy for the case of fixed λ_1=0.5; (b) shows the relationship between hyperparameters λ_1 and λ_3 and the overall classification accuracy for the case of fixed λ_2=1.0; (c) shows the relationship between hyperparameters λ_1 and λ_2 and the overall classification accuracy for the case of fixed λ_3=1.0.

TABLE V
THE COMPLETE ABLATION ON LOSS FORMULATION. METRIC:TOP1 ACCURACY (%).

			Per-class Accuracy				
Methods	Module	Accuracy	W	N1	N2	N3	REM
baseline(v_0)	L_{CE}	52.2	64.7	40.3	48.4	47.8	57.3
v_1	L_{CE}+L_C+L_{JSD}	56.1	67.7	43.8	52.0	50.4	59.2
v_2	L_{CE}+L_D+L_C	62.3	73.4	49.0	58.0	57.2	65.8
v_3	L_{CE}+L_D+L_{JSD}	60.1	71.5	45.6	55.4	54.5	63.6
v_4	L_{CE}+L_D+L_{JSD}+$L_{C'}$	61.1	72.4	46.8	56.7	55.8	64.6
ours	L_{CE}+L_D+L_C+L_{JSD}	**64.4**	**75.6**	**51.0**	**60.2**	**59.3**	**67.8**

- **If the cross-modal distillation method is good enough, the accuracy of distillation from IR video to EEG modality can be achieved with essentially the same accuracy as the feature fusion version.**

VI. ANALYSIS

A. Ablation Analysis

To demonstrate that each loss function we use is necessary, we remove a particular loss function item by item and observe the change in overall accuracy and per-class accuracy.

As shown in TABLE V, the specific experimental setup is as follows. Experiment v_0 removes our proposed L_C, L_D and L_{JSD} losses, keeping only the cross-entropy loss L_{CE}; Experiment v_1 removes the L_D loss compared to the complete SACD; Experiment v_2 removes the L_{JSD} loss compared to the complete SACD; Experiment v_3 removes the two-way K-hard negative loss L_C compared to the complete SACD; It is worth noting that the $L_{C'}$ in Experiment v_4 represents only a single direction of L_C, and this set of experiment is also designed to verify the necessity of the two way K-hard Negative loss. TABLE V shows that Experiment v_2 performed significantly worse than the complete SACD, with 2.1% (64.4%-62.3%) worse in overall accuracy and 2.2% (75.6%-73.4%), 2.0% (51.0%-49.0%), 2.2% (60.2%-58.0%), 2.1% (59.3%-57.2%) and 2.0% (67.8%-65.8%). This suggests that loss L_{JSD} is effective and that the differences between categories on L_{JSD} are not significant. Next, we compare the SACD with two ablation baselines. Experiment v_4 and Experiment v_3, they are each removed one constraint at a time (one-way K-hard Negative loss). It is easy to observe that the first K-hard Negative contrastive learning significantly impacts classification accuracy. Taking overall accuracy as an example, when we progressively remove one-way K-hard Negative loss and two-way K-hard Negative loss, the accuracy degrades 3.3% (64.4%-61.1%) and 4.3% (64.4%-60.1%), respectively. It is worth noting that this loss function L_C seems to be very sensitive to different classes as its effects on the $N1$, $N2$, and $N3$ stages are more significant than on the W and REM stages. This confirms the benefit of our structure-aware coarse-grained semantic alignment, and contrastive learning distills more about $N1$, $N2$, and $N3$ stages information from EEG to IR video modality. As shown in experiment v_1 in TABLE V, removing the L_D loss function reduces the overall accuracy of SACD by 8.3% (64.4% - 56.1%); the reduction in per-class accuracy is similar to the decrease in overall accuracy (7.2% to 8.9% per class). These results indicate that L_D, L_C, and L_{JSD} are complementary, and they work synergistically to distill knowledge across modalities.

B. Quantitative Analysis

1) Hyperparameter Analysis: As described in section V-B, our hyperparameters (λ_1, λ_2 and λ_3) are selected as (0.5, 1, 1), the optimal ratio, reaching a overall classification accuracy of 64.4%, λ_1, λ_2 and λ_3 are the coefficients in front of the loss function L_D, L_C and L_{JSD} , respectively. It is what we have obtained through a lot of comparative experiments. Next, we will give more experimental results about hyperparameters. Based on keeping other experimental settings constant, We set the coefficient of lossL_{CE} to always be only changed the proportion of (λ_1, λ_2 and λ_3).

To visualize the effect of different hyperparameter ratios on the overall classification performance, as shown in Fig. 5, we plot three figures, each fixing one parameter, to observe the relationship between the change in the other two parameters and the overall accuracy.

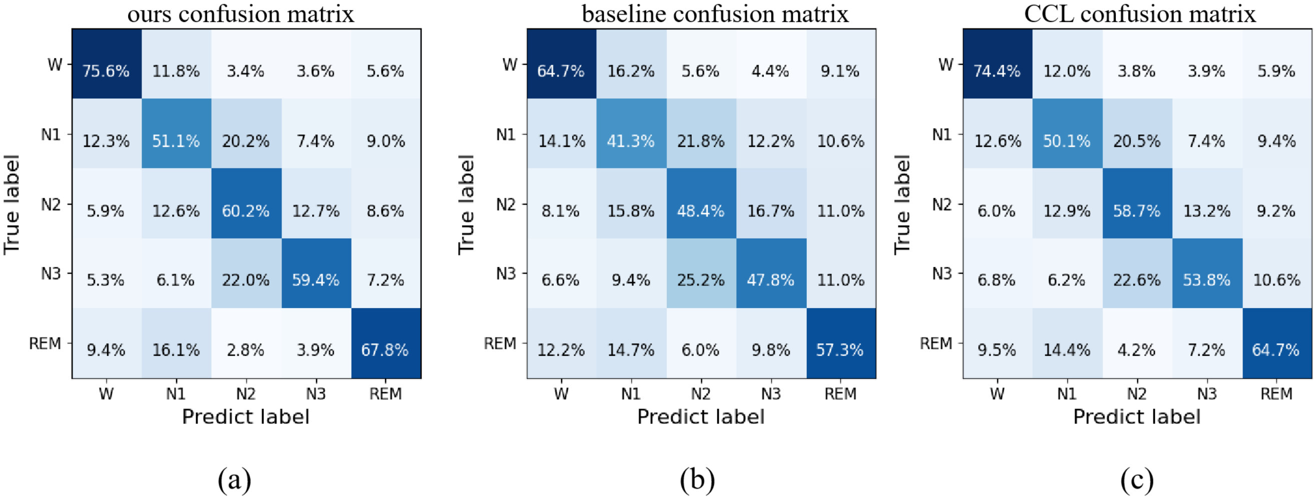

Fig. 6. Normalized confusion matrices of the classification accuracies from our dataset S^3VE. (a). confusion matrix for our method SACD. (b). confusion matrix for baseline R(2+1)D. (c). confusion matrix for CCL [31].

Observing Fig. 5 (a), we can see that when we fix hyperparameter λ_1 to 0.5 and change the value of (λ_2, λ_3) to (0.5, 1.0), the overall accuracy reach the lowest value of 60.8% in this group. When (λ_2=1.0, λ_3=1.0), the accuracy get the highest value of 64.4% in this group of experiments, which shows that when hyperparameter λ_1 is fixed, increasing appropriate hyperparameters λ_2 and λ_3 will improve the contribution to the overall classification accuracy. However, when the two values are rapidly increased to 10, the classification accuracy rapidly decreases to 52%, which is not drawn due to the view scale, thus showing the importance of the appropriate hyperparameters.

Similar to the first set of experiments, we fix hyperparameter λ_2 to 1.0 and vary different values (λ_1, λ_3) to observe the change in overall accuracy. As shown in Fig. 5 (b), when (λ_1=0.5, λ_3=0.1), the overall classification accuracy come to 62.6%, which is the lowest value in this group of experiments; when (λ_2=0.5,λ_2=1.0), the overall classification accuracy increase to 64.4%, which is the highest value in this group of experiments. When the value of (λ_1, λ_3) is changed from (0.5, 1.0) to (1, 1), the accuracy decrease by 0.02% (64.4%-64.2%), which also illustrates that the value of hyperparameter λ_1 cannot be increased without a limit. In Fig. 5 (c), we fix the hyperparameter λ_3= 1 and vary the values of hyperparameter λ_1 and hyperparameter λ_2. When the value of (λ_1, λ_2) is (0.5, 0.1), the accuracy comes to 61.8%, which is the lowest value in this set of experiments, and when the value of (λ_1, λ_2) is (0.5, 1.0), the accuracy rises to 64.4% which is the highest value in this set of experiments.

2) Confusion Matrix Analysis: Fig. 6 shows the results of the confusion matrix predicted by the model. From left to right are (a): our SACD confusion matrix, (b): the IR baseline confusion matrix (c): the CCL confusion matrix. As can be seen with the confusion matrices, the error rate of $N1$ is the highest among all three methods. The poor performance of $N1$ can be attributed to the fact that many samples in $N1$ stage are misclassified as W and $N2$ stages, since most samples in the $N1$ stage belong to the sleep transition period [49] [46]. In addition, it is also possible that the number of data in stage $N1$ itself is relatively small compared to the other sleep stages. Also, observing Fig. 6 (a), it can be found that, except for REM, adjacent sleep stages generally tend to have higher confusion rates than other non-adjacent stages. In contrast, REM is more likely to be misclassified as a $W1$ stage. We guess the possible reason is that $N1$ is a light sleep in the immediate sleep stage, and there may be some small erratic movements, such as manual and eye movements. However, in REM, most dreams occur, and there are obvious rapid eye movements. These similar features in the IR video could lead to the occurrence of misclassification. Comparing Fig. 6 (a) and Fig. 6 (c), we can see that we have a clear advantage over CCL [31] in distinguishing the two sleep stages, REM and $N3$, which will also be mentioned in the visualization analysis section.

3) Statistical Correlation Analysis: To observe the correlation between the model output and the AHI (interest to sleep apnea physicians), a total of 102 data from the training and test sets were subjected to statistical analysis. The details are as follows: averaging all five sleep stage features output for each individual throughout a night according to the ground truth labels. The averaged results represent the "average condition" of the person's five sleep stages, all of which have a dimension of 512 vectors. The AHI was divided into four levels (Normal: <5, Mild: 5-15, Moderate: 16-30, Severe: >30) to create an AHI feature distribution, which was adjusted into four one-hot vectors. Fig. 7 (a) is the Spearman correlation between the "average condition" of each individual's five sleep stages and the overall AHI distribution; Fig. 7 (b)-(e) is the Spearman correlations between the "average condition" of each individual's five sleep stages and the one-hot vector of each AHI level. The p-value values are all less than 0.05, indicating our statistical results are confident.

As shown in Fig. 7 (a), we can observe that the correlations for $N2$ and $N3$ are significantly higher than those for W,

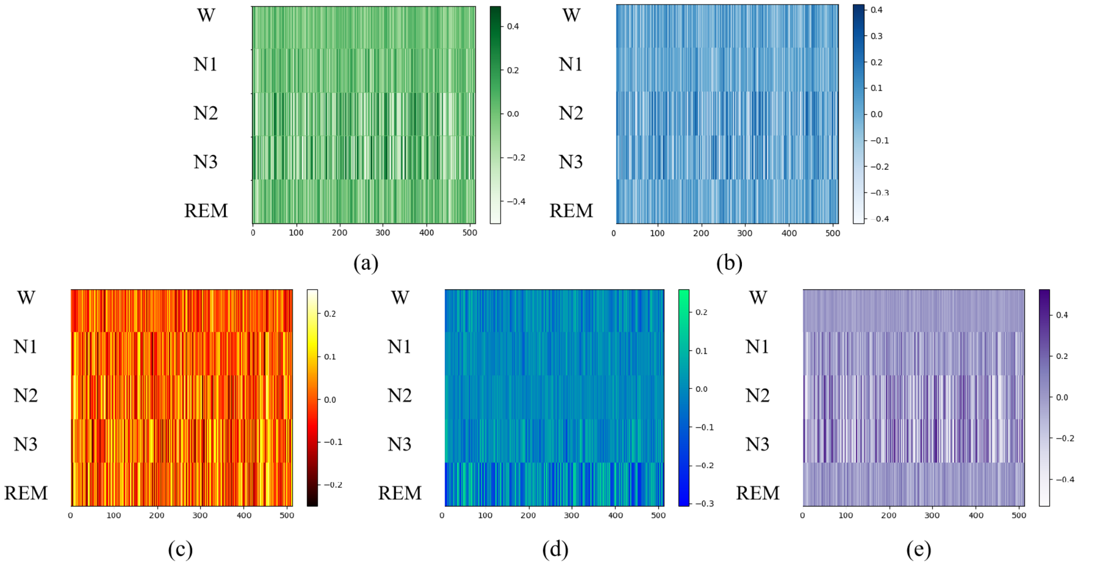

Fig. 7. Spearman correlation analysis of SACD's output and AHI. (a) is the Spearman correlation between the "average condition" of each individual's five sleep stages and the overall AHI distribution; (b)-(e) is the Spearman correlations between the "average condition" of each individual's five sleep stages and the one-hot vector of each AHI class.

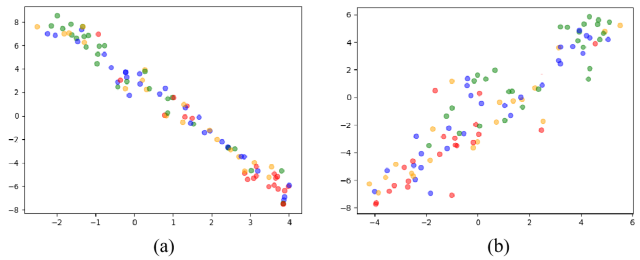

Fig. 8. Selective Tsne visualisation of all samples $N2$ and $N3$. Output channels in $N2$ (Fig. 8 (a)) and $N3$ (Fig. 8 (b)) with AHI speatman correlation coefficients larger than 0.4, green for AHI Normal (AHI <5) , blue for Mild (5<=AHI<15), orange for Moderate (15<=AHI<=30), and red for Severe (AHI >30).

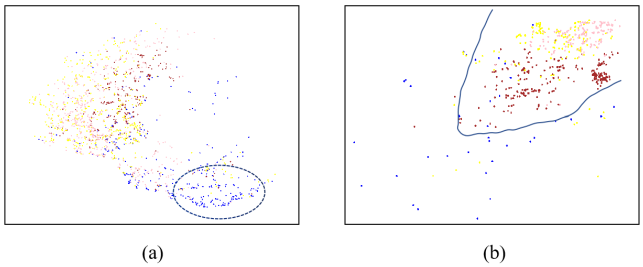

Fig. 9. Tsne visualization of four sample $N2$ and $N3$ clips features. The different colors represent the samples' sleep $N2$ (Fig. 9 (a)) or $N3$ (Fig. 9 (b)) clips; Blue represents sample A; Pink represents sample B; Yellow represents sample C; Brown represents sample D; their basic information is shown in the TABLE VI, and we have also marked the "danger zones" where the AHI is larger.

$N1$, and REM. Consistent with our intuitive sense, there should indeed be a more negligible correlation between the characteristics of W and AHI, as the characteristics of a person while awake are inherently less relevant to sleep apnea events. As shown in Fig. 7 (b) and Fig. 7 (d), similar to the previous analysis, $N2$ and $N3$ are significantly more correlated than W, $N1$, and REM, suggesting that $N2$ and $N3$ are more closely related to whether they are Normal (AHI <5) or Sever (AHI >30). This finding is helpful for clinical medicine, where sleep apnea physicians can look primarily at these two sleep stages. As shown in Fig. 7 (c), the $N2$, $N3$, and REM correlations are more similar than the W and $N1$ stages. As shown in Fig. 7 (d), the correlation between REM is higher than that of the other sleep stages. This finding is inconsistent with our intuition and worthy of further exploration.

We visualize the output features from the following two perspectives to further visualize the correlation between the model output features and the AHI. Fig. 8: Tsne [50] visualization of the high weight channel for the average feature of all samples $N2$ and $N3$, respectively, and Fig. 9: tsne visualization of the $N2$ and $N3$ clips (30s) features for the four samples. As shown in Fig. 8, we make the selective output of the $N2$ and $N3$ feature channels in Fig. 7 (a) that are highly correlated with AHI. It is done by selecting all channels with a spearman correlation greater than 0.4, which is also a selective dimensionality reduction process. For $N2$, there are 17 such channels; for $N3$, there are 36 such channels; finally, Visualization by Tsne, respectively. The different colors in Fig. 8 indicate different AHI levels, and we can observe that in (a), the normal samples (green dots) are primarily located at the top left of the figure, and the severe samples (red dots) are primarily located at the bottom right of the figure. It can be more clearly seen that certain features of the $N2$ output are correlated with the AHI. Similarly, in Fig. 8 (b), the most severe samples are located at the bottom left of the figure, and the most normal samples are located at the top right of the figure, which leads to a similar conclusion.

TABLE VI
MESSAGE OF SAMPLES ABCD ; TST : TOTAL SLEEP TIME.
(NORMAL: <5, MILD : 5-15, MODERATE: 15-30, SEVERE: >30)

Index	Data	Gender	Age	Weight	Sleep-related apnea-hypopnea Index (AHI)
A	20210310	Male	37	108kg	94.2 TST
B	20200117	Male	31	66kg	0.4 TST
C	20210326	Female	58	58kg	6.9 TST
D	20210329	Female	65	64kg	26.2 TST

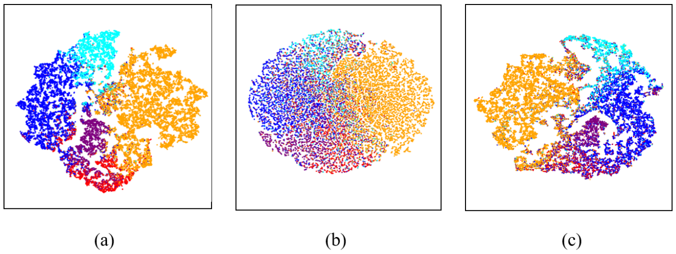

Fig. 10. Visualisation with Tsne [50], (a) is the feature embedding extracted from S^3VE test set using our method. (b) is the feature embedding extracted from S^3VE using baseline Res(2+1)D. (c) is the feature embedding extracted from S^3VE using CCL [31] (the orange represents W; cyan represents $N1$; blue represents $N2$; red represents $N3$; purple represents REM).

While the Fig. 7 visualizations are performed on the average features of all samples in the sleep stages, we next chose four more representative samples and do Tsne [50] visualizations of all their $N2$ and $N3$ clips. As shown in Fig. 9, unlike in Fig. 8: the different colors in the figure represent different samples, and each dot represents the 512-dimensional feature distribution of a single clip for one sample. We also give the basic information of the four individuals as a reference, as shown in the TABLE VI. In Fig. 9 (a), it can be observed that the larger the AHI, the more clips of the sample are distributed in the lower right of the plot, especially the blue dots, with a high AHI of 94.2 TST. A similar conclusion can be drawn in Fig. 9 (b), where the larger the AHI, the more clips of the sample are distributed in the lower left of the plot. These two **“danger zones”** (areas of larger AHI) are drawn in the diagram. Given a new subject, we can do a **“feature portrait”** in the Tsne graph to get a rough guess of which AHI levels it belongs to. We have discussed with sleep apnea physicians the feasibility of this kind of **“feature portrait”** in practical clinical application and have validated it on a number of new examples.

C. Qualitative Results

1) Visualisation Analysis: To qualitatively understand video representation and cross distillation, we analyze the SACD with visualizations. When using Tsne [50] to visualize the 512-dimensional embeddings of the different sleep stages, we can see that embeddings of different categories are grouped into different clusters. The distances between dots in Fig. 10 represent the distances of embeddings in the higher dimensional space, and one can observe that the classification results make the distances between identical classes (sleep stages) closer and the distances between different classes (sleep stages) farther, thus showing the success of our method. Compared with the aggregation cluster of the baseline (Fig. 10 (b)) samples, the cluster of each class of our SACD (Fig. 10 (a)) is significantly more concentrated, which illustrates that our method SACD's embeddings from the same class have a higher similarity. Compared to the aggregated clusters of CCL [31] (Fig. 10 (c)), there are significantly fewer overlapping dots of different colors, implying that our embeddings are more accurate; specifically, the purple (REM) and red ($N2$) parts of Fig. 10 (c) are more confounded, whereas the species in Fig. 10 (a) are significantly more clearly separated, indicating that our method classifies better in REM and $N3$ and this conclusion is corroborated by the section VI-B2 confusion matrix analysis. In summary, the qualitative results show that our SACD can learn discriminative video representations of cross-modal distillation from EEG modality and that the distillation performance is better than current state-of-the-art distillation methods.

D. Qualitative Analysis on Cross-Modal Correspondence

To understand the cross-modal semantic correspondence clearly, we provide examples of video-EEG correspondence in Figure 11. We select five videos of five sleep stages for analysis and gave the EEG waveforms of these five videos to observe together. The five sleep infrared videos are misclassified using baseline R(2+1)D, but after our SACD distillation, the results are correct. First, looking at the EEG waveform figure, we can see that in Fig. 11 (a), the EEG waveform is a $beta$ wave, which usually appears in W [51]. In Fig. 11 (c), we can see the $delta$ wave, which usually appears in $N2$ [51], and in Fig. 11 (e), the we can see the $theta$ wave, which usually appears in REM [51]. This waveform information is easily determined by observing the EEG, indicating that the EEG is the discernible modality with a prior knowledge for these examples. And these information cannot be used without the cross-modal distillation, resulting in “prediction drift” that is prone to occur without distillation, such as baseline R(2+1)D, and the results after distillation are all correct. This demonstrates the usefulness of the EEG modality for our distillation task and the effectiveness of our method.

In addition to these correct examples, we also find some unsolvable examples in the experiment. As shown in Fig. 12 (a), both the baseline and our method have been misjudged. We guess that the reason may be that these videos may be at the critical point of sleep stage change. This speculation has been mentioned in a number of previous studies [49] [46]. It is reflected at the embedding level that the clustering of feature embeddings may not be close enough. Our other speculation is that the error is caused by the PSG labeling process: the PSG gives a unique sleep classification for each clip (30s) based on a combination of physiological signals, with the potential problem that if the clips are partly in one sleep stage and partly in another, then the PSG will give the final label based on their proportion of the sleep stage. This indirectly leads to less accurate labeling, especially when the sleep stage is switched, and is more likely to be misclassified. However, this problem is inherent to PSG and has nothing to do with our algorithm, and we will find ways to correct this

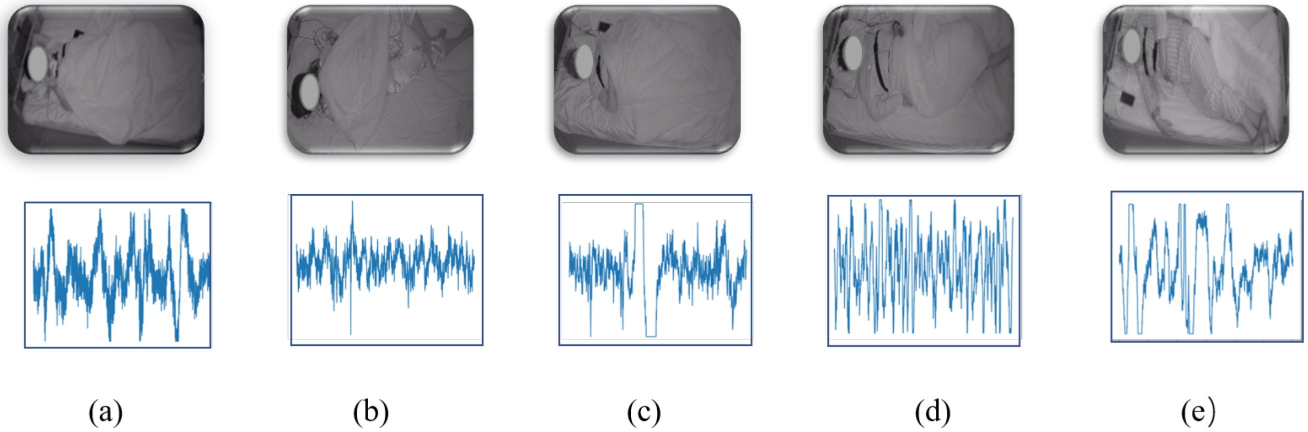

Fig. 11. IR video-EEG correspondence:visual representation of five sleep stages and their EEG waves. In these examples, our predictions are correct and the baseline predictions are wrong. (a) a clip labeled by W . (b) a clip labeled by W . (c) a clip labeled by $N2$. (d) a clip labeled by $N3$. (e) a clip labeled by REM.

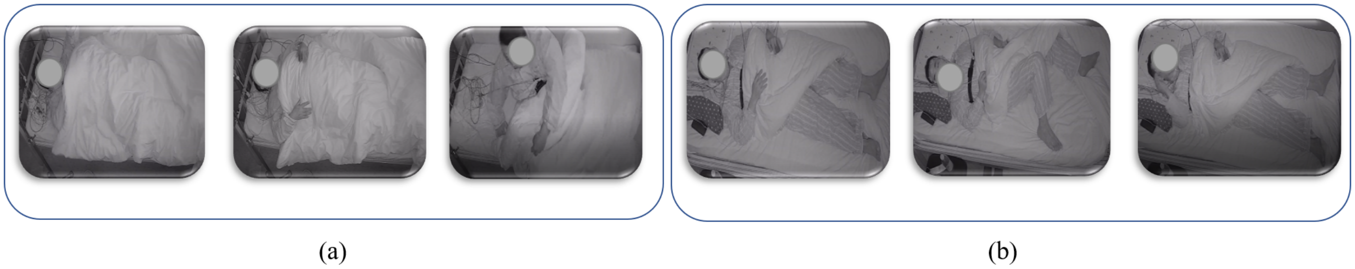

Fig. 12. Bad case examples: (a) example of sleep stage junction, previous clip is $N3$ and next clip is W. (b) Instance-level specificity, this instance's movement changes more frequently than normal instance

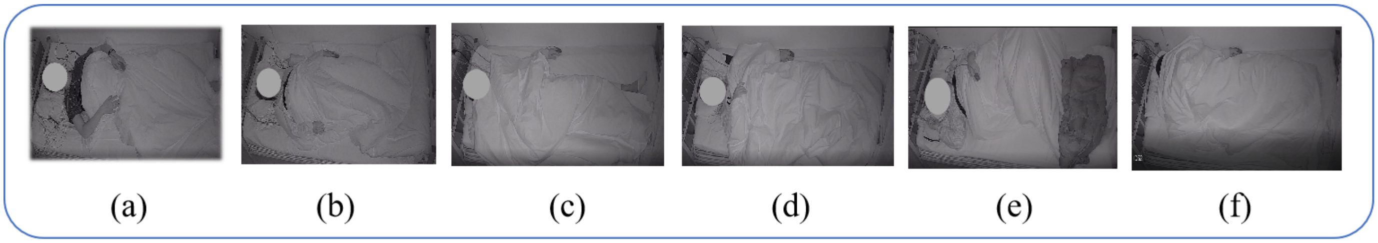

Fig. 13. Analysis of the main error types. (a) The most common error in the stage W, wrongly evaluated as stage $N1$; (b) The most common error in the stage $N1$, wrongly evaluated as stage $N2$; (c) The most common error in the stage $N2$, wrongly evaluated as stage $N1$; (d) The most common error in the stage $N3$, wrongly evaluated as stage $N2$; (e) coexisting errors;

problem in future research. In Fig. 12 (b), this is a special instance that frequently moved throughout sleep. It gives us the heuristic that compares to class-level. We should probably pay more attention to the instance level when faced with these special cases. In addition, sleep is a long-term problem; both our method SACD and baseline do not take advantage of the influence of time correlation, the correlation on the time axis may help us a lot in S^3VE, which is also our direction for future work.

We produce the confusion matrice in VI-B2 for our method SACD on the test set of S^3VE to observe the most dominant error types for each sleep stage. According to the confusion matrix, we list the main error types corresponding to the five stages, as shown in Fig. 13 . About Fig. 13 (a), in stage W, the most dominant error is being misclassified as $N1$; We believe that there may be some sleep-onset phases at the end of W and the beginning of $N1$, where physical representations are remarkably similar and complex to distinguish. About Fig. 13 (b), in stage $N1$, the most dominant error is being misclassified as $N2$; About Fig. 13 (c), in stage $N2$, the most significant errors are misclassified as $N1$ and $N3$, and here we have chosen the example of misclassified N1 for illustration; It may be that the proportion of N1 instances is relatively tiny, and it is difficult to make better use of contrastive learning to widen the gap between $N2$ and $N1$. This is a problem of sample imbalance. Nevertheless, we expect to continue to collect some $N1$ data or use data enhancement methods in the future. About Fig. 13 (d), in stage $N3$, the most dominant error is being misclassified as $N2$; Similar to Fig. 13 (a), this error belongs to the problem that the intersection time is difficult to define. About Fig. 13 (e), in stage REM, the most dominant error is being misclassified as $N1$; This error is more interesting. N1 belongs to the light sleep stage, which is more into sleep. Some samples will show that the body and expression are not completely relaxed, which is similar to the rapid eye movement in REM stage and the body movements during dreaming—leading to misjudgment.

About Fig. 13 (f), this is a possible problem in any sleep stage because the samples sometimes turn their heads to the side of the wall, and the infrared camera cannot capture the face; sometimes, the samples even cover their heads with quilts. We are also unable to obtain more information; the solution is to install an infrared camera on the edge just above the wall to get infrared video information from multiple viewing angles.

VII. Conclusions

In this paper, we propose a novel cross-modal distillation dataset and benchmark for the multi-modal community. This dataset bridges the gap between the clinical and the visual modality and promotes the developments of the point-of-care research. Besides the contribution of the dataset, we also present a novel cross-modal distillation method that can effectively reduce the cross-modal gap and facilitate the usage of visual modality to classify the sleep stage. Experimental results show that our method outperforms other SOTA methods in both our dataset and the other benchmarks. With IR video alone, the proposed method can also achieve considerable sleep stage classification performance. We expect the proposed method to be applied in a wider range of scenarios and hope that more researchers will pay attention to the infrared sleep video modality.

References

[1] R. de Goederen, S. Pu, M. S. Viu, D. Doan, S. Overeem, W. A. Serdijn, K. F. Joosten, X. Long, and J. Dudink, "Radar-based sleep stage classification in children undergoing polysomnography: a pilot-study," *Sleep Medicine*, vol. 82, pp. 1–8, 2021.

[2] B. Deng, B. Xue, H. Hong, C. Fu, X. Zhu, and Z. Wang, "Decision tree based sleep stage estimation from nocturnal audio signals," in *2017 22nd International Conference on Digital Signal Processing (DSP)*. IEEE, 2017, pp. 1–4.

[3] H. Korkalainen, J. Aakko, B. Duce, S. Kainulainen, A. Leino, S. Nikkonen, I. O. Afara, S. Myllymaa, J. Töyräs, and T. Leppänen, "Deep learning enables sleep staging from photoplethysmogram for patients with suspected sleep apnea," *Sleep*, vol. 43, no. 11, p. zsaa098, 2020.

[4] R. Yi, M. Enayati, J. M. Keller, M. Popescu, and M. Skubic, "Non-invasive in-home sleep stage classification using a ballistocardiography bed sensor," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2019, pp. 1–4.

[5] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet *et al.*, "The sleep heart health study: design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[6] G.-Q. Zhang, L. Cui, R. Mueller, S. Tao, M. Kim, M. Rueschman, S. Mariani, D. Mobley, and S. Redline, "The national sleep research resource: towards a sleep data commons," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1351–1358, 2018.

[7] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[8] 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," *Journal of sleep research*, vol. 23, no. 6, pp. 628–635, 2014.

[9] L. Fraiwan, K. Lweesy, N. Khasawneh, M. Fraiwan, H. Wenz, and H. Dickhaus, "Classification of sleep stages using multi-wavelet time frequency entropy and lda," *Methods of information in Medicine*, vol. 49, no. 03, pp. 230–237, 2010.

[10] D. Jiang, Y.-n. Lu, M. Yu, and W. Yuanyuan, "Robust sleep stage classification with single-channel eeg signals using multimodal decomposition and hmm-based refinement," *Expert Systems with Applications*, vol. 121, pp. 188–203, 2019.

[11] S.-F. Liang, C.-E. Kuo, Y.-H. Hu, Y.-H. Pan, and Y.-H. Wang, "Automatic stage scoring of single-channel sleep eeg by using multiscale entropy and autoregressive models," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 6, pp. 1649–1657, 2012.

[12] E. Eldele, Z. Chen, C. Liu, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "An attention-based deep learning approach for sleep stage classification with single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 809–818, 2021.

[13] Z.-y. JIA, Y.-f. LIN, H.-j. ZHANG, and J. Wang, "Sleep stage classification model based ondeep convolutional neural network," *Journal of ZheJiang University (Engineering Science)*, vol. 54, no. 10, pp. 1899–1905, 2020.

[14] X. Cai, Z. Jia, M. Tang, and G. Zheng, "Brainsleepnet: Learning multivariate eeg representation for automatic sleep staging," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 976–979.

[15] Z. Jia, Y. Lin, J. Wang, X. Ning, Y. He, R. Zhou, Y. Zhou, and H. L. Li-wei, "Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1977–1986, 2021.

[16] Z. Jia, Y. Lin, J. Wang, R. Zhou, X. Ning, Y. He, and Y. Zhao, "Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification." in *IJCAI*, 2020, pp. 1324–1330.

[17] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Joint classification and prediction cnn framework for automatic sleep stage classification," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 5, pp. 1285–1296, 2018.

[18] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[19] J. Ba and R. Caruana, "Do deep nets really need to be deep?" *Advances in neural information processing systems*, vol. 27, 2014.

[20] C. Bucila, R. Caruana, and A. Niculescu-Mizil, "Model compression," 2006.

[21] R. Girdhar, D. Tran, L. Torresani, and D. Ramanan, "Distinit: Learning video representations without a single labeled video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 852–861.

[22] N. Komodakis and S. Zagoruyko, "Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer," in *ICLR*, 2017.

[23] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4133–4141.

[24] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[25] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li, "Learning from noisy labels with distillation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1910–1918.

[26] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1365–1374.

[27] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2604–2613.

[28] J. Zhu, S. Tang, D. Chen, S. Yu, Y. Liu, M. Rong, A. Yang, and X. Wang, "Complementary relation contrastive distillation," in *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9260–9269.

[29] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European conference on computer vision*. Springer, 2020, pp. 776–794.

[30] ——, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.

[31] Y. Chen, Y. Xian, A. Koepke, Y. Shan, and Z. Akata, "Distilling audio-visual knowledge by compositional contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7016–7025.

[32] R. B. Berry, R. Brooks, C. E. Gamaldo, S. M. Harding, C. Marcus, B. V. Vaughn *et al.*, "The aasm manual for the scoring of sleep and associated events," *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine*, vol. 176, p. 2012, 2012.

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

[34] K. Soomro, A. R. Zamir, and M. Shah, "A dataset of 101 human action classes from videos in the wild," *Center for Research in Computer Vision*, vol. 2, no. 11, 2012.

[35] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[36] S. Wang, L. Yu, C. Li, C.-W. Fu, and P.-A. Heng, "Learning from extrinsic and intrinsic supervisions for domain generalization," in *European Conference on Computer Vision*. Springer, 2020, pp. 159–176.

[37] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[38] K. Hara, H. Kataoka, and Y. Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 3154–3160.

[39] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987*, 2019.

[40] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick *et al.*, "Moments in time dataset: one million videos for event understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 2, pp. 502–508, 2019.

[41] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[42] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[43] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 268–284.

[44] N. Passalis, M. Tzelepi, and A. Tefas, "Heterogeneous knowledge distillation using information flow modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2339–2348.

[45] A. Supratak, H. Dong, C. Wu, and Y. Guo, "Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 11, pp. 1998–2008, 2017.

[46] Z. Jia, X. Cai, and Z. Jiao, "Multi-modal physiological signals based squeeze-and-excitation network with domain adversarial learning for sleep staging," *IEEE Sensors Journal*, vol. 22, no. 4, pp. 3464–3471, 2022.

[47] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[48] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[49] X. Chen, J. He, X. Wu, W. Yan, and W. Wei, "Sleep staging by bidirectional long short-term memory convolution neural network," *Future Generation Computer Systems*, vol. 109, pp. 188–196, 2020.

[50] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[51] S. C. C. O. T. J. S. O. S. R. S. (JSSR):, T. Hori, Y. Sugita, E. Koga, S. Shirakawa, K. Inoue, S. Uchida, H. Kuwahara, M. Kousaka, T. Kobayashi *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 and clinical neurosciences*, vol. 55, no. 3, pp. 305–310, 2001.