EEG Lightformer-KD: A Lightweight Framework for Motor Imagery EEG Decoding via Knowledge Distillation

1st Zhengxue Huang
School of Opto-Electronic and Communication Engineering
Xiamen University of Technology
Xiamen, China
2322121006@stu.xmut.edu.cn

2nd Hongyi Zhang*
School of Opto-Electronic and Communication Engineering
Xiamen University of Technology
Xiamen, China
zhanghongyi@xmut.edu.cn

3rd Zhenzhe Zhong
Xiamen Key Laboratory of Intelligent Scene Technology in Industrial Metaverse
Xiamen Intretech Inc
Xiamen, China
xmzzzhe@intretech.com

4th Chao Feng
Xiamen Peiyang BCI & Smart Health Innovation Research Institute
Xiamen Peiyang Ruiheng Smart Health Co., LTD
Xiamen, China
charle_feng @intretech.com

5th Jiancheng Chen
Fujian Key Laboratory of Industrial Internet & IoT
Xiamen Intretech Inc
Xiamen, China
xmcjc@intretech.com

6th Hanwei Zheng
School of Opto-Electronic and Communication Engineering
Xiamen University of Technology
Xiamen, China
waurmelonZ@163.com

7th Shiqi Chen
School of Opto-Electronic and Communication Engineering
Xiamen University of Technology
Xiamen, China
2422121003@stu.xmut.edu.cn

8th Shunlin Cai
School of Opto-Electronic and Communication Engineering
Xiamen University of Technology
Xiamen, China
2422121001@stu.xmut.edu.cn

9th Xingen Gao
School of Opto-Electronic and Communication Engineering
Xiamen University of Technology
Xiamen, China
gaoxingen@xmut.edu.cn

***Abstract*—This study proposes a lightweight convolutional-Transformer hybrid architecture named EEG Lightformer, coupled with a knowledge distillation framework based on the EEGPT large model, to address deployment challenges of motor imagery decoding models in resource-constrained brain-computer interface systems. To address the contradiction between insufficient representation capability of conventional shallow convolutional networks and excessive parameters in large models, we innovatively design a two-stage knowledge transfer strategy: First, constructing a lightweight backbone network using depthwise separable convolutions and multi-scale classification mechanisms to compress parameter scale while preserving discriminative features; second, proposing a joint distillation loss function combining KL divergence distribution alignment with cross-entropy supervision to effectively transfer EEGPT's spatiotemporal perception capabilities. Cross-session experiments on three public datasets (BCI Competition IV 2a/2b and OpenBMI) demonstrate that the distilled EEG Lightformer achieves 2.47-5.52 percentage point accuracy improvements compared to baseline models like EEGNet and FBCnet, while maintaining only 127K-205K parameters. The architecture attains 78.67% average accuracy in four-class tasks, reducing parameters by 84% compared to classical models while maintaining comparable performance. This framework provides a practical solution for wearable BCI systems that balances computational efficiency and decoding accuracy.**

This work was supported by Fujian Province Science and Technology Plan Foreign Cooperation Project (No.2024I0045), Fujian Province Middle-aged and Young Teachers' Educational Research Project (JAT220334 and JAT241123), the Key Science and Technology Program Projects of Xiamen (3502Z20234033), and Xiamen University of Technology Graduate Student Innovation and Research Program (YKJCX2024135 and YKJCX2024122).

***Index Terms*—Knowledge Distillation, Brain-Computer Interface (BCI), Motor Imagery, Lightweight, Kullback-Leibler Divergence.**

I. Introduction

The groundbreaking advancements in deep learning have significantly propelled the development of motor imagery (MI) decoding in brain-computer interface (BCI) systems. Models like EEGPT [1], which employ dual self-supervised pre-training through spatio-temporal representation alignment and masked reconstruction mechanisms, have successfully constructed hierarchical feature representations from massive EEG data, demonstrating exceptional cross-scenario generalization capabilities. However, the model's single inference requires computations involving over ten million parameters, rendering it unsuitable for resource-constrained deployment scenarios such as wearable devices and real-time neurorehabilitation systems. This gives rise to a critical challenge: How to effectively compress EEGPT's spatio-temporal perception capabilities into lightweight models through knowledge transfer mechanisms while maintaining robust MI decoding performance?

Traditional MI decoding approaches typically rely on shallow convolutional networks that prioritize computational efficiency at the expense of representational capacity[2]. While meeting edge deployment requirements, these models exhibit significant performance degradation when confronted with session variability or non-stationary EEG patterns. Knowledge distillation, pioneered by Hinton et al.[3], has proven effective in compressing large models while preserving discriminative capabilities. This methodology aligns well with portable BCI requirements, as resource constraints necessitate models that maintain accuracy while operating efficiently.

We propose a knowledge distillation framework based on the EEGPT large model and present EEG Lightformer – a lightweight convolutional-Transformer hybrid architecture specifically optimized for MI decoding. The implementation involves two key innovations: 1) A lightweight backbone network employing depthwise separable convolutions and multi-scale classification mechanisms; 2) A knowledge distillation strategy combining KL divergence-based unsupervised distribution alignment with cross-entropy-based supervised label learning, enabling effective transfer of EEGPT's representational knowledge to the lightweight model.

Extensive experiments on three widely-used public MI datasets (BCI Competition IV 2a, 2b, and OpenBMI) demonstrate that the distilled EEG Lightformer achieves 2.47-5.52 percentage point improvements in cross-session decoding accuracy compared to conventional shallow convolutional networks, while maintaining lightweight computational requirements.

II. Methods

In this paper, we propose a distillation framework based on the EEGPT large model and a lightweight convolutional Transformer. As illustrated in Figure 1, we first introduce the architectures of the teacher and student models, followed by a detailed description of the distillation process.

A. Teacher Model

The teacher model employs a pre-trained EEGPT architecture, which consists of three core components: an adaptive spatial filter, a local spatio-temporal embedding layer, and a hierarchical Transformer encoder. The adaptive spatial filter utilizes learnable 1×1 convolutions to map input EEG signals into the standardized channel space of the pre-trained model, enabling compatibility with diverse electrode configurations across recording devices. The local spatio-temporal embedding layer adopts a block-wise partitioning strategy to divide raw signals into fixed-time-window spatio-temporal blocks. This layer dynamically encodes electrode names into low-dimensional vectors using a channel embedding codebook, achieving cross-dataset spatial generalization. The hierarchical Transformer encoder extracts multi-scale features through stacked multi-head self-attention modules: short-term spatial patterns (e.g., local neural rhythm variations) are captured in early stages, while long-term temporal dependencies (e.g., dynamic transitions between brain states) are modeled in deeper layers. All parameters of the teacher model remain frozen during the distillation process.

B. Student Model

The student model, named EEG Lightformer, is a lightweight convolutional Transformer designed for efficient decoding of motor imagery EEG signals. It follows a streamlined three-stage architecture that integrates spatio-temporal feature encoding, attention enhancement, and multi-scale classification, achieving an 80% parameter reduction compared to classical EEG Conformer frameworks. The spatio-temporal feature encoding stage begins with temporal convolutions to model time-varying dynamics of the input signals, followed by depthwise separable convolutions that decouple spatial feature extraction into channel-wise and pointwise operations to minimize redundancy. Temporal average pooling then compresses the features while smoothing temporal variations. In the attention enhancement stage, Transformer layers adaptively model long-range dependencies and non-local relationships across the encoded spatio-temporal features. Finally, the multi-scale classification stage combines global average pooling (to preserve holistic signal characteristics) and global max pooling (to amplify discriminative local patterns) on the Transformer outputs, feeding the aggregated features into a compact fully connected layer that reduces classifier parameters by 72%. This architecture ensures computational efficiency without compromising representational capacity through hierarchical abstraction and multi-scale fusion.

C. Knowledge Distillation

Suppose the logits output by the teacher network and the student network are$z_t, z_s \in \mathbb{R}^K$ (where K is the number of categories). By introducing a temperature coefficient $\tau \in \mathbb{R}^K$, the softened probability distribution is calculated as:

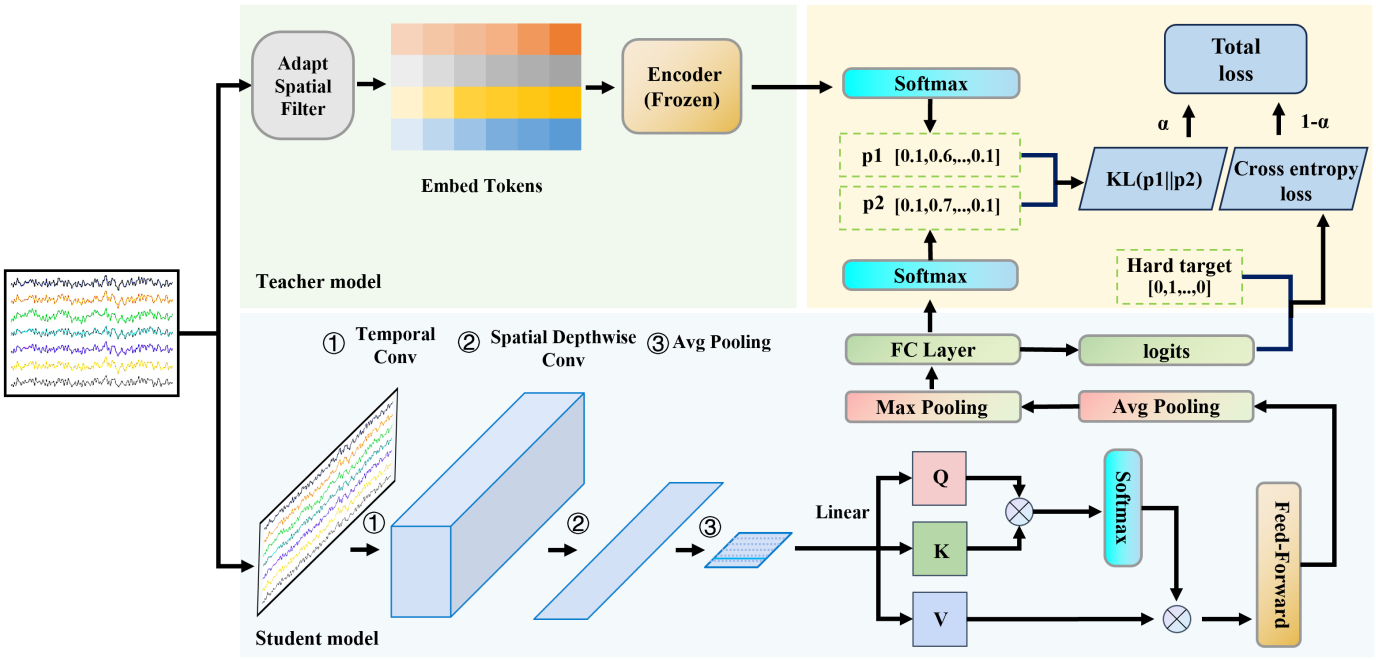

Fig. 1. Distillation from EEGPT into EEG Lightformer

$$\boldsymbol{p}_t = \text{Softmax}\left(\frac{\boldsymbol{z}_t}{\tau}\right) = \left[\frac{e^{z_t^{(k)}/\tau}}{\sum_{j=1}^{K} e^{z_t^{(j)}/\tau}}\right]_{k=1}^{K} \tag{1}$$

$$\boldsymbol{p}_s = \text{Softmax}\left(\frac{z_s}{\tau}\right) = \left[\frac{e^{z_s^{(k)}/\tau}}{\sum_{j=1}^{K} e^{z_s^{(j)}/\tau}}\right]_{k=1}^{K} \tag{2}$$

The student network is trained to mimic the soft label distribution of the teacher network through Kullback-Leibler (KL) divergence, while simultaneously optimizing for ground-truth alignment via cross-entropy loss. The joint training objective is formulated as:

$$\mathcal{L}_{\text{KD}} = \tau^2 \cdot D_{\text{KL}}\left(\boldsymbol{p}_t \parallel \boldsymbol{p}_s\right) = \tau^2 \cdot \sum_{k=1}^{K} p_t^{(k)} \log\left(\frac{p_t^{(k)}}{p_s^{(k)}}\right) \tag{3}$$

$$\mathcal{L}_{\text{CE}} = -\sum_{k=1}^{K} y^{(k)} \log p_s^{(k)} \tag{4}$$

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{KD}} + (1-\alpha)\mathcal{L}_{\text{CE}} \tag{5}$$

Here,$\alpha \in [0,1]$,denotes the distillation weight coefficient, and $y^{(k)}$ represents the one-hot encoding of ground-truth labels. During the optimization process, all parameters of the teacher network remain frozen and only participate in forward propagation to obtain soft labels $\boldsymbol{p}_t$,without engaging in gradient updates. The student network performs forward computations to generate logits z_t and probability outputs $\boldsymbol{p}_s$.Through backpropagation, it optimizes the total loss $\mathcal{L}_{\text{total}}$ to update the student network parameters θ_{s} :

$$\theta_s \leftarrow \theta_s - \eta \nabla_{\theta_s} \mathcal{L}_{\text{total}} \tag{6}$$

III. EXPERIMENTS

A. datasets

We evaluate our method on three widely-used motor imagery EEG datasets: the BCI Competition IV Dataset 2a[4], the BCI Competition IV Dataset 2b[5], and the Korea University OpenBMI Motor Imagery Dataset[6]. These datasets employ distinct acquisition devices, experimental paradigms, numbers of subjects, and sample sizes, thereby providing a comprehensive framework to validate the adaptability of our method across heterogeneous experimental conditions.

1) Dataset I : The BCI Competition IV Dataset 2b (2008) contains EEG recordings from nine healthy subjects performing left-hand and right-hand motor imagery tasks. The experiment comprises two separate sessions per subject, each consisting of five runs with 48 trials per run, yielding a total of 240 trials per session. EEG signals were recorded using three Ag/AgCl electrodes positioned at standard locations C3, Cz, and C4 according to the 10-20 international system, referenced to the left earlobe with a ground electrode placed on the right earlobe. Data acquisition was conducted at a sampling rate of 250 Hz with 0.5-100 Hz bandpass filtering, supplemented by a 50 Hz notch filter to eliminate power-line interference. To mitigate ocular artifacts, monopolar electrooculography (EOG) signals were simultaneously recorded.

2) Dataset II: The BCI Competition IV Dataset 2a (2008) comprises EEG recordings from nine subjects, each participating in two experimental sessions conducted on separate days. The study employs a cue-guided brain-computer interface paradigm involving four motor imagery tasks: left hand, right hand, feet, and tongue movement imagination. Each session consists of six runs, with 48 trials per run (12 trials per class),

yielding a total of 288 trials per session. EEG signals were acquired through 22 Ag/AgCl electrodes arranged according to the international 10-20 system, sampled at 250 Hz and bandpass filtered between 0.5 and 100 Hz, supplemented by a 50 Hz notch filter to suppress mains interference. Additionally, three monopolar electrooculography (EOG) channels were recorded to facilitate subsequent ocular artifact removal procedures.

3) Dataset III: The Korea University OpenBMI Motor Imagery Dataset comprises EEG recordings from 54 healthy subjects performing two distinct motor imagery (MI) tasks: left-hand and right-hand movement imagination. Each subject completed two experimental sessions (training and testing phases), with each phase containing 100 trials balanced equally between both hand conditions. Original EEG signals were acquired using 62 Ag/AgCl electrodes sampled at 1000 Hz, referenced to the nose tip with a ground electrode at AFz. Following 8-30 Hz bandpass preprocessing, our study specifically utilized 20 electrodes over the sensorimotor cortex (FC-5/3/1/2/4/6, C-5/3/1/z/2/4/6, and CP-5/3/1/z/2/4/6) to optimize task-related signal analysis while maintaining computational efficiency.

B. Baseline Models

To validate the synergistic efficacy of our knowledge distillation strategy integrated with lightweight architecture, we conducted a comparative analysis against five established baseline models spanning conventional and state-of-the-art approaches. The selected benchmark models encompass:

1) EEGNet : EEGNet is a compact convolutional neural network specifically designed for EEG signal processing[7], which innovatively employs depthwise separable convolutions to structure the model into two core modules: spatiotemporal feature extraction and temporal feature fusion. The spatiotemporal module first captures frequency band features through temporal convolution, followed by spatial convolution to optimize signal characteristics across brain regions. The temporal fusion module utilizes separable convolutions to compress parameters while effectively integrating temporal information. Notably, EEGNet achieves a groundbreaking reduction in parameter count to 1/100 of conventional CNN models, significantly enhancing computational efficiency without compromising feature representation capabilities.

2) Shallow ConvNet : Shallow ConvNet is a shallow architecture designed for raw EEG signals[8]. The first two layers perform temporal and spatial convolutions, using a larger kernel in the temporal convolution to increase the range of transformation. After the two convolutions, square nonlinearity, mean pooling, and a logarithmic activation function are introduced, ultimately outputting classification results.

3) FBCnet : FBCNet is a compact, neurophysiologically inspired convolutional neural network (CNN) architecture designed for motor imagery (MI) classification[9]. Drawing from the concept of FBCSP, FBCNet adopts a multi-view data representation approach. Specifically, it generates multi-view data by applying 9 narrowband Chebyshev Type II bandpass filters, each with a bandwidth of 4 Hz, covering a frequency range of 4 to 40 Hz (including 4-8 Hz, 8-12 Hz, ..., 36-40 Hz), with a transition bandwidth of 2 Hz and a stopband ripple of 30 dB. These filters assist in extracting discriminative features from multiple frequency bands of EEG signals. Moreover, FBCNet introduces a novel variance layer to effectively extract temporal information from EEG signals, further enhancing classification performance.

4) LightConvNet[10] : LightConvNet employs the same multi-view EEG data representation processing method as FBCNet. The first two convolution modules are consistent with those in FBCNet, but LightConvNet is designed with a novel temporal attention module to capture temporal dependencies between features extracted from different time periods (variance layer), enhancing the representation of features in the temporal dimension.

5) EEG Conformer[11] : EEG Conformer uses a 6th order Chebyshev filter to retain the frequency band from 4 to 40 Hz. It captures local features through one-dimensional temporal and spatial convolution layers, and extracts global correlations through self-attention mechanisms to improve the classification accuracy of EEG signals.Through optimizing the modular design of this architecture, our study achieves significant improvement in computational efficiency while maintaining model performance, ultimately establishing the lightweight EEG Lightformer framework.

C. Experimental Details

To validate the effectiveness of the proposed knowledge distillation architecture in motor imagery electroencephalogram (MI-EEG) signal decoding, this study conducted cross-session experiments on three publicly available benchmark datasets. In the experimental design, all trials from Session 1 of each subject were used as the training set, while all trials from Session 2 served as the test set. Notably, the acquisition intervals between two sessions exceeded 24 hours across all three datasets, a setup that effectively evaluates the model's adaptability to time-varying characteristics of neural signals across sessions.

For network parameter configuration, the temperature coefficient $\tau = 3$ was set to optimize the smoothness of class probability distributions, while the distillation loss weight $\alpha = 0.3$ regulated the knowledge transfer intensity between teacher and student models. Regarding input normalization for the teacher model EEGPT, cubic spline interpolation was employed to adjust the spatial dimensions of raw EEG signals.

All comparative experiments were implemented using the PyTorch deep learning framework, with model training executed on an NVIDIA GeForce RTX 4070 GPU (8GB memory).

IV. RESULTS

As shown in Tables I to III, our proposed EEG-Lightformer-KD model exhibits the following characteristics in the within-subject cross-session EEG classification task:

TABLE I
COMPARATIVE PERFORMANCE (ACCURACY AND PARAMETER SIZE) ON BCIC-IV-2A

datasets	methods	S01	S02	S03	S04	S05	S06	S07	S08	S09	Average(%)	Parm(K)
BCIC-IV-2a	EEGnet	79.51	**61.11**	88.54	71.53	**71.18**	59.03	71.53	80.56	75.35	73.15	1.7
	FBCnet	85.42	60.42	90.63	76.39	74.31	53.82	84.38	79.51	80.90	76.20	11
	EEG-Conformer	87.85	60.07	**93.40**	81.25	52.78	62.50	**93.06**	**87.84**	**88.54**	78.59	789
	Ours(EEG-Lightformer)	85.76	53.47	89.93	74.65	47.22	58.33	88.54	85.42	84.03	74.15	127
	Ours(EEG-Lightformer-KD)	**89.93**	60.07	93.06	79.51	55.56	**62.50**	91.66	87.50	88.19	**78.67**	127

TABLE II
COMPARATIVE PERFORMANCE (ACCURACY AND PARAMETER SIZE) ON BCIC-IV-2B

datasets	methods	S01	S02	S03	S04	S05	S06	S07	S08	S09	Average(%)	Parm(K)
BCIC-IV-2b	Shallownet	69.38	58.93	**74.06**	93.75	85.62	68.44	80.62	86.87	83.44	77.90	39
	EEGnet	75.94	57.64	58.43	98.13	81.25	88.75	84.06	93.44	89.69	80.48	1.2
	EEG-Conformer	**81.25**	62.50	60.31	**98.75**	85.32	**90.31**	87.50	94.69	**91.88**	**83.61**	839
	Ours(EEG-Lightformer)	77.50	63.21	59.06	97.81	79.06	90.00	82.19	94.38	90.31	81.50	205
	Ours(EEG-Lightformer-KD)	77.81	**63.21**	58.75	98.44	**85.94**	90.00	**88.75**	**95.00**	91.25	83.24	205

TABLE III
COMPARATIVE PERFORMANCE (ACCURACY AND PARAMETER SIZE) ON OPENBMI

datasets	methods	Accuracy(%)	Parm(K)
OpenBMI	LightConvNet	65.31	11
	FBCnet	67.19	8
	EEG-Conformer	70.12	786
	Ours(EEG-Lightformer)	68.23	127
	Ours(EEG-Lightformer-KD)	**70.36**	127

A. Performance evaluation of the four-class classification task (BCIC-IV-2A dataset)

The EEG-Lightformer-KD model achieved a superior average accuracy of 78.67% (Table I), outperforming all baseline models while requiring only 1/6 of the parameters (127K vs. 789K) compared to EEG-Conformer (78.59%), demonstrating a 0.08% accuracy improvement. It exhibited significant performance advantages over EEG-Conformer in specific subsets: S01 (89.93% vs. 87.85%) and S05 (55.56% vs. 52.78%), while maintaining minimal performance gaps ($< 0.5\%$) in S03 (93.06% vs. 93.40%) and S09 (88.19% vs. 88.54%). These results confirm that the lightweight design preserves core feature extraction capabilities. Compared to its non-distilled counterpart EEG-Lightformer (74.15%), the knowledge distillation (KD) technique delivered an absolute improvement of +4.52%, with particularly notable gains in low-performance subsets: S02 (60.07% vs. 53.47%), S05 (55.56% vs. 47.22%), and S06 (62.50% vs. 58.33%).

B. Performance evaluation of the binary classification task (BCIC-IV-2B and OpenBMI datasets)

a) BCIC-IV-2B: In the BCIC-IV-2B dataset, the EEG-Lightformer-KD achieves a marginally lower mean accuracy of 83.24% (Table II) compared to EEG-Conformer (83.61%), yet still surpasses classical models such as Shallow ConvNet and EEGNet. Notably, it demonstrates superior performance in specific subsets: S05 (85.94% vs. 85.32%), S07 (88.75% vs. 87.50%), and S08 (95.00% vs. 94.69%). Remarkably, this competitive performance is achieved with only 205K parameters (24.4% of EEG-Conformer's parameter count), while also exhibiting a 1.74% absolute accuracy improvement over its non-distilled counterpart (81.50%).

b) OpenBMI: In the OpenBMI dataset, the EEG-Lightformer-KD achieves a superior accuracy of 70.36%, outperforming EEG-Conformer (70.12%), while validating the generalization capability of its lightweight design in large-scale subject cohorts (n=54) with only 127K parameters (16.1% of EEG-Conformer's parameter count). It demonstrates substantial advancements over FBCNet (67.19%) and LightConvNet (65.31%), achieving absolute improvements of 3.17%-5.05%. Furthermore, the model exhibits a 2.13% accuracy enhancement compared to its non-distilled counterpart (68.23%), underscoring the efficacy of knowledge distillation in cross-subject scenarios.

C. Parameter Efficiency

In the BCIC-IV-2A dataset, the EEG-Lightformer-KD achieves an accuracy of 78.67% with 127K parameters, outperforming EEGNet (1.7K parameters/73.15% accuracy) by a 5.52% absolute improvement and FBCNet (11K parameters/76.20%) by 2.47%, thereby validating the advantage of its "moderate parameter scale and efficient architecture design". This trend is consistently observed in the BCIC-IV-2B and OpenBMI datasets. Notably, the knowledge distillation (KD) technique delivers a 4.52% accuracy gain in the four-class classification task (BCIC-IV-2A), significantly surpassing its benefits in binary classification tasks (+1.74% on BCIC-IV-2B and +2.13% on OpenBMI). This disparity suggests that the teacher model (EEG-Conformer) provides richer inter-class discriminative knowledge in multi-class scenarios, leading to higher distillation efficacy.

V. CONCLUSION

This paper proposes a knowledge distillation framework based on an EEGPT large model and a lightweight convolutional Transformer. By freezing teacher model parameters and designing a lightweight student model (EEG Lightformer), combined with temperature-smoothed KL divergence distillation and cross-entropy supervision, the framework achieves effective decoding on three major motor imagery datasets including BCIC-IV-2A. Experimental results demonstrate that the proposed framework outperforms mainstream models on multiple public datasets with significantly reduced parameter scale. The distillation-enhanced lightweight design exhibits superior discriminative knowledge transfer capabilities in multi-classification tasks while maintaining stable generalization performance in binary classification scenarios. This work provides a practical solution for resource-constrained brain-computer interface systems that effectively balances computational efficiency and model performance.

ACKNOWLEDGMENT

The authors are grateful to the anonymous reviewers for their constructive comments.

REFERENCES

[1] G. Wang, W. Liu, Y. He, C. Xu, L. Ma, and H. Li, "Eegpt: Pretrained transformer for universal and reliable representation of eeg signals," *Advances in Neural Information Processing Systems*, vol. 37, pp. 39 249–39 280, 2024.

[2] R. Tang, Y. Lu, L. Liu, L. Mou, O. Vechtomova, and J. Lin, "Distilling task-specific knowledge from bert into simple neural networks," *arXiv preprint arXiv:1903.12136*, 2019.

[3] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[4] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "Bci competition 2008–graz data set a," *Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology*, vol. 16, no. 1-6, p. 1, 2008.

[5] R. Leeb, C. Brunner, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "Bci competition 2008–graz data set b," *Graz University of Technology, Austria*, vol. 16, pp. 1–6, 2008.

[6] M.-H. Lee, O.-Y. Kwon, Y.-J. Kim, H.-K. Kim, Y.-E. Lee, J. Williamson, S. Fazli, and S.-W. Lee, "Eeg dataset and openbmi toolbox for three bci paradigms: An investigation into bci illiteracy," *GigaScience*, vol. 8, no. 5, p. giz002, 2019.

[7] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.

[8] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human brain mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[9] R. Mane, E. Chew, K. Chua, K. K. Ang, N. Robinson, A. P. Vinod, S.-W. Lee, and C. Guan, "Fbcnet: A multi-view convolutional neural network for brain-computer interface," *arXiv preprint arXiv:2104.01233*, 2021.

[10] X. Ma, W. Chen, Z. Pei, J. Liu, B. Huang, and J. Chen, "A temporal dependency learning cnn with attention mechanism for mi-eeg decoding," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 3188–3200, 2023.

[11] Y. Song, Q. Zheng, B. Liu, and X. Gao, "Eeg conformer: Convolutional transformer for eeg decoding and visualization," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 710–719, 2022.