# Teacher Assistant-Based Knowledge Distillation Extracting Multi-level Features on Single Channel Sleep EEG

**Heng Liang**[1] , **Yucheng Liu**[1] , **Haichao Wang**[2] and **Ziyu Jia**[1] *

[1] Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] Tsinghua-Berkeley Shenzhen Institute, Shenzhen, China
{hengliang01, yuchengliu1214, hychaowang, jia.ziyu}@outlook.com

## Abstract

Sleep stage classification is of great significance to the diagnosis of sleep disorders. However, existing sleep stage classification models based on deep learning are usually relatively large in size (wider and deeper), which makes them hard to be deployed on wearable devices. Therefore, it is a challenge to lighten the existing sleep stage classification models. In this paper, we propose a novel general knowledge distillation framework for sleep stage classification tasks called SleepKD. Our SleepKD, composed of the multi-level module, teacher assistant module, and other knowledge distillation modules, aims to lighten large-scale sleep stage classification models. Specifically, the multi-level module is able to transfer the multi-level knowledge extracted from sleep signals by the teacher model (large-scale model) to the student model (lightweight model). Moreover, the teacher assistant module bridges the large gap between the teacher and student network, and further improves the distillation. We evaluate our method on two public sleep datasets (Sleep-EDF and ISRUC-III). Compared to the baseline methods, the results show that our knowledge distillation framework achieves state-of-the-art performance. SleepKD can significantly lighten the sleep model while maintaining its classification performance. The source code is available at https://github.com/HychaoWang/SleepKD.

## 1 Introduction

In recent years, sleep disorders are becoming a worrying problem that affects human health. Sleep stage classification is helpful for the diagnosis of sleep disorders. The experts complete the analysis of sleep quality by inferring each sleep stage with the signals from sensors on different parts of the body. Specifically, the signals include Electroencephalogram(EEG), Electromyography(EMG), Electrooculography(EOG), etc. Then, these signals are segmented into 30-second data samples which are called sleep epoch for the sleep stage classification. Finally, experts classify each sleep epoch into a specific stage according to the criteria of the American Academy of Sleep Medicine (AASM) [Berry *et al.*, 2012] or other sleep manuals such as the Rechtschaffen and Kales(R&K) [Rechtschaffen, 1968]. Therefore, manual stage classification is a very time-consuming task.

To automate the sleep stage classification, some deep learning methods are applied [Sekkal *et al.*, 2022; Jia *et al.*, 2022a; Liu and Jia, 2023]. For example, DeepSleepNet [Supratak *et al.*, 2017] and SailentSleepNet [Jia *et al.*, 2021b] are used to automatically extract multi-level features from sleep signals. Specifically, there are two kinds of important features in the sleep signals, which are epoch-level features and sequence-level features. The epoch-level features represent the local characteristics of a single sleep epoch. For example, the N2 stage includes mainly sleep spindles and K complexes. The sequence level features are the transition rules between multiple sleep epochs. For instance, the N1 stage often serves as a transition stage between the W stage and other stages. To capture these features, the intermediate layers of existing sleep models are usually relatively large in size (wider and deeper). To the best of our knowledge, some sleep stage classification models usually have parameters up to the order of 100k or even 1M. The deployment of models may cost a large amount of computing resources.

In order to lighten the large-scale model, some knowledge distillation methods are applied [Wang and Yoon, 2021]. The teacher model (complex and large-scale model) can transfer knowledge to the student model (lightweight model) with the knowledge distillation framework. However, the performance of most knowledge distillation methods directly applied to sleep stage classification is unsatisfactory. These methods ignore the valuable information in the multiple levels of sleep epoch features and sleep sequence features shown in Figure 1. Therefore, it is a challenge to design a distillation framework that transfers multi-level sleep knowledge.

Another challenge is how to bridge the gap between teacher and student network with the smallest loss of knowledge. Specifically, in most cases, the teacher network is deep while the student network is shallow as shown in Figure 2. In some circumstances where teacher and student network have too much difference, knowledge may be transferred inefficiently [Mirzadeh *et al.*, 2020]. Moreover, intermediate features extracted by the sleep stage classification model is rela-

*Ziyu Jia is the corresponding author.

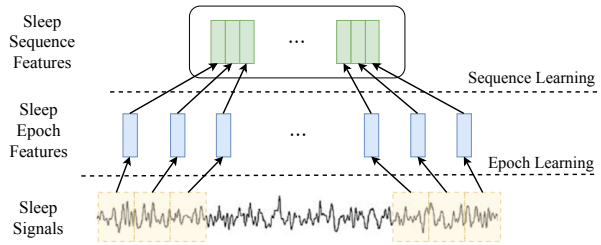


Figure 1: Most of the deep learning methods in sleep stage classification focus on the multi-level features. Epoch-level features and sequence-level features are extracted separately in different parts of the model.

tively more complex and can not be fully conveyed to student because of the dimension alignment [Aguilar *et al.*, 2020].

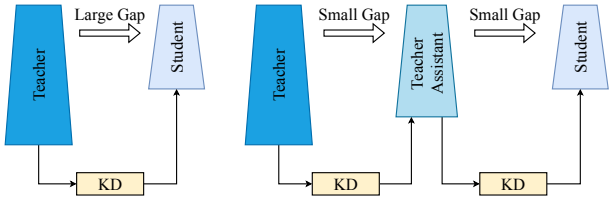


Figure 2: In general, the difference between teacher and student networks limits the reduction of network complexity. The teacher assistant is designed to bridge the gap between teacher and student networks.

To address the above challenges, we propose a general knowledge distillation framework called SleepKD to lighten the EEG-based sleep stage classification model as shown in Figure 3. This framework consists of the multi-level distillation module, the teacher assistant module, and the other knowledge distillation module. The main contributions are summarized as follows:

- To the best of our knowledge, it is the first attempt to use multi-level knowledge distillation in the sleep stage classification model using EEG. Our proposed multi-level knowledge distillation module can efficiently boost epoch-level and sequence-level knowledge transfer from the teacher network to the student.
- We design the teacher assistant module for different kinds of sleep stage classification models. It bridges the excessive gap between teacher and student network for the gap-sensitive multi-level knowledge transfer.
- Experimental results show that our knowledge distillation framework achieves SOTA performance compared to existing knowledge distillation methods. In addition, the proposed knowledge distillation framework can significantly lighten the sleep model while maintaining its classification performance.

## 2 Related Works

### 2.1 Sleep Stage Classification

Sleep stage classification is widely used to diagnose diseases such as sleep disorders. In early studies, machine learning methods are utilized to classify the sleep stages [Tzimourta *et al.*, 2018; Basha *et al.*, 2021; Sundararajan *et al.*, 2021; Jia *et al.*, 2022b]. However, these methods need a large amount of prior knowledge, which means they require a lot of manual costs to extract features. Therefore, many researchers start to implement automatic sleep stage classification by using deep learning methods.

Currently, there are two typical deep learning architectures that are widely used for sleep stage classification, CNN-based [Zhang and Wu, 2017; Cui *et al.*, 2018; Phan *et al.*, 2019] and a hybrid of CNN and RNN [Yang *et al.*, 2018; Back *et al.*, 2019; Fan *et al.*, 2021]. CNN-based architectures are widely applied to sleep stage classification models. CNN is used to extract sequence level information in sleep signals [Chambon *et al.*, 2018]; SleepUtime [Perslev *et al.*, 2019] is proposed with a fully feed-forward deep learning method based on physiological time series segmentation by using the U-Time module; SalientSleepNet [Jia *et al.*, 2021b] is devised with a $U^2$-structure stream by nesting multiple U-units and capturing more useful information for sleep stage classification. It extracts multi-level features with Multi-Scale Extractor (MSE) for sleep epoch and sequence features.

Also, researchers propose a series of sleep stage classification models based on the hybrid architecture of CNN and RNN. DeepSleepNet [Supratak *et al.*, 2017] is applied to extract sleep epochs features and sequence features by using CNN and Bi-directional Long Short-Term Memory (BiLSTM); A hierarchical neural network is designed to learn both comprehensive features and sequential features for sleep stage classification [Sun *et al.*, 2019]; SleepEEGNet [Mousavi *et al.*, 2019] is devised to extract time-invariant features from the original signal, and to capture both long-term and short-term context dependencies by using Bidirectional Recurrent Neural Network (BRNN). In addition, there are other architectures that can be used for sleep stage classification. For example, GraphSleepNet [Jia *et al.*, 2020] is composed of a deep graph neural network for sleep stage classification. MSTGCN [Jia *et al.*, 2021a] is constructed to extract the sleep features by using ST-GCN. An improved real-time sleep stage estimation is devised to have a better sleep stage classification [Harada *et al.*, 2017]. An evolutionary algorithm is used to complete the age-based sleep stage estimation [Matsushima *et al.*, 2012]. Relative evaluation is employed to score the sleep stage by heart rate [Tobaru *et al.*, 2019].

Although these approaches yield good results in the field of sleep stage classification, the parameters of the network model grow. This leads to high computational and storage costs for the models at the industrial level, making deployment difficult to be achieved. Therefore, lightweight sleep stage classification models are particularly important. There have been a few related studies such as [Joshi *et al.*, 2021], which ignore the intermediate features of the sleep signals or the large gap between teacher and student.

### 2.2 Knowledge Distillation

Knowledge distillation can lighten large-scale models. Most of the knowledge distillation approaches can be classified into two types: distillation from logits [Furlanello *et al.*, 2018; Cho and Hariharan, 2019; Tian *et al.*, 2019; Zhao *et al.*, 2022]

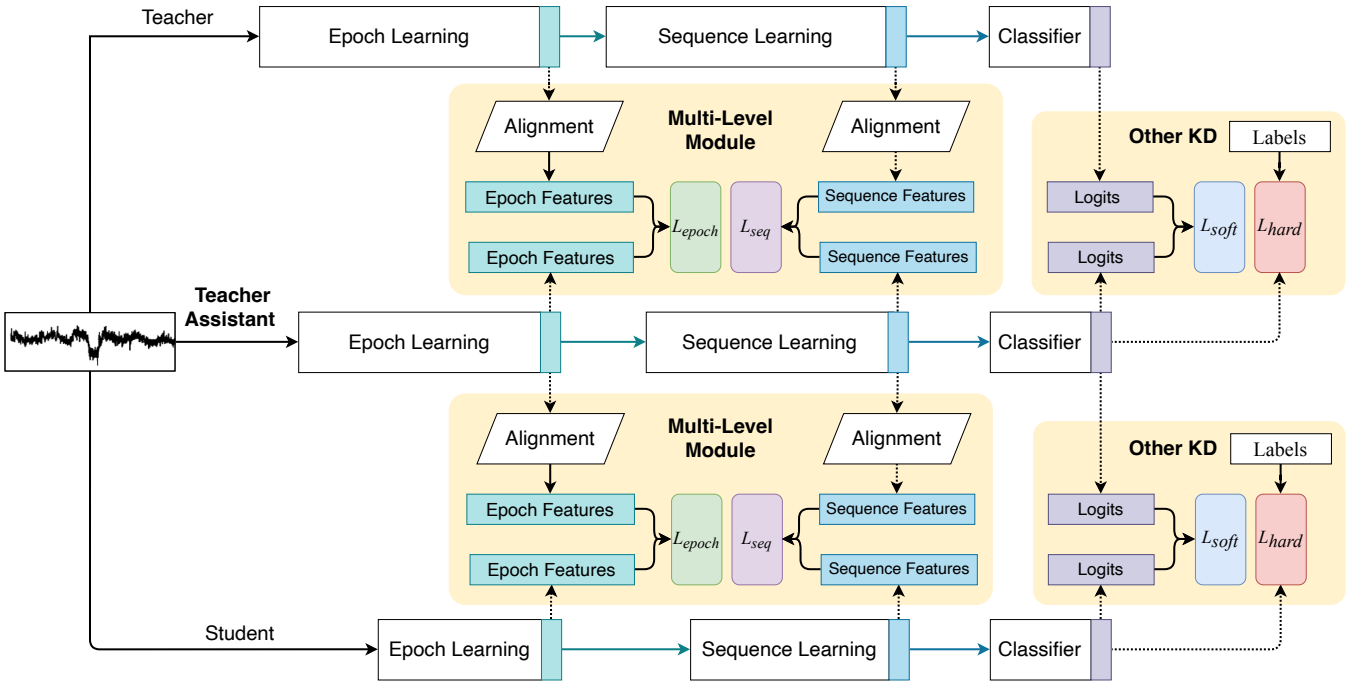


Figure 3: The overview of SleepKD. It includes three modules, which are the multi-level knowledge distillation module ($\mathcal{L}_{epoch}$ and $\mathcal{L}_{seq}$), the teacher assistant module, and other knowledge distillation module ($\mathcal{L}_{soft}$ and $\mathcal{L}_{hard}$). The multi-level knowledge distillation module transfers knowledge in the features from the sleep epochs and sleep sequences. The teacher assistant module is designed to bridge the gap between the teacher and student network.

and distillation from intermediate layers [Yim *et al.*, 2017; Kim *et al.*, 2018; Heo *et al.*, 2019; Chen *et al.*, 2021]. Specifically, the distillation from logits focuses on the logits distribution. The concept of Knowledge Distillation is proposed [Hinton *et al.*, 2015], which aims to transfer knowledge through the difference of the logits distribution of the classification from teacher and student network.

In addition, there are some studies focusing on distilling intermediate features of teacher network. The features of intermediate layers are used to guide the student model during distillation [Romero *et al.*, 2014]. It uses a wide and shallow teacher model to train a narrow and deep student model from intermediate layers with the mean squared error.

Most of the studies are based on the above types of methods. However, there are a small number of special distillation approaches [Zhang *et al.*, 2018; Mirzadeh *et al.*, 2020; Xu *et al.*, 2020; Son *et al.*, 2021]. Actually, different types of distillation methods are complementary. Therefore, the proposed distillation framework combines two major types of distillation methods. We train the student model with knowledge not only from single and multiple epochs of sleep signals but also from the true label and the logits distribution of the teacher network. To improve the performance of distillation even further, we also introduce the teacher assistant to the distillation to help bridge the gap between the teacher and student network.

## 3 Preliminary

The input of the model is a sequence of sleep epochs, and the output is a sequence of corresponding labels. Each sleep epoch is defined as $x \in \mathbb{R}^n$, where $n$ denotes the number of samples in a sleep epoch. The input sequence of sleep epochs is defined as $S = \{x_1, x_2, \cdots, x_L\}$, where $x_i$ $(i \in [1, 2, \cdots, L])$ denotes a sleep epoch and $L$ is the number of sleep epochs from the input sequence.

We distill different sleep stage classification models based on the proposed distillation framework. Then, we evaluate them based on the final classification performance and model compression ratio. We define the predicted output of the model as $\hat{Y} = \{\hat{y_1}, \hat{y_2}, \cdots, \hat{y_L}\}$, where $\hat{y_i} \in \{0, 1, 2, 3, 4\}$ denotes the classification result of $x_i$, corresponding to the five sleep stages W, N1, N2, N3, and REM in the AASM manual, respectively.

## 4 The Proposed SleepKD

Figure 3 presents our SleepKD framework for sleep stage classification. It can be summarized into three key points. 1) We develop a novel multi-level knowledge distillation that simultaneously conveys epoch-level knowledge and sequence-level knowledge in sleep EEG signals. 2) We design the teacher assistant module to bridge the excessive gap between the teacher and student network for the gap-sensitive multi-level knowledge transfer, which makes the distillation more effective. 3) We employ the Kullback-Leibler divergence between the output of teacher and student to transfer the knowl-

edge from logits distribution to the student network.

### 4.1 Multi-Level Module

There are two kinds of important features in the EEG, which are epoch-level features and sequence-level features. They represent the local characteristics of a single sleep epoch and the transition rules between multiple sleep epochs, respectively. In order to capture these two types of features, the intermediate layers of existing models are usually designed to be large. In this paper, we use the knowledge distillation technique to transfer the knowledge extracted from the teacher network to the student network.

Because epoch-level and sequence-level knowledge are extracted from the intermediate layers, we distill these two kinds of knowledge in the intermediate layers of the network. Figure 4 shows the epoch-level knowledge distillation. Specifically, we minimize the difference between teacher's and student's epoch features at the epoch level. This enables the student to learn epoch features from the teacher network. Thus, the student can better capture the features of each single sleep epoch and improve classification with them. The loss at the epoch level is defined as follows:

$$\mathcal{L}_{epoch} = \mathcal{L}_{MSE}\left(\Phi(\boldsymbol{F}_e^T), \boldsymbol{F}_e^S\right) \tag{1}$$

where $\boldsymbol{F}_e^T$ denotes the epoch features of the teacher and $\boldsymbol{F}_e^S$ denotes the epoch features of the student. Because of the dimension difference between the teacher and student features, we use an alignment function $\Phi$. It can be max-pooling or $1 \times 1$ convolution. $\mathcal{L}_{MSE}$ denotes the loss function calculated by mean square error.

In addition, as shown in Figure 4, we calculate the difference between the teacher's and student's sequence features at the sequence level. By minimizing the difference, the student is allowed to learn the sequence-level features from the teacher. The student is able to learn sleep transition rules, further improving classification performance. The loss at the sequence level is defined as follows:

$$\mathcal{L}_{seq} = \mathcal{L}_{MSE}\left(\Phi(\boldsymbol{F}_s^T), \boldsymbol{F}_s^S\right) \tag{2}$$

where $\boldsymbol{F}_s^T$ denotes the sequence features of the teacher. $\boldsymbol{F}_s^S$ denotes the sequence features of the student. $\Phi$ is the alignment function mentioned above. To measure the difference, we choose mean square error as the loss function, which is denoted as $\mathcal{L}_{MSE}$.

### 4.2 Teacher Assistant Module

Existing studies have shown that knowledge transfer is hindered when teacher and student network are too much different. In the multi-level module, we introduce multi-level knowledge from intermediate layers in the network. However, the dimensions of intermediate features are mismatched because teacher and student have different architectures. To transfer the knowledge, dimension alignment between multi-level features is necessary. During the dimension alignment, the complex knowledge in these features could be lost, which makes it more sensitive to the gap. To smooth the transfer

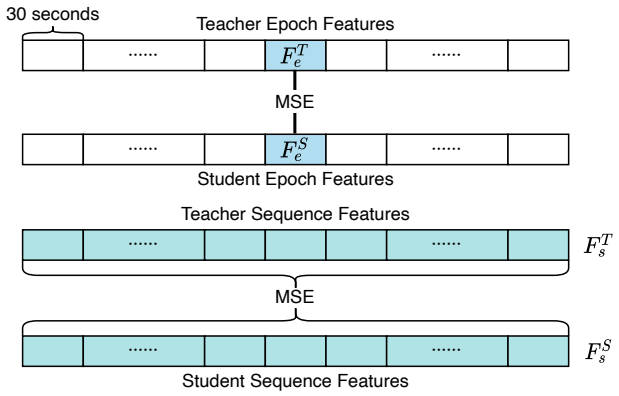


Figure 4: The diagram of multi-level knowledge distillation. Two levels of knowledge are extracted from EEG by the sleep models. By minimizing the difference between representations of epochs and sequences from teacher and student, the rich knowledge at these two levels is conveyed efficiently.

of multi-level knowledge, we design a teacher assistant module between teacher and student network to bridge the gap. It enhances the transfer of multi-level knowledge and makes distillation more effective for sleep stage classification.

We design the teacher assistant module for two kinds of typical sleep stage classification architectures. One is CNN-based architectures and the other is hybrid architectures based on CNN and RNN. For the model using CNNs to extract epoch and sequence features of EEG signals, we design a medium-sized teacher assistant model by reducing the convolution layers of the teacher assistant model between teacher and student network. For instance, the number of CNN layers of the teacher model is 6, and the number of CNN layers of the student model is 2. We set the number of CNN layers of the teacher assistant model to 4. The details of the design are shown in Figure 5.

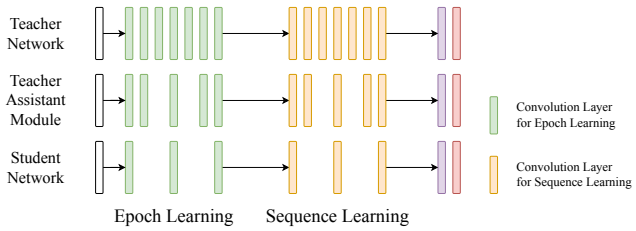


Figure 5: Design of the teacher assistant module based on CNN architecture.

For the hybrid architecture with CNN and RNN, we design a teacher assistant module with a scale between the teacher model and the student model. For the CNN layers of the teacher assistant module, we reduce it as CNN-based architectures above, which means the CNN depth of the teacher assistant model is between the teacher and student network. Meanwhile, we design a medium-sized RNN component to extract the sequence-level features. For example, the number of units of the RNN layer in the teacher network is 512, and the student network is 128. We set the number of units of the

RNN layer in the teacher assistant network to 256. We design the teacher assistant module through such a strategy, and the details are shown in Figure 6.

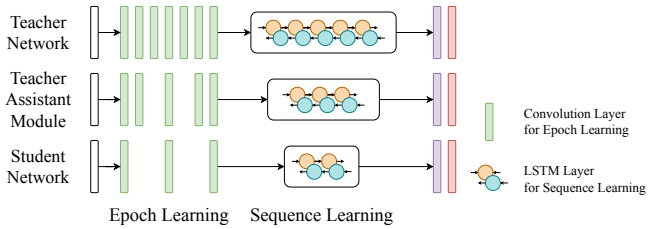


Figure 6: Design of the teacher assistant module based on the hybrid architecture of CNN and RNN.

### 4.3 Other Knowledge Distillation Module

The soft labels, which are the probability distribution for each stage from the teacher model output, also contain useful knowledge. Therefore, we introduce Kullback-Leibler divergence to compute $\mathcal{L}_{soft}$ between the teacher and student network. This allows the teacher network to transfer knowledge from its logits distribution to the student network. $\mathcal{L}_{soft}$ is defined as follows:

$$\mathcal{L}_{soft} = D_{KL}\left(\boldsymbol{p}^T \| \boldsymbol{p}^S\right) \tag{3}$$

where $D_{KL}$ denotes the Kullback-Leibler divergence, which is used to calculate the relative entropy of the output distribution between the teacher model and the student model. $\boldsymbol{p}^T$ denotes the output of the teacher model and $\boldsymbol{p}^S$ denotes the output of the student model. Moreover, we calculate $\mathcal{L}_{hard}$ using the cross-entropy loss function to obtain knowledge of hard labels. $\mathcal{L}_{hard}$ is defined as follows:

$$\mathcal{L}_{hard} = \mathcal{L}_{CE}\left(\boldsymbol{y}, \boldsymbol{p}^S\right) \tag{4}$$

where $\mathcal{L}_{CE}$ denotes the cross-entropy loss function and $\boldsymbol{y}$ denotes the true label. Finally, the total loss $\mathcal{L}_{Total}$ is defined as follows:

$$\mathcal{L}_{Total} = \alpha\mathcal{L}_{epoch} + \beta\mathcal{L}_{seq} + \gamma\mathcal{L}_{soft} + \delta\mathcal{L}_{hard} \tag{5}$$

where $\alpha$, $\beta$, $\gamma$, $\delta$ denote the weights of $\mathcal{L}_{epoch}$, $\mathcal{L}_{seq}$, $\mathcal{L}_{soft}$, $\mathcal{L}_{hard}$, respectively.

## 5 Experiments

### 5.1 Datasets and Data Processing

We evaluate our method on two public datasets: ISRUC-III [Khalighi *et al.*, 2016] and Sleep-EDF [Kemp *et al.*, 2018].

***ISRUC-III*** collects the PSG data samples from 10 subjects (1 for males and 9 for females) for a whole night in 8 hours. The annotations of this dataset are scored by two professional experts.

***Sleep-EDF*** is a very famous public dataset that contains the PSG data samples from 20 subjects (10 for males and 10 for females) in 2 days. The ages of the subjects range from 25 to 34 years old. These recordings were manually classified into one of the eight classes (W, N1, N2, N3, N4, REM, Movement, Unknown) by sleep experts according to the R&K standard. For a fair comparison, we remove the Movement and Unknown stage, and merge the N3 and N4 stage into a single N3 stage according to the AASM manual.

In the experimental data, the EEG is typically segmented into sleep epochs of 30 seconds. We finally extract the sleep epochs of the EEG signal from the Fpz-Cz channel in the Sleep-EDF and the sleep epochs of the EEG signal from F3-A2 in ISRUC-III. The EEG data from each dataset is downsampled to 100Hz.

### 5.2 Baseline Methods

We select some classical and well-performing knowledge distillation methods as baseline methods from three aspects: distillation from logits, distillation from intermediate features, and distillation with teacher assistants.

- KD [Hinton *et al.*, 2015]: Propose a simple way to improve the performance by distilling the knowledge of the complex model into a compact model with the output of the former.
- Fitnets [Romero *et al.*, 2014]: Extend the idea of the traditional knowledge distillation by using both the output of the teacher network and the intermediate representation as a hint to the student.
- NST [Huang and Wang, 2017]: Implement a knowledge transfer loss function by minimizing the Maximum Mean Discrepancy between the feature map of the sophisticated model and the slimming model.
- TAKD [Mirzadeh *et al.*, 2020]: Introduce a multi-step knowledge distillation by using teacher assistant (TA) whose size is between the teacher and student model.
- DGKD [Son *et al.*, 2021]: Devise the densely-guided knowledge method using multiple teacher assistant to fill the large gap between teacher and student model gradually.
- DKD [Zhao *et al.*, 2022]: Reformulate the classical KD method with non-target class knowledge distillation (NCKD) and target class knowledge distillation (TCKD).

### 5.3 Experiment Settings and Implementation

We split the datasets into the train, validation, and test sets by a ratio of 8:1:1 on Sleep-EDF and ISRUC-III separately. The input sleep sequence is 20-epoch long. Each epoch lasts 30 seconds. We implement the teacher, teacher assistant, and student models with TensorFlow. We use Adam as the optimizer in each experiment. In experiments of the CNN framework, we choose SalientSleepNet as a representative. The learning rate of SalientSleepNet is 0.001. The number of training epochs is 60 and the batch size is 8. The weights are $\alpha$ = 0.3, $\beta$ = 0.2, $\gamma$ = 0.4 and $\delta$ = 0.1. In experiments of the CNN and RNN framework, we choose DeepSleepNet as a representative. DeepSleepNet has a learning rate of 0.00001. The number of training epochs is 200 for SleepEDF, 300 for ISRUC-III and the batch size is 20. The weights are $\alpha$ = 1.0, $\beta$ = 0.1 and $\gamma$ = $\delta$ = 1.0. We set these hyperparameters according to the performance on validation sets.

We design corresponding TA and student network for SalientSleepNet and DeepSleepNet. As for SalientSleepNet, we take the number of U-units and Multi-Scale Extractors (MSE) as the proxy of the model capacity since they are designed to extract epoch-level and sequence-level knowledge, respectively. In the implementation of DeepSleepNet, we consider the number of convolution layers in each stream's epoch extractor and units in BiLSTM, which is employed to capture the sequence features from a series of epoch features, as the complexity of our models. As a result, we decide to reduce the number of these components above to design corresponding teacher assistants and students. The specific numbers of layers are presented in Table 1.

| Model | SalientSleepNet | DeepSleepNet |
|---|---|---|
| Teacher | 5 U-units & 5 MSE | 4 CNN & 512 BiLSTM |
| TA | 3 U-units & 3 MSE | 2 CNN & 256 BiLSTM |
| Student | 2 U-units & 2 MSE | 1 CNN & 128 BiLSTM |

Table 1: Summary of the architecture for teacher, TA, and student designing for different kinds of sleep stage classification models.

In order to evaluate the classification performance of the model and compare it with the other baseline models, we employ Accuracy and F1-Score as the evaluation metrics, which are defined as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$F1\text{-}Score = \frac{2TP}{2TP + FP + FN} \tag{7}$$

where $TP$ is a true positive, $TN$ is a true negative, $FP$ is a false positive, and $FN$ is a false negative.

Besides, we also use compression ratio to evaluate our method, which is calculated as follows:

$$Compression\ Ratio = \frac{P_{Teacher} - P_{Student}}{P_{Teacher}} \tag{8}$$

where $P_{Teacher}$ and $P_{Student}$ represent the number of parameters in teacher and student network, respectively.

## 5.4 Experiment Results

As shown in Table 2 and Table 3, we perform several experiments with SleepKD and baseline methods on SalientSleepNet and DeepSleepNet, which are classic models with a CNN framework and a hybrid framework based on CNN and RNN, respectively. Our SleepKD achieves the SOTA knowledge distillation results.

As for KD and DKD, they only focus on the knowledge from logits distribution. These kinds of approaches introduce the output of the teacher network as the extra label to help the student to reach better performance. By contrast, there is a limitation of information transfer in these kinds of methods. Hence, their classification performance is relatively lower than the other approaches. Besides, Fitnets and NST concentrate on the knowledge in the feature map of intermediate layers. However, the knowledge from intermediate layers is gap-sensitive. When facing a student network with a high compression ratio, these methods are restricted by the significant gap between the teacher and student network. As a result, information in the teacher network may not be conveyed efficiently. In addition, the TAKD and DGKD present a great performance because they realize that the gap between teacher and student can be bridged by teacher assistant. However, these types of knowledge distillation ignore the multi-level features from intermediate layers. Student can not learn to extract information in epochs and sequences from the teacher which limits their performance. Because of the consideration of the knowledge at multiple levels (epoch knowledge and sequence knowledge) and the huge gap between the teacher and student network, our SleepKD achieves the best performance. Take the performance on Sleep-EDF as an example: the accuracy of SleepKD reaches up to 87.05% and 85.66% on SalientSleepNet and DeepSleepNet.

| Method | ISRUC-III | | Sleep-EDF | |
|---|---|---|---|---|
| | Acc | F1-Score | Acc | F1-Score |
| KD | 74.65 | 73.74 | 83.62 | 78.93 |
| Fitnets | 75.00 | 73.33 | 85.33 | 80.21 |
| NST | 75.68 | 75.46 | 83.67 | 77.85 |
| TAKD | 77.27 | 76.19 | 85.57 | 80.74 |
| DGKD | 76.70 | 73.68 | 85.19 | 78.86 |
| DKD | 76.70 | 73.73 | 84.64 | 78.96 |
| **SleepKD** | **79.66** | **78.57** | **87.05** | **81.40** |

Table 2: The comparison of the knowledge distillation approaches applied on SalientSleepNet.

| Method | ISRUC-III | | Sleep-EDF | |
|---|---|---|---|---|
| | Acc | F1-Score | Acc | F1-Score |
| KD | 80.22 | 74.54 | 81.28 | 64.41 |
| Fitnets | 81.11 | 75.05 | 80.59 | 65.83 |
| NST | 81.59 | 76.48 | 84.71 | 68.53 |
| TAKD | 81.59 | 76.46 | 83.97 | 67.87 |
| DGKD | 81.36 | 75.75 | 84.47 | 68.46 |
| DKD | 79.88 | 75.37 | 83.88 | 67.78 |
| **SleepKD** | **83.29** | **77.29** | **85.66** | **69.46** |

Table 3: The comparison of the knowledge distillation approaches applied on DeepSleepNet.

Furthermore, we evaluate SleepKD in different aspects (which include parameters, compression ratio, acceleration, etc.). Table 4 and Table 5 present that the student models achieve 74.68% and 71.78% on the compression ratio while the reduction of the accuracy are less than 1%. These data reveal that our framework is able to compress the model the most with the least cost of accuracy. Therefore, the performance in different aspects gets a significant improvement.

## 5.5 Ablation Experiments

To evaluate the effectiveness of each module, the ablation experiments are designed. With the same experiment set-

| Metric | Teacher | Student |
|---|---|---|
| Accuracy | 80.34% | 79.66% |
| Memory | 632.88MB | 160.24MB |
| Parameters | 474,662 | 120,181 |
| Compression Ratio | 74.68% | |
| Acceleration | 6.85x | |

Table 4: Performance of SleepKD on SalientSleepNet. SleepKD can significantly accelerate the inference and reduce the cost of memory and parameters while maintaining the accuracy on SalientSleepNet.

| Metric | Teacher | Student |
|---|---|---|
| Accuracy | 83.97% | 83.29% |
| Memory | 21.46MB | 6.04MB |
| Parameters | 5,502,474 | 1,552,906 |
| Compression Ratio | 71.78% | |
| Acceleration | 5.59x | |

Table 5: Performance of SleepKD on DeepSleepNet. SleepKD can significantly accelerate the inference and reduce the cost of memory and parameters while maintaining the accuracy on DeepSleepNet.

tings, we select different combinations of loss terms to verify the effectiveness of each loss term in SleepKD loss function $\mathcal{L}_{Total}$.

- $\mathcal{L}_1 = \mathcal{L}_{Total} - \mathcal{L}_{seq}$
- $\mathcal{L}_2 = \mathcal{L}_{Total} - \mathcal{L}_{epoch}$
- $\mathcal{L}_3 = \mathcal{L}_{Total} - \mathcal{L}_{soft}$
- $\mathcal{L}_4 = \mathcal{L}_{Total} - \mathcal{L}_{hard}$

Figure 7 demonstrates that each loss term in SleepKD is useful and effective. These modules transfer valuable information. Multi-level information of EEG signals transferred by $\mathcal{L}_{epoch}$ and $\mathcal{L}_{seq}$ significantly improve student performance. Students learn the teacher's probability distribution by $\mathcal{L}_{soft}$ and knowledge from labels by $\mathcal{L}_{hard}$.

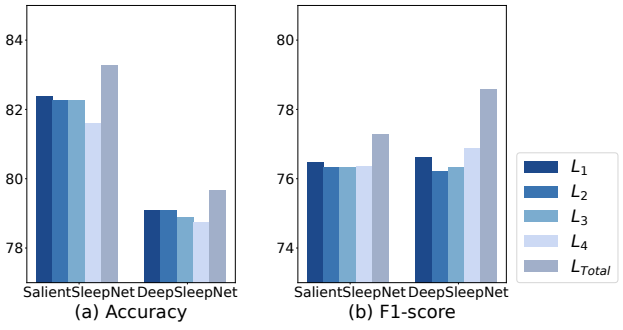


Figure 7: The results of ablation experiments on ISRUC-III dataset. Histogram (a) and Histogram (b) show the accuracy and F1-Score of ablation experiments for the multi-level module.

To demonstrate the effectiveness of the teacher assistant module, the following distillation methods are designed:

- *varient a)*: Multi-Level Module
- *varient b)*: Multi-Level Module + TA Module

Figure 8 illustrates that the teacher assistant module improves the efficiency of distillation knowledge transfer and enhances the performance of the student model by about 1%. The teacher assistant module protects the gap-sensitive multi-level knowledge and helps transfer logits knowledge by smoothing the gap between the teacher and student network.

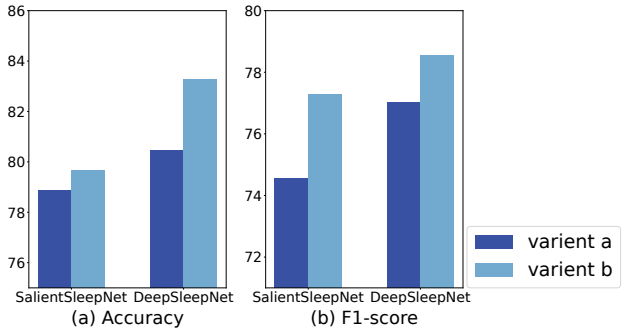


Figure 8: The results of ablation experiments on ISRUC-III dataset. Histogram (a) and Histogram (b) show the accuracy and F1-Score of ablation experiments for the teacher assistant module.

## 6 Conclusion

In this paper, we propose a knowledge distillation framework for the sleep stage classification model. We employ knowledge distillation on the multi-level sleep stage classification model for the first time and introduce the teacher assistant module to improve the distillation. The proposed SleepKD framework can adapt well to the current mainstream multi-level sleep stage classification model. It is able to transfer the features of single sleep stages and transition rules between multiple sleep stages in sleep signals. Meanwhile, we design corresponding teacher assistant modules for different architectures. This can bridge the excessive gap between teacher and student network and further enhance knowledge distillation. Experiments show that our distillation framework achieves excellent results on two popular architectures (CNN-based and hybrid of CNN and RNN). Moreover, SleepKD achieves state-of-the-art distillation performance compared to other distillation methods. The proposed method is a general distillation framework for time series classification. In the future, we can apply the proposed method to other large-scale time series models.

## Acknowledgments

This work was supported by STI2030-Major Projects 2021ZD0200200 and China Postdoctoral Science Foundation 2023M733738.

## Contribution Statement

Heng Liang, Yucheng Liu, and Haichao Wang have equal contributions to this paper.

## References


[Aguilar *et al.*, 2020] Gustavo Aguilar, Yuan Ling, Yu Zhang, Benjamin Yao, Xing Fan, and Chenlei Guo. Knowledge distillation from internal representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7350–7357, 2020.

[Back *et al.*, 2019] Seunghyeok Back, Seongju Lee, Hogeon Seo, Deokhwan Park, Tae Kim, and Kyoobin Lee. Intra-and inter-epoch temporal context network (iitnet) for automatic sleep stage scoring. *arXiv preprint arXiv:1902.06562*, 2019.

[Basha *et al.*, 2021] A Jameer Basha, B Saravana Balaji, S Poornima, M Prathilothamai, and K Venkatachalam. Support vector machine and simple recurrent network based automatic sleep stage classification of fuzzy kernel. *Journal of ambient intelligence and humanized computing*, 12(6):6189–6197, 2021.

[Berry *et al.*, 2012] Richard B Berry, Rita Brooks, Charlene E Gamaldo, Susan M Harding, C Marcus, Bradley 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*, 176:2012, 2012.

[Chambon *et al.*, 2018] Stanislas Chambon, Mathieu N Galtier, Pierrick J Arnal, Gilles Wainrib, and Alexandre Gramfort. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4):758–769, 2018.

[Chen *et al.*, 2021] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2021.

[Cho and Hariharan, 2019] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019.

[Cui *et al.*, 2018] Zhihong Cui, Xiangwei Zheng, Xuexiao Shao, and Lizhen Cui. Automatic sleep stage classification based on convolutional neural network and fine-grained segments. *Complexity*, 2018, 2018.

[Fan *et al.*, 2021] Jiahao Fan, Chenglu Sun, Meng Long, Chen Chen, and Wei Chen. Eognet: A novel deep learning model for sleep stage classification based on single-channel eog signal. *Frontiers in Neuroscience*, 15, 2021.

[Furlanello *et al.*, 2018] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018.

[Harada *et al.*, 2017] Tomohiro Harada, Takahiro Kawashima, Morito Morishima, and Keiki Takadama. Improving accuracy of real-time sleep stage estimation by considering personal sleep feature and rapid change of sleep behavior. In *2017 AAAI Spring Symposium Series*, 2017.

[Heo *et al.*, 2019] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930, 2019.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[Huang and Wang, 2017] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[Jia *et al.*, 2020] Ziyu Jia, Youfang Lin, Jing Wang, Ronghao Zhou, Xiaojun Ning, Yuanlai He, and Yaoshuai Zhao. Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In *IJCAI*, pages 1324–1330, 2020.

[Jia *et al.*, 2021a] Ziyu Jia, Youfang Lin, Jing Wang, Xiaojun Ning, Yuanlai He, Ronghao Zhou, Yuhan Zhou, and H Lehman Li-wei. Multi-view spatial-temporal graph convolutional networks with domain generalization for sleep stage classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1977–1986, 2021.

[Jia *et al.*, 2021b] Ziyu Jia, Youfang Lin, Jing Wang, Xuehui Wang, Peiyi Xie, and Yingbin Zhang. Salientsleepnet: Multimodal salient wave detection network for sleep staging. *arXiv preprint arXiv:2105.13864*, 2021.

[Jia *et al.*, 2022a] Ziyu Jia, Xiyang Cai, and Zehui Jiao. Multi-modal physiological signals based squeeze-and-excitation network with domain adversarial learning for sleep staging. *IEEE Sensors Journal*, 22(4):3464–3471, 2022.

[Jia *et al.*, 2022b] Ziyu Jia, Junyu Ji, Xinliang Zhou, and Yuhan Zhou. Hybrid spiking neural network for sleep electroencephalogram signals. *Science China Information Sciences*, 65(4):140403, 2022.

[Joshi *et al.*, 2021] Vaibhav Joshi, Sricharan Vijayarangan, Preejith SP, and Mohanasankar Sivaprakasam. A deep knowledge distillation framework for eeg assisted enhancement of single-lead ecg based sleep staging. *arXiv preprint arXiv:2112.07252*, 2021.

[Kemp *et al.*, 2018] Bob Kemp, A Zwinderman, B Tuk, H Kamphuisen, and J Oberyé. Sleep-edf database expanded. *physionet. org*, 2018.

[Khalighi *et al.*, 2016] Sirvan Khalighi, Teresa Sousa, José Moutinho Santos, and Urbano Nunes. Isruc-sleep: A comprehensive public dataset for sleep researchers. *Computer methods and programs in biomedicine*, 124:180–192, 2016.

[Kim *et al.*, 2018] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018.

[Liu and Jia, 2023] Yuchen Liu and Ziyu Jia. Bstt: A bayesian spatial-temporal transformer for sleep staging. In *The Eleventh International Conference on Learning Representations*, 2023.

[Matsushima *et al.*, 2012] Hiroyasu Matsushima, Shogo Minami, and Keiki Takadama. Age-based sleep stage estimation by evolutionary algorithm. In *2012 AAAI Spring Symposium Series*, 2012.

[Mirzadeh *et al.*, 2020] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020.

[Mousavi *et al.*, 2019] Sajad Mousavi, Fatemeh Afghah, and U Rajendra Acharya. Sleepeegnet: Automated sleep stage scoring with sequence to sequence deep learning approach. *PloS one*, 14(5):e0216456, 2019.

[Perslev *et al.*, 2019] Mathias Perslev, Michael Jensen, Sune Darkner, Poul Jørgen Jennum, and Christian Igel. U-time: A fully convolutional network for time series segmentation applied to sleep staging. *Advances in Neural Information Processing Systems*, 32, 2019.

[Phan *et al.*, 2019] Huy Phan, Fernando Andreotti, Navin Cooray, Oliver Y. Chén, and Maarten De Vos. Joint classification and prediction cnn framework for automatic sleep stage classification. *IEEE Transactions on Biomedical Engineering*, 66(5):1285–1296, 2019.

[Rechtschaffen, 1968] Allan Rechtschaffen. A manual for standardized terminology, techniques and scoring system for sleep stages in human subjects. *Brain information service*, 1968.

[Romero *et al.*, 2014] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[Sekkal *et al.*, 2022] Rym Nihel Sekkal, Fethi Bereksi-Reguig, Daniel Ruiz-Fernandez, Nabil Dib, and Samira Sekkal. Automatic sleep stage classification: From classical machine learning methods to deep learning. *Biomedical Signal Processing and Control*, 77:103751, 2022.

[Son *et al.*, 2021] Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. Densely guided knowledge distillation using multiple teacher assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9395–9404, 2021.

[Sun *et al.*, 2019] Chenglu Sun, Chen Chen, Wei Li, Jiahao Fan, and Wei Chen. A hierarchical neural network for sleep stage classification based on comprehensive feature learning and multi-flow sequence learning. *IEEE journal of biomedical and health informatics*, 24(5):1351–1366, 2019.

[Sundararajan *et al.*, 2021] Kalaivani Sundararajan, Sonja Georgievska, Bart HW Te Lindert, Philip R Gehrman, Jennifer Ramautar, Diego R Mazzotti, Séverine Sabia, Michael N Weedon, Eus JW van Someren, Lars Ridder, et al. Sleep classification from wrist-worn accelerometer data using random forests. *Scientific Reports*, 11(1):1–10, 2021.

[Supratak *et al.*, 2017] Akara Supratak, Hao Dong, Chao Wu, and Yike Guo. Deepsleepnet: A model for automatic sleep stage scoring based on raw single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(11):1998–2008, 2017.

[Tian *et al.*, 2019] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

[Tobaru *et al.*, 2019] Akari Tobaru, Yusuke Tajima, and Keiki Takadama. Sleep stage estimation using heart rate variability divided by sleep cycle with relative evaluation. In *AAAI Spring Symposium: Interpretable AI for Well-being*, 2019.

[Tzimourta *et al.*, 2018] Katerina D Tzimourta, Athanasios Tsilimbaris, Katerina Tzioukalia, Alexandros T Tzallas, Markos G Tsipouras, Loukas G Astrakas, and Nikolaos Giannakeas. Eeg-based automatic sleep stage classification. *Biomed J*, 1(6), 2018.

[Wang and Yoon, 2021] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[Xu *et al.*, 2020] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *European Conference on Computer Vision*, pages 588–604. Springer, 2020.

[Yang *et al.*, 2018] Yang Yang, Xiangwei Zheng, and Feng Yuan. A study on automatic sleep stage classification based on cnn-lstm. In *Proceedings of the 3rd International Conference on Crowd Science and Engineering*, pages 1–5, 2018.

[Yim *et al.*, 2017] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo 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*, pages 4133–4141, 2017.

[Zhang and Wu, 2017] Junming Zhang and Yan Wu. A new method for automatic sleep stage classification. *IEEE transactions on biomedical circuits and systems*, 11(5):1097–1110, 2017.

[Zhang *et al.*, 2018] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018.

[Zhao *et al.*, 2022] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11953–11962, 2022.