SOUND-BASED SLEEP STAGING BY EXPLOITING REAL-WORLD UNLABELED DATA

JongMok Kim[1,2] **Daewoo Kim**[1] **Eunsung Cho**[1] **Hai Hong Tran**[1] **Joonki Hong**[1]
Dongheon Lee[1] **JungKyung Hong**[1] **In-Young Yoon**[4,5] **Jeong-Whun Kim**[4,5]
Hyeryung Jang[3] **Nojun Kwak**[2*]

[1]Asleep Inc. [2]Seoul National University [3]Dongguk University
[4]Seoul National University College of Medicine [5]Seoul National University Bundang Hospital

ABSTRACT

With a growing interest in sleep monitoring at home, sound-based sleep staging with deep learning has emerged as a potential solution. However, collecting labeled data is restrictive in home environments due to the inconvenience of installing medical equipment at home. To handle this, we propose novel training approaches using accessible real-world sleep sound data. Our key contributions include a new semi-supervised learning technique called ***sequential consistency loss*** that considers the time-series nature of sleep sound and a ***semi-supervised contrastive learning*** method which handles out-of-distribution data in unlabeled home recordings. Our model was evaluated on various datasets including a labeled home sleep sound dataset and the public PSG-Audio dataset, demonstrating the robustness and generalizability of our model across real-world scenarios.

1 INTRODUCTION

Sleep is an essential factor for human health and well-being, and an in-depth comprehension of sleep patterns and sleep stages is imperative for diagnosing and treating sleep disorders. There exists a gold-standard polysomnography (PSG) for sleep assessment in hospitals (Bloch, 1997), but it is inconvenient and costly, making sound-based solutions at home more preferable (Hong et al., 2022). However, sleep stage annotations at home are limited because the PSG is mostly conducted in hospitals. Moreover, the sound-based model validated in hospitals is not guaranteed to perform the same at home because of various types of noise from residential environments. Therefore, to enhance the representational power of sleep sound data recorded at home, it is crucial to make better use of unlabeled home recordings by such techniques as semi-supervised learning (SSL).

Sound-based sleep staging involves identifying patterns in respiratory and body movement sounds. The pattern characteristics are reflected over a long period of time, making it difficult to fully comprehend a sleep stage from just a single snapshot of sound. Thus, modeling sleep sound necessitates a comprehensive understanding of the time-series nature of sleep sound, presenting a unique challenge in applying semi-supervised learning for time-series data. Additionally, the use of unlabeled home recordings can result in the inclusion of out-of-distribution data (i.e. absence of a person, two people in bed, and playing music), as they are self-conducted by participants using their own smartphones in real-world environments without quality control measures.

In this work, we present a novel approach to address the unique challenges of modeling sleep sound data captured in a real-world setting. Our key contributions include a new SSL technique, called ***sequential consistency loss***, which enhances the temporal correlation of the model when handling time-series sleep sound data. In addition, we propose a ***semi-supervised contrastive learning*** (SSCL) method to handle real-world data, which contains out-of-distribution (OOD) data. We thoroughly evaluate our approach on two datasets: a labeled home sleep sound dataset recorded using portable PSG devices with 45 participants, and the PSG-Audio dataset (Korompili et al., 2021), which includes a diverse range of participants and environments. Our results demonstrate the robustness and generalizability of our model across various real-world scenarios.

*Correspondence to: Nojun Kwak (nojunk@snu.ac.kr)

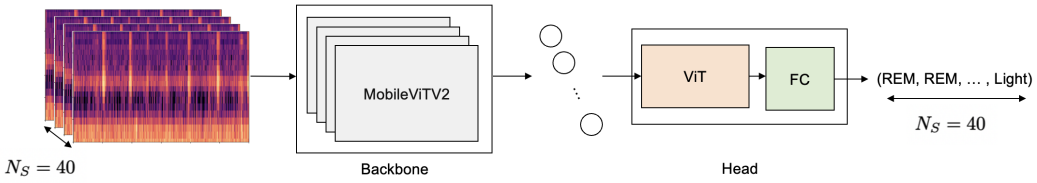

Figure 1: Framework of our SleepFormer model with N_S sequence of Mel spectrograms input.

Related work. Consistency-based SSL approach seeks to reduce the prediction discrepancy between various augmentations of a single data sample (Berthelot et al., 2019; Tarvainen & Valpola, 2017). Kim et al. (2020) introduced a structured consistency loss that considers the spatial correlation between pixels in the segmentation task. Our methodology extends this idea by proposing a novel consistency loss that accounts for the time-series properties of sleep data.

Contrastive learning is a technique for learning representations by maximizing the similarity between positive samples and minimizing the similarity between negative samples (Chen et al., 2020; He et al., 2020; Khosla et al., 2020). Yang et al. (2022) utilized high-confidence unlabeled samples for contrastive learning to address Out-of-Distribution (OOD) samples in a robust SSL scenario. Our approach incorporates labeled samples as a reliable reference point since unlabeled samples may contain a considerable proportion of OOD data.

2 METHOD

Preliminary. The sound-based sleep staging task which requires temporal analysis can be framed as a sequence prediction task, where a sequence of Mel spectrogram samples, denoted by $\mathbf{x} = (x_i)_{i=1}^{N_s}$, is used as an input to predict a corresponding sequence of sleep stage labels, $\mathbf{y} = (y_i)_{i=1}^{N_s}$, where x_i and y_i are the i^{th} sample of the sequences $\mathbf{x}$ and $\mathbf{y}$ respectively, and N_s is the number of samples in the sequence. The sleep stage labels, y_i, have four possible classes (Wake, REM, Light, Deep) represented as one-hot labels. The previous work (Hong et al. (2022)) tackles this problem by building the sequence-to-sequence SoundSleepNet model, which consists of a backbone for low-level feature extraction and a head for learning the temporal correlation between the Mel spectrograms. Built upon it, we have replaced the backbone and head networks with MobileVitV2 (Mehta & Rastegari (2022)) and ViT (Dosovitskiy et al. (2020)) respectively as shown in Fig.1, which we call SleepFormer. The model generates predictions for the sleep stage logits of the sequence $\hat{\mathbf{y}} = (\hat{y}_i)_{i=1}^{N_s}$, and the supervised baseline is trained exclusively using the cross-entropy loss ($\mathcal{L}_{\text{SUP}}$).

2.1 HARNESSING TIME-SERIES REAL-WORLD UNLABELED DATA

To achieve high performance in home environments, using unlabeled data ($\mathbf{u} = (u_i)_{i=1}^{N_s}$) from home settings is crucial. This section outlines our semi-supervised method for handling real-world sleep data, which exhibits both time-series and noisy characteristics.

Sequential Consistency Loss. To employ the consistency training, we create two different augmented samples (u_i^a, u_i^b) from u_i, and the model outputs the corresponding logits $(\hat{y}_i^a, \hat{y}_i^b)$. We first take sample-wise consistency loss $\mathcal{L}_{\text{C}} = \sum_{i=1}^{B_u N_s} JS(\hat{y}_i^a, \hat{y}_i^b)$, where Jensen-Shannon divergence is used and B_u is the batch size of unlabeled sequences. The consistency loss makes the model more generalized for sample-wise prediction, but it can not exploit the temporal correlation in sleep sound data and their corresponding labels. Therefore, we propose sequential consistency loss $\mathcal{L}_{\text{SC}}$ that matches the similarity of the prediction sequence as follows:

$$\mathcal{L}_{\text{SC}} = \sum_{s=1}^{B_u} (\mathbf{C}_s^a - \mathbf{C}_s^b)^{\circ 2} \odot \mathbf{W}, \tag{1}$$

where $\circ$ is Hadamard power and $\odot$ is the element-wise multiplication of the two matrices which are then averaged into a value. We adopt cosine similarity between the logits of i^{th} and j^{th} samples in a sequence to catch the amount of sleep stage variation over time. $\mathbf{C}_s^a$ and $\mathbf{C}_s^b$ are $N_s \times N_s$ symmetric cosine similarities matrices of two different augmented sequences from the same sequence

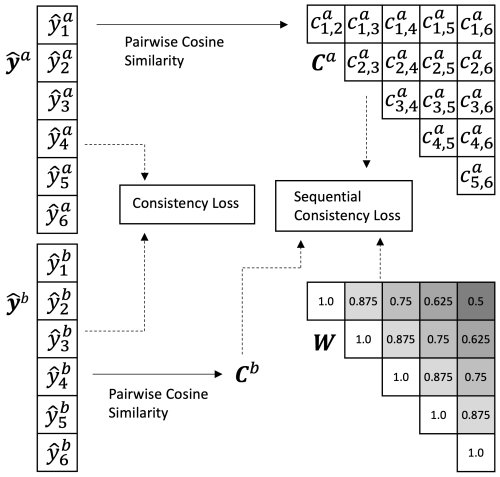

Figure 2: Illustration of two consistency losses when $N_S = 6$. It presents upper triangular matrices of $\mathbf{C}^a$ and $\mathbf{W}$ for a better understanding.

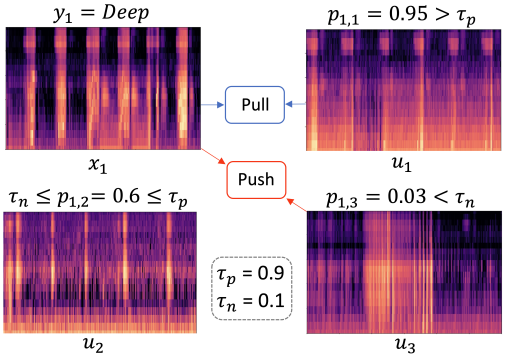

Figure 3: Illustration of SSCL working mechanism. For instance, SSCL pulls u_1 to anchor x_1 since u_1 has high enough confidence for the Deep class due to the clear and regular respiratory pattern. Otherwise, SSCL pushes u_3 since u_3 has no evidence to be a Deep class (other class or possibly OOD).

$\mathbf{u}_s$. Moreover, we define the weighted mask matrix $\mathbf{W} = [w_{i,j}|w_{i,j} = \frac{(N_S-1-|i-j|)(1-w_{\min})}{N_S-2} + w_{\min}]_{1\le i,j\le N_s}$ that weighs close sample pairs more than far ones, where $w_{\min}$ indicates the minimum weight value for the furthest pair. Thus, the loss enforces the predictions of two different augmentations from the same sequence to have a similar sequential trend. Fig.2 describes the whole process of calculating the consistency losses.

Semi-supervised Contrastive Loss. In order to fully utilize the unlabeled data which possibly includes OOD samples, we adopt Class-aware Contrastive Semi-Supervised Learning (CCSSL) from Yang et al. (2022), known as the state-of-the-art robust SSL method, which calculates the supervised contrastive loss (Khosla et al. (2020)) between the unlabeled samples using the pseudo label, and pushes the OOD samples from the in-class feature representation clusters. In a batch of B_u unlabeled sequences with a total of $N = B_u N_s$ samples, CCSSL considers two strong augmentations for each unlabeled sample u_i. Denoting by $i \in I \equiv \{1, \dots, 2N\}$, the index of an arbitrary augmented sample, the contrastive loss of CCSSL is defined as $\mathcal{L}_{\text{CCSSL}} = \sum_{i\in I} \mathcal{L}_{\text{CCSSL},i}$, where we have a loss of the *anchor* i as

$$\mathcal{L}_{\text{CCSSL},i} = -\log \frac{\exp(z_i \cdot z_i^*/\tau)}{\sum_{j\in I} \mathbb{1}_{i\neq j} \exp(z_i \cdot z_j/\tau)} - \sum\nolimits_{k\in\mathcal{K}(i)} \rho_{i,k} \log \frac{\exp(z_i \cdot z_k/\tau)}{\sum_{j\in I} \mathbb{1}_{j\neq i} \exp(z_i \cdot z_j/\tau)}. \quad (2)$$

Here, z_i is an embedding of the anchor i, obtained through ViT (see Fig.1); $z_i^* = z_{(i+N) \bmod 2N}$ is the embedding of the other augmented sample originating from the same unlabeled sample; $\cdot$ denotes the inner product; τ is a temperature; $k \in \mathcal{K}(i)$ represents the index of an augmented sample associated with the same pseudo label; and $\rho_{i,k} := \max(\hat{y}_i)\max(\hat{y}_k)$ is a re-weighting factor.

In the presence of heavily contaminated unlabeled data, the CCSSL can be unreliable, as OOD samples from the unlabeled data can be sampled with high confidence, leading to confusion in class clustering. To address this issue, we introduce Semi-Supervised Contrastive Learning (SSCL), a method that leverages reliable labeled data as anchor points for class clustering. Considering the labeled sample as an anchor, SSCL can use trustworthy positive and negative samples and achieve better class clusters in an embedding space by pushing away the OOD samples from the in-class embedding cluster. The contrastive loss for a labeled anchor m can be defined as:

$$\mathcal{L}_{\text{SSCL},m} = -\sum\nolimits_{i\in\mathcal{P}(m)} p_{m,i} \log \frac{\exp(z_m \cdot z_i/\tau)}{\sum_{j\in\mathcal{N}(m)} \exp(z_m \cdot z_j/\tau)}, \quad (3)$$

where m indicates the index of a sample in a batch of B_l labeled sequences; and $p_{m,i} := y_m \cdot \hat{y}_i$ means the similarity of the predicted class for the i-th augmented unlabeled sample i and the m-th anchor's class y_m. We only consider the pseudo labels with high confidence to construct positive and negative samples, i.e., denote by $\mathcal{P}(m) = \{i \in I | p_{m,i} > \tau_p\}$ and $\mathcal{N}(m) = \{j \in I | p_{m,i} < \tau_n\}$

Table 1: Home-PSG test result of our proposed method compared to the existing methods. C, SC, CC, SS, and WA denote consistency, sequential consistency, CCSSL, SSCL, and weight average, respectively.

Model	C	SC	CC	SS	WA	F1-score
SoundSleepNet						0.5718
SleepFormer						0.6332
	✓					0.6597
	✓	✓				0.6751
	✓	✓	✓			0.6751
	✓	✓	✓	✓		**0.6780**
	✓	✓	✓	✓	✓	**0.6804**

Table 2: Result of PSG-Audio (top) and Lab-PSG (bottom). SleepFormer$^-$ and SleepFormer$^+$ denote the model trained with only supervised learning and the proposed SSL methods.

Model	Accuracy	F1-Score
SleepFormer$^-$	0.6496	0.4832
SleepFormer$^+$	0.6933	0.5015

Model	Accuracy	F1-Score
SleepFormer$^-$	0.7231	0.7000
SleepFormer$^+$	0.7302	0.7070

the set of positive and negative augmented unlabeled samples, respectively, where τ_p and τ_n are filtering thresholds (Fig.3). As a result, the contrastive loss of SSCL can be obtained as $\mathcal{L}_{\text{SSCL}} = \sum_{m=1}^{B_l N_s} \mathcal{L}_{\text{SSCL},m}$, with the batch size of labeled sequence as B_l. Note that we detach the gradient of the labeled embeddings z_m in SSCL since the goal of SSCL is to push OOD samples in unlabeled sequences far away from the labeled sample, not to train the labeled features. Finally, our overall training loss can be summarized as:

$$\mathcal{L} = \mathcal{L}_{\text{SUP}} + \lambda_{\text{C}}\mathcal{L}_{\text{C}} + \lambda_{\text{SC}}\mathcal{L}_{\text{SC}} + \lambda_{\text{CCSSL}}\mathcal{L}_{\text{CCSSL}} + \lambda_{\text{SSCL}}\mathcal{L}_{\text{SSCL}}, \tag{4}$$

where λ_A is a weighing value for the corresponding loss $\mathcal{L}_A$.

3 Experiments and Results

Data. The model is trained with labeled sleep sounds from 2574 in-lab PSG nights (Lab-PSG), and 2731 unlabeled sleep sounds self-collected by participants at home. The model was then evaluated using three datasets: (i) Lab-PSG (454 nights), (ii) Home-PSG (45 nights), and (iii) PSG-Audio (282 nights). To check the performance of the model in a real-world setting, 45 volunteers underwent PSG tests at home (Home-PSG), and the results were compared with the model's predictions. The generalization ability of the model was tested by evaluating its performance on the open dataset PSG-Audio (Korompili et al., 2021), which predominantly comprises data from apnea patients.

Training Details. We set N_S as 40 following that a sleep technician usually checks $\pm$10 minute when they allocate a label of each 30-second unit. We use labeled and unlabeled batch sizes (B_l, B_u) as 4. The weighing value of each unsupervised loss, $\lambda_{\text{C}}, \lambda_{\text{SC}}, \lambda_{\text{CCSSL}}$, and λ_{SSCL} are 1.5, 0.1, 0.1, and 0.1, respectively. For unsupervised training, the filtering threshold of τ_p, τ_n are 0.9 and 0.2.

Results. We assess the performance trends by incorporating semi-supervised learning methods one by one, as shown in Tab.1. For the supervised baseline, SleepFormer achieved an F-1 score of 0.6332, which is an improvement of 0.0614 compared to SoundSleepNet. Adding consistency and sequential consistency loss resulted in further improvement of 0.0419 (0.6332 $\rightarrow$ 0.6597 $\rightarrow$ 0.6751). Interestingly, incorporating CCSSL didn't improve the performance, but our SSCL significantly boosted it to 0.6780. Finally, by averaging the weights of three models trained with different seeds, we achieved a score of 0.6804, which is an improvement of 0.1085 compared to SoundSleepNet.

Additionally, we evaluated the test performance on PSG-Audio and Lab-PSG datasets, which represent unseen target and labeled source distributions, respectively (Tab. 2). Our proposed method improved the accuracy by 0.0437 even on the unseen PSG-Audio dataset. It should be noted that the F1-score on PSG-Audio is lower than that on our data due to the fact that the dataset mainly consists of heavy apnea patients and is highly imbalanced in terms of sleep stage class distribution. The improvement on Lab-PSG was relatively small since the supervised baseline already achieved good performance on the labeled source distribution data.

4 CONCLUSION

In this paper, we presented a semi-supervised learning approach for processing time-series sleep sound data in real-world scenarios. Our proposed sequential consistency loss enhances the temporal correlation of the model while the semi-supervised contrastive loss helps to improve the cluster of the feature representation with labeled samples and effectively filter out out-of-distribution samples. Our proposed method, built on top of SleepFormer, showed significant and consistent improvements in test data from home environments, unseen target distributions, and even labeled source distributions.

5 ACKNOWLEDGEMENTS

Nojun Kwak was supported by NRF grant (2021R1A2C3006659) and IITP grant (2021-0-01343) funded by Korean Government. Hyeryung Jang was supported by NRF grant (2021R1F1A1063288) funded by the Korea government (MSIT).

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Konrad E. Bloch. Polysomnography: a systematic review. *Technology and health care : official journal of the European Society for Engineering and Medicine*, 5 4:285–305, 1997.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Joonki Hong, Hai Hong Tran, Jinhwan Jung, Hyeryung Jang, Dongheon Lee, In Young Yoon, Jung Kyung Hong, and Jeong Whun Kim. End-to-End Sleep Staging Using Nocturnal Sounds from Microphone Chips for Mobile Devices. *Nature and Science of Sleep*, 14:1187–1201, 2022. ISSN 11791608. doi: 10.2147/NSS.S361270.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 646–661. Springer, 2016.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Jongmok Kim, Jooyoung Jang, Hyunwoo Park, and SeongAh Jeong. Structured consistency loss for semi-supervised semantic segmentation. *arXiv preprint arXiv:2001.04647*, 2020.

Georgia Korompili, Anastasia Amfilochiou, Lampros Kokkalas, Stelios A Mitilineos, Nicolas-Alexander Tatlas, Marios Kouvaras, Emmanouil Kastanakis, Chrysoula Maniou, and Stelios M Potirakis. Psg-audio, a scored polysomnography dataset with simultaneous audio recordings for sleep apnea studies. *Scientific Data*, 8(1):197, 2021.

Sachin Mehta and Mohammad Rastegari. Separable self-attention for mobile vision transformers. *arXiv preprint arXiv:2206.02680*, 2022.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Fan Yang, Kai Wu, Shuyi Zhang, Guannan Jiang, Yong Liu, Feng Zheng, Wei Zhang, Chengjie Wang, and Long Zeng. Class-aware contrastive semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14421–14430, 2022.

A APPENDIX

A.1 DEMOGRAPHICS OF THE PARTICIPANTS

Our paper used three different datasets: Lab-PSG dataset (2574 and 454 nights for train and test, respectively), which is a clinical dataset from the sleep center of Seoul National University Bundang Hospital (SNUBH). We gathered the Home-PSG test dataset (45 nights) from the adult volunteers by setting the portable PSG device in their homes. With this, we successfully built the home-environment test dataset with trustworthy labels. The demographics of the above two datasets are described in Tab.3. In order to verify the generalizable performance of our model, we also used the public sound-based PSG dataset, namely PSG-Audio dataset (Korompili et al., 2021). Demographics and AHI statistics of the PSG-Audio dataset are introduced in Korompili et al. (2021). Notice that the proportion of severe apnea ($30 \leq$ AHI) in PSG-Audio is 88.7%, which is extremely high compared to the Lab-PSG (14.6%, 10.6%) and Home-PSG (11.1%). Therefore, PSG-Audio can be considered as a heavily imbalanced dataset, which significantly reduces the F-1 score while keeping comparably good accuracy performance in Tab.2.

Table 3: Demographics of Lab-PSG and Home-PSG dataset.

Demographics	Lab-PSG-Train	Lab-PSG-Test	Home-PSG
Age	52.6 ± 13.8	54.3 ± 13.6	44.7 ± 15.8
Male, n (%)	1841 (71.5%)	246 (54.2%)	19 (42.2%)
BMI, kg/m^2	25.7 ± 3.9	25.4 ± 3.6	24.0 ± 3.9
AHI	22.8 ± 23.0	21.0 ± 19.5	11.8 ± 16.4
AHI < 5, n (%)	665 (25.8%)	115 (25.3%)	22 (48.9%)
$5 \leq$ AHI < 15, n (%)	613 (23.8%)	115 (25.3%)	11 (24.4%)
$15 \leq$ AHI < 30, n (%)	535 (20.8%)	112 (24.7%)	7 (15.6%)
$30 \leq$ AHI, n (%)	761 (29.6%)	112 (24.7%)	5 (11.1%)

A.2 TRAINING DETAILS

In our SleepFormer model architecture, we used the MobileVitV2-075 as the backbone and ViT-tiny as the prediction head. To evaluate the performance of our model, we reported the test results using its Exponential Moving Average (EMA) model. We used the ImageNet (Deng et al., 2009) pretrained checkpoint available in the open-source library timm (Wightman, 2019). To match the channel size of MobileVit-V2-075 (384) and ViT-tiny (192), we designed the customized intermediate block (LayerNorm (Ba et al., 2016) - Fully Connected - DropPath (Huang et al., 2016)). Other training details and hyperparameters are summarized in Tab.4.

Table 4: Hyperparameters of training SleepFormer$^+$ with the proposed SSL

Config	
Optimizer	AdamW
Optimizer momentum	β_1, β_2 = (0.9, 0.999)
Weight decay	1E-3
Learning rate	3E-4
Warmup epochs	5
Training epochs	15
Drop path	0.2
EMA decay	0.996