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

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

1 Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China

2 Tsinghua-Berkeley Shenzhen Institute, Shenzhen, China

Introduction

Sleep stage classification:

- The American Academy of Sleep Medicine classifies sleep into five main stages: W, N1, N2, N3, and REM.
- Help doctors correctly diagnose narcolepsy, snoring, Alzheimer‘s, diabetes, depression, and other diseases.

Two typical deep learning architectures:

- CNN-based: SalientSleepNet, MMCNN, etc.
- Hybrid architecture of CNN and RNN: DeepSleepNet, XsleepNet, etc.

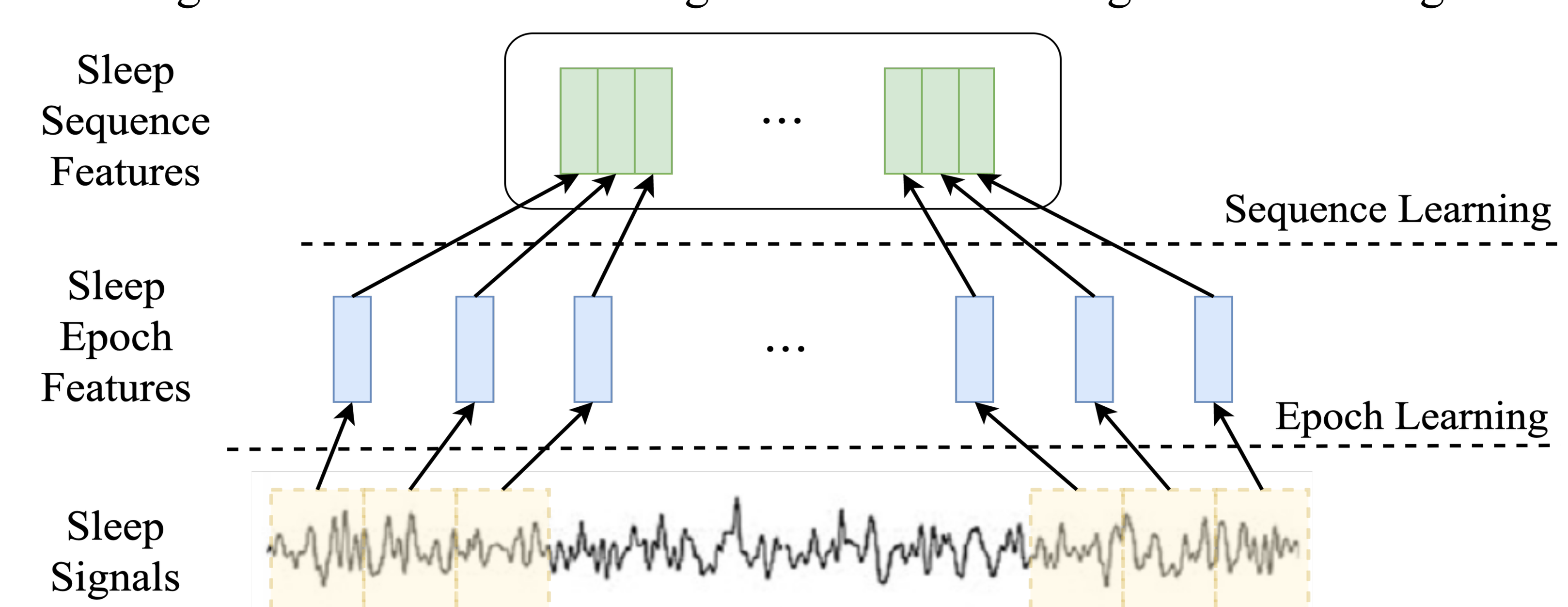
Difficulty in applying deep learning models on wearable devices:

- Large number of parameters, long training time.

Motivation

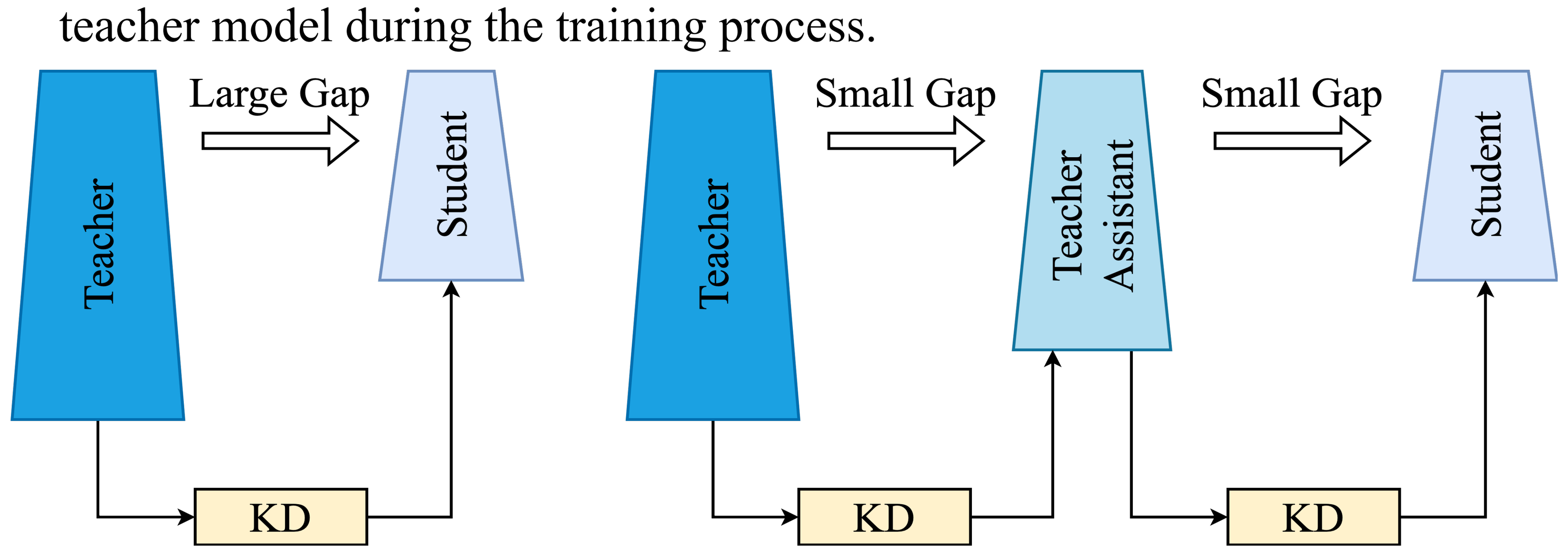
M1: **How can better transfer two important types of knowledge?**

- Epoch-level features: local characteristics of a single sleep epoch. The N2 stage includes mainly sleep spindles and K complexes.
- Sequence-level features: transition rules between multiple sleep epochs. The N1 stage is often a transition stage between the W stage and other stages.

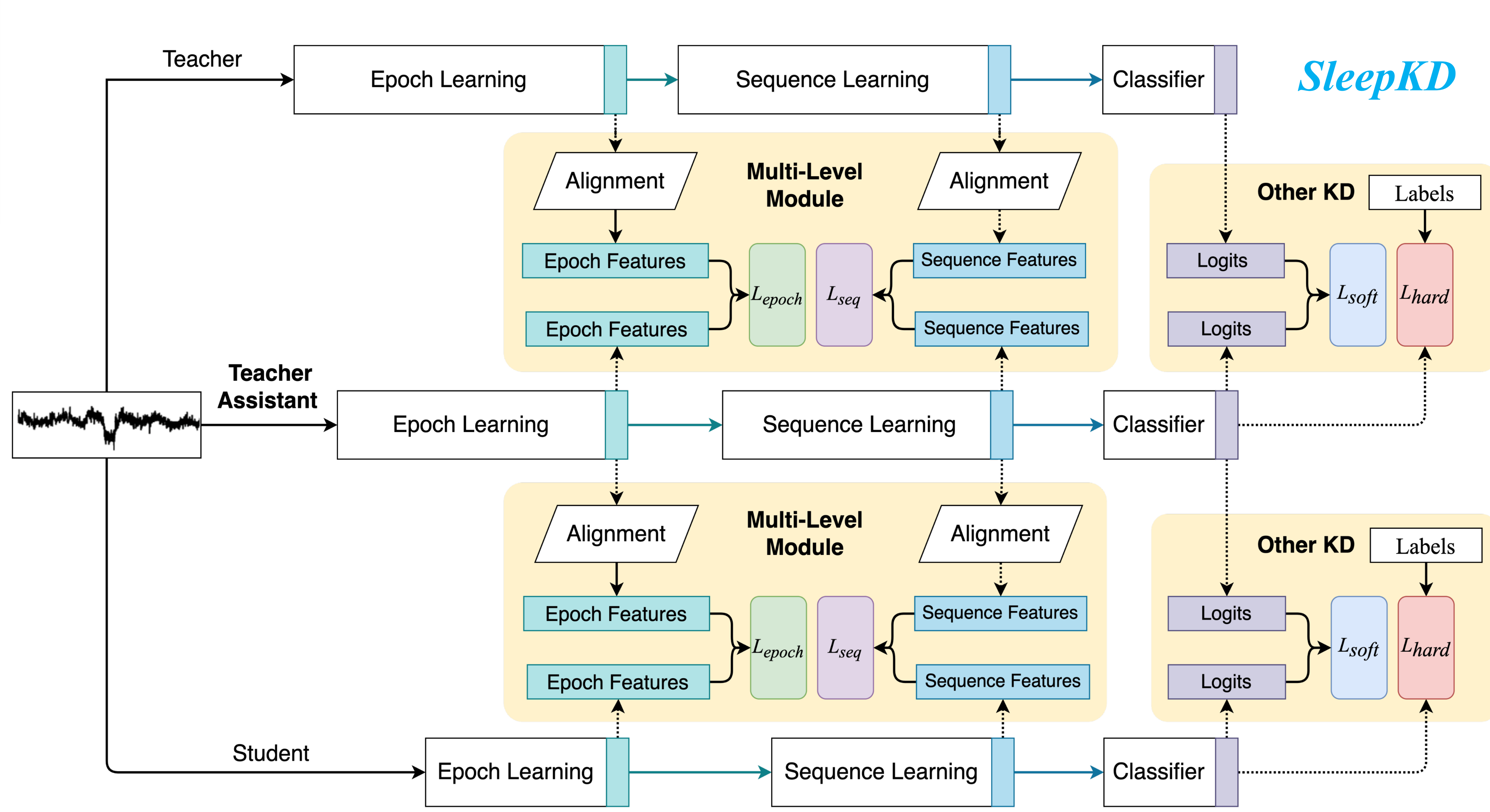


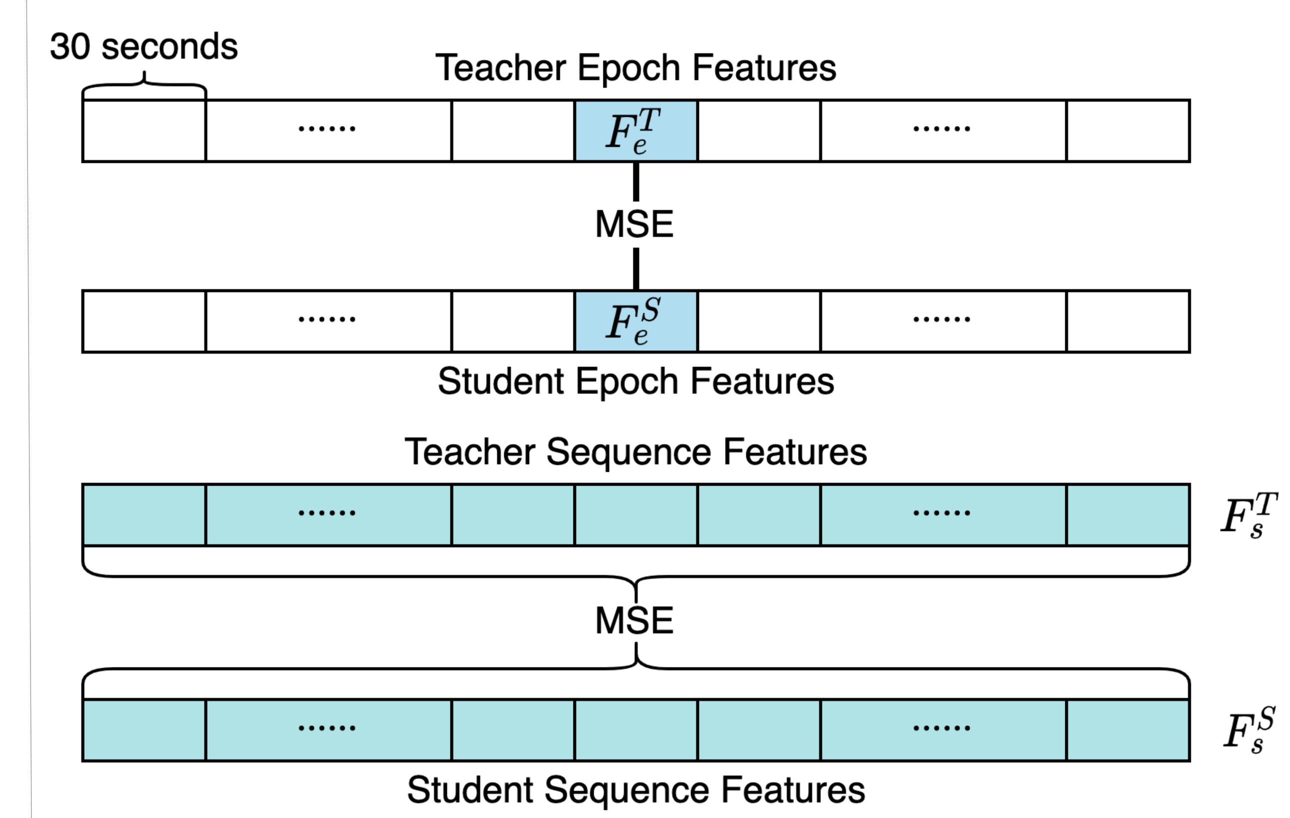
M2: **How to bridge the gap between teacher and student model?**

- The teacher network is often deep while the student network is shallow.
- Excessive gap leads to a difficulty for the student model to learn from the teacher model during the training process.



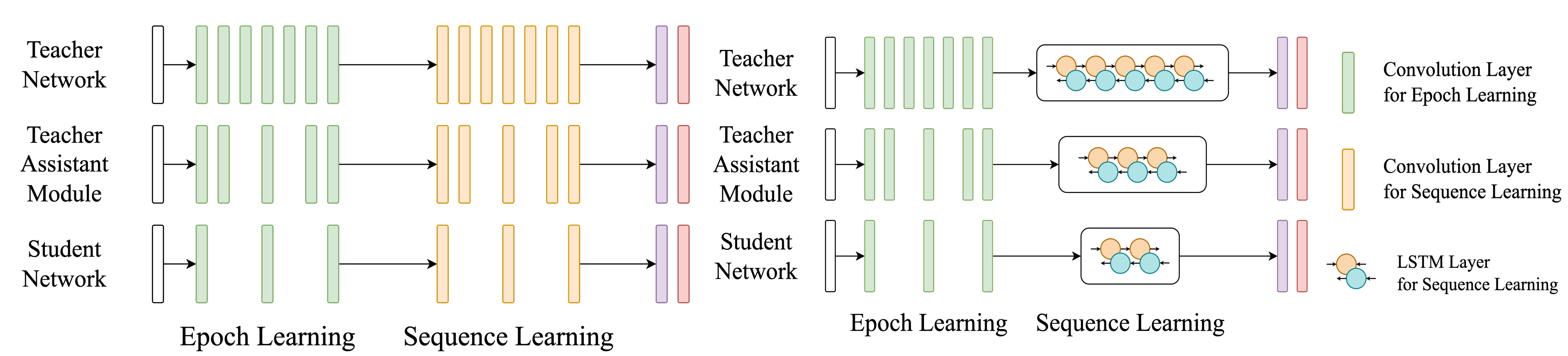
Methods





M1: **Multi-Level Module**

- Capture these two types of features: Epoch-level features and Sequence-level features.
- Calculate the Mean Squared Error and guide the student model to learn the epoch and sequence knowledge of the teacher model.



M2: **Teacher Assistant Module (TA Module)**

- Traditional distillation: knowledge transfer is hindered when teacher and student network have too much difference.
- TA Module: bridges the gap between teacher and student models and improves knowledge transfer.
- We design different TA architectures for both CNN and CNN+RNN structures.

Results

- We apply SleepKD on SalientSleepNet and DeepSleepNet and evaluate the performance on ISRUC-III and Sleep-EDF datasets.
- The baseline comparisons are shown in the tables, SleepKD achieves the SOTA results.

Method	ISRUC-III Acc	ISRUC-III F1-Score	Sleep-EDF Acc	Sleep-EDF 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 Acc	ISRUC-III F1-Score	Sleep-EDF Acc	Sleep-EDF 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.

Conclusion

- We first employ knowledge distillation on the multi-level sleep stage classification model.
- We design the Multi-Level Module to better transfer the epoch-level features and sequence-level features.
- We design the TA Module for two architectures to bridge the gap between teacher and student network.
- SleepKD achieves SOTA distillation performance compared with other methods.

Webpage:https://hychaowang.github.io/SleepKD/; Code:https://github.com/HychaoWang/SleepKD