M-FANet: Multi-Feature Attention Convolutional Neural Network for Motor Imagery Decoding

Yiyang Qin, Banghua Yang, Sixiong Ke, Peng Liu, Fenqi Rong, and Xinxing Xia

***Abstract*— Motor imagery (MI) decoding methods are pivotal in advancing rehabilitation and motor control research. Effective extraction of spectral-spatial-temporal features is crucial for MI decoding from limited and low signal-to-noise ratio electroencephalogram (EEG) signal samples based on brain-computer interface (BCI). In this paper, we propose a lightweight Multi-Feature Attention Neural Network (M-FANet) for feature extraction and selection of multi-feature data. M-FANet employs several unique attention modules to eliminate redundant information in the frequency domain, enhance local spatial feature extraction and calibrate feature maps. We introduce a training method called Regularized Dropout (R-Drop) to address training-inference inconsistency caused by dropout and improve the model's generalization capability. We conduct extensive experiments on the BCI Competition IV 2a (BCIC-IV-2a) dataset and the 2019 World robot conference contest-BCI Robot Contest MI (WBCIC-MI) dataset. M-FANet achieves superior performance compared to state-of-the-art MI decoding methods, with 79.28% 4-class classification accuracy (kappa: 0.7259) on the BCIC-IV-2a dataset and 77.86% 3-class classification accuracy (kappa: 0.6650) on the WBCIC-MI dataset. The application of multi-feature attention modules and R-Drop in our lightweight model significantly enhances its performance, validated through comprehensive ablation experiments and visualizations.**

***Index Terms*— Brain–computer interface, motor imagery, convolutional neural networks, multi-feature attention.**

Manuscript received 11 July 2023; revised 17 November 2023 and 1 January 2024; accepted 6 January 2024. Date of publication 9 January 2024; date of current version 19 January 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFC3602700 and Grant 2022YFC3602703, in part by the National Natural Science Foundation of China under Grant 62376149, in part by the National Defense Basic Scientific Research Program of China (Defense Industrial Technology Development Program) under Grant JCKY2021413B005, in part by the Shanghai Major Science and Technology Project under Grant 2021SHZDZX, and in part by the Shanghai Industrial Collaborative Technology Innovation Project under Grant XTCX-KJ-2022-2-14. *(Corresponding author: Banghua Yang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted in part by the Institute for Knowledge Discovery, Graz University of Technology (for approving the BCIC-IV-2a dataset), and in part by the Tsinghua University Medical Ethics Committee (for approving the WBCIC-MI dataset).

Yiyang Qin, Sixiong Ke, Peng Liu, Fenqi Rong, and Xinxing Xia are with the School of Mechanical and Electrical Engineering and Automation, Shanghai University, Shanghai 200444, China.

Banghua Yang is with the School of Mechatronic Engineering and Automation, School of Medicine, Research Center of Brain Computer Engineering, Shanghai University, Shanghai 200444, China, and also with the Engineering Research Center of Traditional Chinese Medicine Intelligent Rehabilitation, Ministry of Education, Shanghai 201203, China (e-mail: yangbanghua@shu.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2024.3351863

I. Introduction

BRAIN-COMPUTER Interface (BCI) constitutes a communication conduit connecting the nervous system to the external environment [1]. Motor Imagery (MI) entails the internal rehearsal of a movement before its execution [2], and it serves as a cornerstone of BCI research [3]. Electroencephalography (EEG), a non-invasive technique renowned for its cost-effectiveness and convenience, facilitates the recording of neural activity with high temporal resolution [4], [5]. When participants visualize moving parts of the body, the phenomenon that specific areas of the brain experience energy changes called event-related desynchronization/synchronization (ERD/ERS) could be recorded via EEG and then used to discriminate motor intent [6], [7]. MI-based BCI systems have advanced notably, enabling the control of exoskeletons [8], [9] and cursors [10]. Furthermore, the integration of MI with virtual reality technology [11] has shown promising prospects for stroke rehabilitation [12]. The success of these systems hinges on high-performance MI decoding methods [13]. However, improving the classification performance of spontaneous MI, as compared to other BCI paradigms reliant on external stimuli, such as event-related potential [14] and steady-state visual evoked potential [15], poses a formidable challenge due to factors like a low signal-to-noise ratio and cross-subject variability [16].

The EEG signals, distinguished by their excellent temporal resolution, encompass rich spectral and spatial features [4]. The integration of these temporal-spatial-spectral features can bolster classification accuracy. Traditional machine learning methods generally extract neurophysiological features from MI-induced EEG signals. For instance, the Common Spatial Pattern (CSP) is a spatial-domain filtering feature extraction method that distills spatially distributed components from multichannel EEG signals corresponding to each class [17]. The Filter Bank CSP (FBCSP) method enhances CSP performance by segmenting EEG data into multiple frequency bands [18]. However, traditional machine learning methods highly rely on handcrafted features and, thereby, may fail to exploit the latent information embedded in the data fully.

With the advent of deep learning methods, a wealth of novel methods have emerged for EEG signal classification [19], such as Convolutional Neural Networks (CNN) [20]. DeepConvNet [21], proposed by Schirrmeister et al., employed multiple convolutional layers with larger temporal and spatial feature extraction kernels. Sakhavi et al. [22]

utilized FBCSP for feature extraction, followed by CNN-based classification. These comparatively larger networks have demonstrated remarkable classification performance compared to traditional methods. Nevertheless, larger models pose challenges in hyperparameter selection, as they are more prone to overfitting and demand longer training times.

Recently, novel lightweight deep-learning networks based on CNN have achieved better performance. These networks are efficient and outperform traditional deep learning models in accuracy when applied to MI-BCIs. FBCNet [23], proposed by Mane et al., initiated with spectral filtering of raw EEG data, followed by a spatial convolutional layer, and eventually computed temporal variance. FBCNet effectively encapsulates time-frequency domain feature information. Lawhern et al. [24] introduced a compact network, EEGNet, employing depthwise and separable convolution for spatial-temporal feature extraction. EEGNet wholly learned frequency filters via deep learning. However, hand-crafted frequency bands also encompass a substantial amount of frequency domain information, necessitating their effective combination with automatic learning frequency filters. In addition, these networks typically utilize a convolution kernel identical in size to the number of electrodes for direct spatial feature extraction. Nevertheless, considering that MI primarily activates particular brain regions [25], [26], it is crucial to place an additional focus on these responsive areas rather than solely relying on global spatial feature extraction.

The highly regarded attention mechanisms, which have gained significant attention in natural language processing and image processing, have recently been successfully applied to MI-EEG decoding. The MI-DABAN [27], proposed by Li et al., leveraged domain-specific attention modules based on CNN to capture critical features in both the source and target domains, achieving higher transfer performance. Song et al. introduced the Conformer [28], which applied Transformer modules based on self-attention to MI decoding, achieving strong classification performance. Self-attention modules have the advantage of globally extracting information from the entire sequence without being constrained by distance, and they apply weights separately. However, self-attention modules come with drawbacks, such as a high number of parameters and demanding computational resources. Moreover, they are not tailored for EEG data, leading to suboptimal resource utilization.

EEG signals often suffer from limited training samples, leading to overfitting. Besides adopting more lightweight network structures, exploring EEG signal-specific training methodologies is essential. Recently, Regularized Dropout (R-Drop) [29], a novel training method for tackling deep learning overfitting issues, has shown promise in natural language and image processing. R-Drop enhances the model's generalization ability by narrowing the inconsistency between the complete model and sub-models, thereby boosting the final performance of the model. This training method is apt for MI task classification networks prone to overfitting.

To tackle the above issues, we propose a lightweight network incorporating multi-feature attention named Multi-Feature Attention Neural Network (M-FANet). This model comprises multiple convolutional layers for feature extraction and several distinct attention modules for calibrating relevant information from three perspectives: frequency, local space, and feature map. The multi-feature attention modules adaptively extract valid frequency band features specific to individual subjects, heightening the perception of the spatial response of MI-related channel groups and filtering out redundant feature maps. Additionally, we employ R-Drop as a training method, reducing the discrepancy between various sub-models resulting from dropout by constraining the probability distribution differences of their outputs, thereby enhancing the performance of the complete model. We conduct detailed comparative experiments and ablation studies on two MI datasets to demonstrate the significant performance of M-FANet.

The contributions are summarized as follows:

- We propose a lightweight network named Multi-Feature Attention Neural Network (M-FANet) by extracting a broader range of features from EEG signals and using multiple attention modules to effectively leverage them to enhance the classification performance of MI tasks.
- We introduce R-Drop as a model training method to mitigate overfitting, an issue stemming from the limited and noisy nature of EEG samples.
- We conduct experiments on two datasets to evaluate the effectiveness and superiority of the proposed M-FANet against state-of-the-art MI decoding methods. In addition, we delve into the impact of the regularization term introduced by R-Drop on the network performance.

II. Related Works

A. Traditional Methods

Many traditional machine learning methods for MI-EEG classification have been proposed. Barachant et al. introduced a brain-computer interface classification framework based on MI, employing Riemannian geometry for direct classification using spatial covariance matrices as EEG signal descriptors [30]. CSP [17] was one of the most effective and popular feature extraction methods in EEG signal processing. It differentiated brain activities under various tasks by analyzing the spatial distribution of multi-channel EEG signals.

However, the effectiveness of CSP was greatly influenced by the selected frequency bands as well as selected features. To address this issue, mutual information-based selection of optimal spatial–temporal patterns (OSTP) [31] utilized a feature selection method based on mutual information to optimize the CSP method, facilitating the automatic selection of frequency bands and time segments for CSP filtering. Ozdenizci et al. introduced a method that leverages information theoretic learning to enhance neural feature interpretation and overcome the constraints of traditional feature ranking and selection techniques in model training [32].

FBCSP [18] divided the EEG data into multiple frequency bands and applied CSP to these segments. Subsequently, a feature selection algorithm was employed to automatically select features specific to each subject, thereby significantly enhancing classification accuracy. The FBCSP method stands as one of the most triumphant in the field, being extensively utilized for comparative method analysis.

B. Deep Learning Methods

In recent years, a surge of deep learning methods has demonstrated promising advancements in the sphere of MI-BCI [19]. Among these, EEGNet [24] has emerged as a notably compact deep learning network. By leveraging depthwise and pointwise convolutions, EEGNet significantly reduced the number of parameters, thereby ensuring a lightweight network structure. The model initiates by convolution along the temporal dimension, broadening the feature maps. Then, it carried out a depthwise convolution along the spatial dimension, compressing the entirety of the spatial information. Ultimately, it used depthwise separable convolution to extract features.

Compared to EEGNet, FBCNet manually extracted temporal features by calculating variance rather than employing temporal filters. Moreover, FBCNet utilized a bank of band-pass filters to segment the data into multiple frequency bands. Owing to their reliable performance, they serve as baseline methods in our experiments.

Attention mechanisms are being widely applied in EEG decoding. More recently, Conformer [28], which leveraged self-attention modules to learn global temporal features, has reached the highest classification accuracy on the BCIC-IV-2a dataset. MI-DABAN approached EEG feature transfer with the design of compact attention modules, leading to improved classification performance. The advantage of attention modules is that their design basis is the original input characteristics, which is suitable for MI-EEG, ensuring both lightweight implementation and effective feature extraction. Inspired by these observations and considerations, we propose the Multi-Feature Attention Neural Network (M-FANet), aimed at enhancing the extraction of features related to the principles of MI.

III. Methods

A. Proposed M-FANet Architecture

In the realm of deep learning methods employed for MI, EEGNet is known for its subtle architecture and lucid explanation. However, it does present limitations in robust feature extraction capabilities. Building upon the foundation provided by EEGNet, we propose M-FANet, which firstly incorporates data selectively from hand-crafted frequency bands by utilizing a frequency band attention module. Then, the M-FANet's capacity for extracting local spatial features is enhanced by introducing a local spatial attention module.

The chosen kernel length for the following temporal convolution layer, K_T, is equivalent to one-eighth of the data sampling rate, outputting multi-feature maps of size F_T, which is numerically equivalent to the number of temporal filters. A depthwise convolution follows, where each kernel solely connects to one preceding feature map. This depthwise convolution layer reduces the number of trainable parameters, enabling the integration of all spatial information, where D is the spatial filter multiplier and the number of spatial filters is $D * F_T$. By adjusting the values of D and F_T, the number of feature maps can be modified, allowing M-FANet to adapt to various datasets. Furthermore, a Squeeze-and-Excitation Block (SEBlock) [33] also is introduced, facilitating automatic feature calibration and allowing the network to prioritize feature maps dynamically.

Following the introduction of the feature map attention mechanism, a separable convolution comprising depthwise convolution and pointwise convolution is employed. Specifically, this first extracts temporal features within each feature map, followed by the extraction of inter-feature map features. This approach enables the network to capture local temporal patterns within individual feature maps and global relationships between different feature maps with fewer parameters. Ultimately, a convolution layer is used in the classification stage to aggregate the features.

The visual representation and comprehensive description of our M-FANet method are presented in Fig.1 and Table I, respectively.

1) Frequency Band Attention Module: In machine learning algorithms for MI, the band-pass filter parameters significantly influence the decoding results, which suggests that while the frequency domain is replete with valuable information, it is difficult to extract it accurately. To capture the pertinent information within the frequency domain, we design a frequency band attention module based on the attention mechanism, facilitating adaptive frequency filtering.

A single-trial raw EEG sample can be represented as $X \in R^{C\times T}$, where C denotes the number of EEG channels, and T is the time points. Initially, we implement multiple band-pass filters to the raw EEG data utilizing the Chebyshev Type II filter with a stopband ripple of -30dB and a fixed step size of L (set at 4Hz here). This yields multi-band EEG data X_{MB}:

$$X_{\mathrm{MB}}(n) = X * h(n) \in R^{N_b\times C\times T} \tag{1}$$

where $h(n)$ represents the bandpass filter corresponding to the n_{th} frequency sub-band, N_b is the number of sub-band.

Then, we employ a pointwise convolution to amalgamate the multi-band information. This process enables the network to harness the complementary information from each frequency band. Simultaneously, an adaptive weight is designated to each frequency band to attenuate noise in redundant frequency bands and amplify efficient information in other frequency bands. After fusing frequency band information, we derive X_{FS}:

$$X_{\mathrm{FS}}(i, j) = \sum_{n=1}^{N_b} X_{MB}(n, i, j) \cdot w(n) \in R^{C\times T} \tag{2}$$

where $w(n)$ represents the feature weight corresponding to the n_{th} sub-band.

2) Local Spatial Attention Module: Based on the neuroscientific prior knowledge regarding MI [25], [26], it's known that MI predominantly activates motor-related brain regions, especially near the electrodes C3, Cz, and C4. Hence, when extracting features in the spatial dimension, channels shouldn't be treated as equal, but focus should be placed on areas relevant to MI. Consequently, leveraging the attention mechanism,

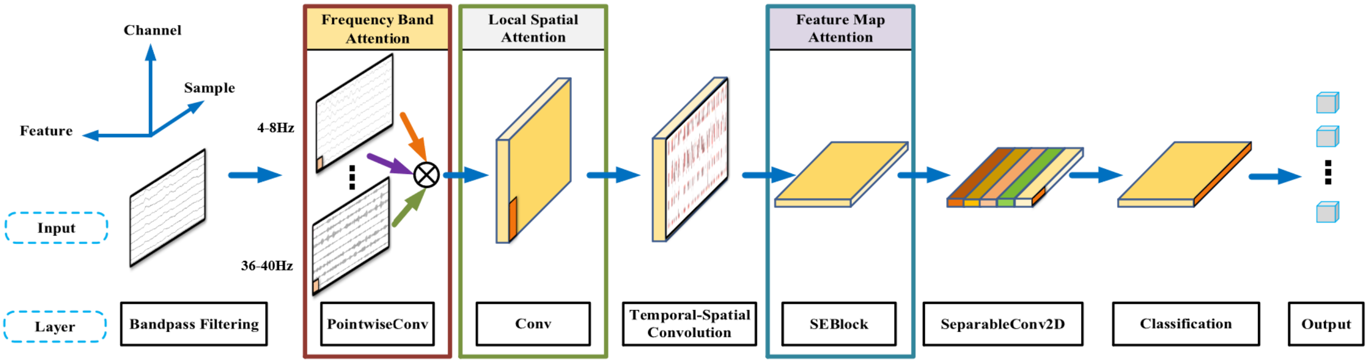


Fig. 1. The architecture of our proposed Multi-Feature Attention Convolutional Neural Network (M-FANet) employs convolution for feature extraction and classification, augmented with three distinct attention modules from the perspectives of frequency domain, spatial domain, and feature maps.

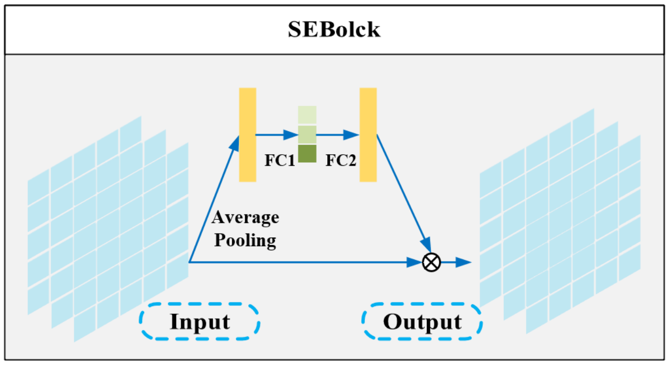


Fig. 2. The structure of SEBlock.

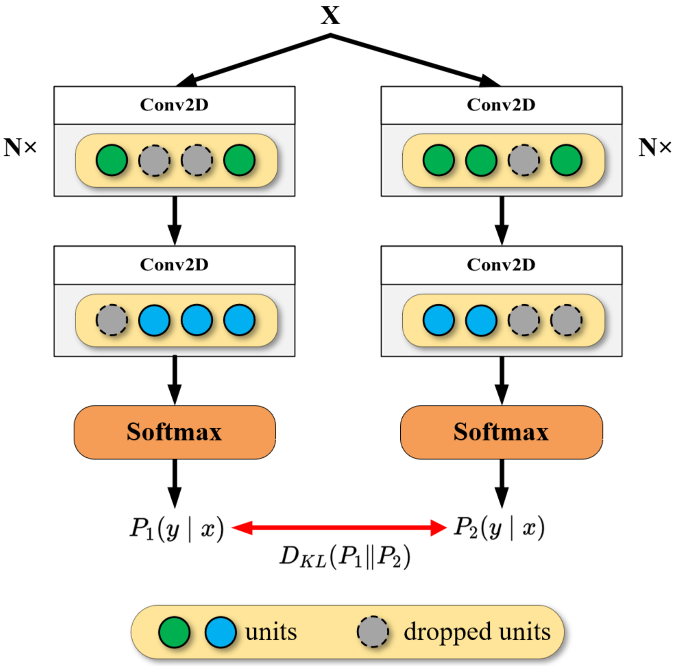


Fig. 3. The overall framework of R-Drop. We take CNN structure for illustration. The network conducts dual forward passes and utilizes KL divergence to enforce consistency between the results of these passes.

we design a local spatial attention module to extract effective spatial information from local electrode groups.

We employ a convolution layer to extract local spatial information by using a small kernel size (K_s, 1) and stride to enable the detailed extraction of local spatial features from EEG data, convolving solely along the spatial dimension. After local spatial feature extraction, we derive X_{LS}:

$$X_{\mathrm{LS}}(i, j) = \sum_{m=0}^{K_s-1} X_{\mathrm{FS}}(i+m, j) \cdot w_s(m) + b \in R^{C\times T} \quad (3)$$

where w_s represents the weight matrix of the kernel, m denotes the index of the weights within the kernel, and b is the bias term.

The small convolution kernel facilitates the focus of M-FANet on local receptive fields, promoting the identification of intricate spatial patterns related to various MI tasks. In terms of MI EEG signals, this results in attention being directed towards particular sets of electrodes within the motor area, as evident in the output feature maps.

3) Feature Map Attention Module: A SEBlock, as shown in Fig. 2, is utilized to apply weights to the feature maps. Consider input with F feature maps, C channels and T time points that are represented as $X_{SE} \in R^{F\times C\times T}$. We squeeze the global spatial and temporal information into a feature descriptor by using global average pooling to generate feature-wise statistics:

$$z_i = \frac{1}{C\times T}\sum_{j=1}^{C}\sum_{k=1}^{T} X_{SE\,i,j,k}, \quad i = 1, 2, \cdots, F \quad (4)$$

Then, two fully connected layers are used to learn the nonlinear relations between different feature maps:

$$W = \sigma\left(W_{F2}, \delta\left(W_{F1}, Z\right)\right) \quad (5)$$

where W is scale vector, $Z = \{z_1, z_2, \cdots, z_F\}$ denotes the feature-wise statistics vector, $W_{F1} \in R^{\frac{F}{r}\times F}$ is a weight matrix of the front fully connected layer whose reduction ratio is r, a hyperparameter, and $W_{F2} \in R^{F\times\frac{F}{r}}$ is a weight matrix of the other fully connected layer which elevates the feature back to its original dimension, $\delta(\cdot)$ is rectified linear unit (ReLU) activation function, $\sigma(\cdot)$ is the sigmoid activation function.

TABLE I
THE ARCHITECTURE OF M-FANET

Block	Layer	#Filters	Kernel Size	Output	Options
Frequency Band Attention	Input			$(C, T, 1)$	
	Band Pass Filter			(C, T, N_b)	
	PointwiseConv2D		$(1,\ 1)$	$(C, T, 1)$	
Local Spatial Attention	Conv2D		$(K_s,\ 1)$	$(C, T, 1)$	padding = same
Temporal-Spatial Convolution	Conv2D	F_T	$(1,\ K_T)$	(C, T, F_T)	padding = same
	BatchNorm2D			(C, T, F_T)	
	DepthwiseConv2D	$D * F_T$	$(C,\ 1)$	$(1, T, D * F_T)$	max_norm = 1, groups = F_T
	AveragePool2D		$(1, 4)$	$(1, T\ //4, D * F_T)$	
Feature Map Attention	SEBlock			$(1, T\ //4, D * F_T)$	
Temporal Convolution	SeparableConv2D	$D * F_T$	$(1, 25)$	$(1, T\ //4, D * F_T)$	padding = same, groups = $D * F_T$
	AveragePool2D		$(1, 8)$	$(1, T\ //32, D * F_T)$	
Classifer	Conv2D	N_c	$(1, T\ //32)$	$(1, 1, N_c)$	

B. R-Drop Module

The dropout technique, randomly dropping a fraction of units with a given probability p (set at 0.5 here) during forward propagation, suppresses overfitting [34]. During the training process, dropout randomly disables some units, thereby enhancing the generalization ability of the model. However, each training iteration uses a sub-model due to the dropout while using a complete model during inference. This discrepancy between the models used in training and inference can lead to a circumstance where the model performs well on the training set but still demonstrates an inescapable gap on the test set. This issue is especially pronounced in EEG signal datasets with small sample sizes highly susceptible to overfitting.

To address the abovementioned problem, we chose R-Drop as the training method, which constrains the different output predictions of sub-models caused by dropout [29]. The core idea of R-Drop is adding a regularization term to the loss function, which mitigates the discrepancies among different sub-models, narrowing the gap between the complete model and its sub-models. For each sample, the Kullback-Leibler (KL) divergence between the output probability distributions of two randomly chosen sub-models is computed and added to the loss function. The KL divergence encourages sub-models to align as closely as possible in their output probability distributions, narrowing the inconsistency between the complete model and the sub-models.

The framework of R-Drop is shown in Fig. 3. We obtain two distributions of the model predictions by performing forward propagation twice, denoted as $\mathcal{P}_1(y_i \mid x_i)$ and $\mathcal{P}_2(y_i \mid x_i)$, which are different from each other because the dropout operator randomly dropped units. Next, we calculate the bidirectional KL divergence between these two sub-model outputs to regularize the model output, where the input data pair (x_i, y_i) is the same:

$$\mathcal{L}_{KL}^{i} = \frac{1}{2}\Big(\mathcal{D}_{KL}\big(\mathcal{P}_1(y_i \mid x_i)\|\mathcal{P}_2(y_i \mid x_i)\big) + \mathcal{D}_{KL}\big(\mathcal{P}_2(y_i \mid x_i)\|\mathcal{P}_1(y_i \mid x_i)\big)\Big) \tag{6}$$

With the cross-entropy loss to minimize the classification error between the predicted labels and the ground-truth labels:

$$\mathcal{L}_{CE}^{i} = \frac{1}{2}\Big(\mathcal{D}_{CE}\big(y_i\|\mathcal{P}_1(y_i \mid x_i)\big) + \mathcal{D}_{CE}\big(y_i\|\mathcal{P}_2(y_i \mid x_i)\big)\Big) \tag{7}$$

The final training objective is to minimize the following:

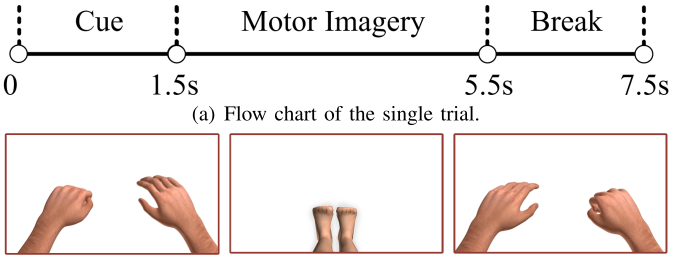
$$\begin{aligned} \mathcal{L}^{i} &= \mathcal{L}_{CE}^{i} + \alpha \cdot \mathcal{L}_{KL}^{i} \\ &= \frac{1}{2}\Big(\mathcal{D}_{CE}\big(y_i\|\mathcal{P}_1(y_i \mid x_i)\big) + \mathcal{D}_{CE}\big(y_i\|\mathcal{P}_2(y_i \mid x_i)\big)\Big) \\ &\quad + \frac{\alpha}{2}\Big(\mathcal{D}_{KL}\big(\mathcal{P}_1(y_i \mid x_i)\|\mathcal{P}_2(y_i \mid x_i)\big) \\ &\quad + \mathcal{D}_{KL}\big(\mathcal{P}_2(y_i \mid x_i)\|\mathcal{P}_1(y_i \mid x_i)\big)\Big) \end{aligned} \tag{8}$$

where the KL divergence is represented by $\mathcal{D}_{KL}(\mathcal{P}_1\|\mathcal{P}_2)$ between two probability distributions $\mathcal{P}_1$ and $\mathcal{P}_2$, the cross-entropy loss is represented by $\mathcal{D}_{CE}\big(Y\|\mathcal{P}(Y \mid X)\big)$ between the distribution of the ground-truth label Y and the predicted labels probability distribution $\mathcal{P}(Y \mid X)$, α is the coefficient weight to control the degree of regularization, $\mathcal{L}^{i}$ is used for gradient updates. We will discuss the impact of α in the next chapter.

IV. EXPERIMENTS AND RESULTS

A. Datasets

1) Dataset I: The 2008 BCI Competition (BCIC) IV-2a EEG data set is also used to evaluate method performance [35]. It consists of EEG data from nine subjects on four types of MI tasks: left hand, right hand, foot, and tongue. Two recording sessions are collected on separate days, utilizing twenty-two Ag/AgCl electrodes with a sampling rate of 250 Hz. Each session contained 288 EEG trials, with 72 trials per task. The first session is designated for training while using second for testing.



(a) Flow chart of the single trial.

(b) Single motor imagery task (L: left-handed fist M: both-ankles bend R: right-handed fist)

Fig. 4. Illustration of the single motor imagery task and the flow chart of the single trial.

2) Dataset II: The 2019 World robot conference contest-BCI Robot Contest MI (WBCIC-MI) dataset contains 3-class MI-EEG data from 12 healthy subjects provided by Shanghai University. Twelve healthy right-handed students from the school participated in the experiment. All participants are naive BCI users and provide informed consent before the experiment. This experimental study, which received ethical approval (No.20190002) from the Tsinghua University Medical Ethics Committee, complies with the Declaration of Helsinki. Drinking alcohol 24 hours before the test, coffee, or tea within 4 hours is prohibited. The subjects are required to sit on a comfortable chair roughly 1m in front of the computer screen and remain as still as possible when performing the tasks.

In a single session, there is a 1-minute eye-opening time when subjects need to keep calm and gaze at the "+" on the screen without MI. After the opening eyes phase, a 1-minute closing eyes phase is subsequent. Then, the subject can press any button to start MI, including three-class MI tasks-left-handed fist, right-handed fist, and both-ankles bend, as shown in Fig. 4(a). In a single trial, shown in Fig. 4(b), a virtual reality video as "Cue" appears on the screen for 1.5 seconds when subjects should prepare for MI. In the next 4 seconds, subjects perform corresponding MI tasks according to the "Cue." At the end of the trial, the subjects have a break for 2 seconds.

Each set of data (block) contains five sessions, a total of 300 trials divided equally into three-class MI tasks, and between each session, there is a 2 minutes rest time for subjects.

The EEG signals are based on CPz and collected by Neuracle 64-lead equipment whose wet electrodes are arranged following the international 10-20 electrode placement method, with a 1000 Hz sampling rate, maintaining impedance below 10kΩ during recording and subsequently downsampling to 250 Hz. Ten-fold cross-validation is used in the WBCIC-MI dataset.

3) Preprocessing: In this study, we preprocess the EEG data with the Mne [36]. We use [2, 6] seconds of each trial and perform the 0.5-40Hz band-pass filtering on the original data to remove interference signals such as ocular electricity and electromyography, improving the signal-to-noise ratio.

B. Experiment Details

In this study, we implement our M-FANet using the PyTorch library, based on Python 3.7, with a Geforce 3090 GPU. We train M-FANet employing the Adam optimizer [37] with default settings. The batch size and learning rate are configured at 16 and 0.0001, respectively.

We utilize classification accuracy and Cohen's kappa as two key performance metrics for method evaluation. The computation formula for kappa is as follows:

$$kappa = \frac{p_o - p_e}{1 - p_e} \tag{9}$$

where p_o denotes the classification accuracy, p_e denotes the accuracy of random guesses. To analyze statistical significance, we apply the Wilcoxon Signed Rank Test.

C. Baseline Comparison

We conduct extensive subject-dependent experiments and compare our method with several state-of-the-art approaches across two separate datasets. Dataset I is a widely adopted dataset for four-class MI tasks, and many notable methods have proven their efficacy on this dataset. For instance, OSTP [31] adapted parameters specifically for the selection within the CSP framework; FBCSP [18] initially subdivided the signal into frequency bands and subsequently applied individual spatial filtering to each band; DeepConvNet [21] and EEGNet [24] were both convolution-based methods, designed for feature extraction and classification, with EEGNet being more lightweight than DeepConvNet. FBCNet [23], on the other hand, leveraged narrowband filters to capture features across different frequency bands, followed by a convolution with a large kernel to learn the spatial pattern for each band and subsequently extracted temporal feature through variance computation. Conformer [28], which extracts features through a self-attention module.

Table II displays the classification performance of all methods on Dataset I. Our M-FANet method achieves the highest average accuracy of 79.28% and the highest average kappa value of 0.7259. The results demonstrate that the classification accuracy of our proposed M-FANet is not only 11.5% higher than FBCSP ($p < 0.05$), which was the victor of the BCI Competition IV but also significantly surpasses that of OSTP ($p < 0.01$). Deep learning methods entirely based on CNN, such as DeepConvNet and EEGNet, also garnered commendable classification results, further testifying the effectiveness of CNN in feature extraction. However, these methods lack effective feature selection. For example, FBCNet does not selectively concentrate on the frequency bands it partitions, and EEGNet compresses all spatial information directly with large convolutional kernels, neglecting attention to local channel groups. In addition, they also lack calibration of feature maps. In contrast, our proposed M-FANet enhances CNN with frequency band attention, local spatial attention, and feature map attention, hence selecting informative features. M-FANet, employing compact yet accurate attention modules, outperformed the larger model Conformer [28], which utilized self-attention modules. As a result, M-FANet achieves the best average accuracy ($p < 0.05$) and kappa.

Further, we compare several state-of-the-art methods on Dataset II in Table III. In comparison to other methods, M-FANet significantly enhances the overall classification performance, with improvements of 32.81%, 16.75%, 12.17%,

TABLE II
CLASSIFICATION ACCURACY(%) COMPARISONS WITH STATE-OF-THE-ART METHODS ON DATASET I

Dataset	Methods	S01	S02	S03	S04	S05	S06	S07	S08	S09	Average	Kappa
I	OSTP [31]	73.10	39.80	78.70	57.40	41.20	25.50	82.90	75.00	62.00	59.50	0.4830
	FBCSP [18]	76.00	56.50	81.25	61.00	55.00	45.25	82.75	81.25	70.75	67.75	0.5700
	ConvNet [21]	76.39	55.21	89.24	74.65	56.94	54.17	**92.71**	77.08	76.39	72.53	0.6337
	EEGNet [24]	85.76	61.46	88.54	67.01	55.90	52.08	89.58	83.33	79.51	74.50	0.6600
	FBCNet [23]	85.42	60.42	90.63	76.39	74.31	53.82	84.38	79.51	80.90	76.20	0.6827
	Conformer [28]	**88.19**	61.46	**93.40**	**78.13**	52.08	**65.28**	92.36	**88.19**	**88.89**	78.66	0.7155
	M-FANet	86.81	**75.00**	91.67	73.61	**76.39**	61.46	85.76	75.69	87.15	**79.28**	**0.7259**

TABLE III
CLASSIFICATION ACCURACY(%) COMPARISONS WITH STATE-OF-THE-ART METHODS ON DATASET II

Dataset	Methods	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	Avg	Kappa
II	OSTP [31]	43.33	89.33	37.00	38.00	35.33	40.33	49.00	48.00	38.00	34.00	54.54	33.67	45.05	0.3099
	FBCSP [18]	63.67	94.33	42.00	52.33	34.33	40.00	82.00	81.33	71.67	43.67	71.60	59.33	61.11	0.4166
	ConvNet [21]	68.67	91.00	79.67	51.33	81.00	69.33	97.67	76.00	**72.00**	**70.00**	79.93	41.33	73.16	0.5950
	EEGNet [24]	62.33	91.67	82.33	49.67	74.33	57.67	88.67	53.67	62.67	59.33	68.91	37.00	65.69	0.4867
	FBCNet [23]	72.00	**94.67**	53.33	58.33	45.67	53.00	83.67	**83.33**	64.00	36.67	75.52	62.00	65.18	0.4768
	Conformer [28]	64.67	89.67	61.33	57.33	66.67	63.67	94.67	70.00	52.67	68.00	80.28	41.33	67.52	0.5131
	M-FANet	**75.33**	91.00	**82.00**	**65.67**	**84.33**	**72.33**	**98.33**	80.67	71.33	65.00	**82.61**	**65.67**	**77.86**	**0.6650**

12.68%, 10.34%, and 4.7%, respectively, for OSTP ($p < 0.01$), FBCSP ($p < 0.01$), EEGNet ($p < 0.01$), FBCNet ($p < 0.01$), Conformer ($p < 0.05$), and DeepConvNet ($p < 0.05$).

D. Training Process

We employ a two-stage training strategy [21]. In the first stage, the training data is divided into training and validation sets. The model is trained exclusively on the training set, and its accuracy is monitored using the validation set. If the validation set accuracy does not improve for 400 consecutive epochs, the training is stopped. Once the stopping criterion is met, the network parameters with the best validation set accuracy are loaded. These loaded model parameters serve as the starting point for the second stage, where the model is further trained using the complete training data (training + validation sets). The second stage of training is stopped when the validation set loss decreases below the training set loss from stage 1. To prevent unlimited training in cases where convergence is not achieved, the maximum number of training epochs is limited to 1500 for stage 1 and 600 for stage 2. For the Dataset I, 20% of the training data is reserved as the validation set. In Dataset II, one of nine training folds is retained as the validation set.

E. Ablation Study

The significant advancement of M-FANet lies in its multi-feature attention modules compared to other state-of-the-art methods. Therefore, we conduct an ablation study using Dataset I, as presented in Fig. 5, involving the sequential removal of frequency band attention, local spatial attention, and feature map attention from our method. The results show a corresponding average accuracy decrease of 5.98%, 2.81%, and 2.05%, respectively. When we remove the frequency band attention module, three participants (S04, S07, S09) exhibit a decreased classification accuracy of over 10%. The most pronounced decrease in accuracy is S07, with a reduction of 13.89%. A comparison of the confusion matrix for S07, as shown in Fig. 6, reveals that the frequency band attention module substantially enhances the classification accuracy for tasks involving the right hand and both feet.

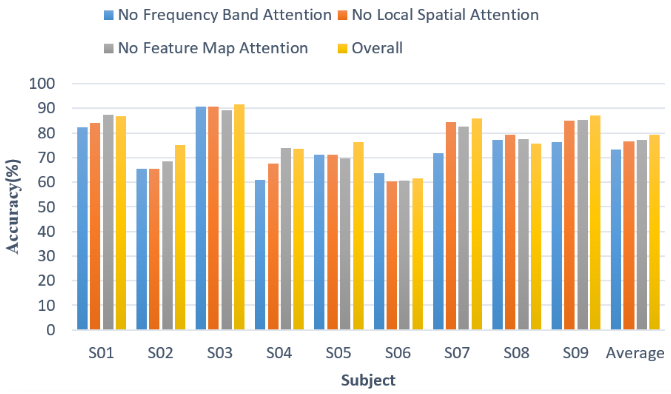
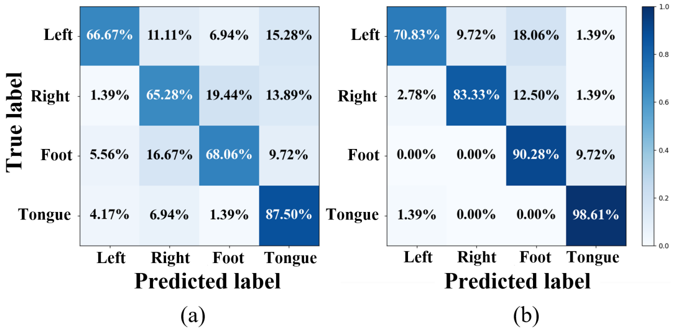


Fig. 5. Ablation study on the impact of frequency band attention, local spatial attention, and feature map attention on classification accuracy.



(a) (b)

Fig. 6. Confusion matrix for S07. (a) No frequency band attention. (b) Overall.

F. Parameter Sensitivity

R-Drop also contributes to our method in terms of model training. It introduces a regularization term to the loss

TABLE IV
THE CLASSIFICATION ACCURACY(%) OF M-FANET WITH DIFFERENT α ON DATASET I

	S01	S02	S03	S04	S05	S06	S07	S08	S09	Average	Kappa
α = 0	83.68	67.71	90.63	73.61	72.22	56.60	74.31	78.82	81.94	75.50	0.6734
α = $1e^{-1}$	85.76	69.79	87.50	65.63	73.96	57.29	76.74	79.86	83.33	75.54	0.6739
α = $2e^{-1}$	85.76	71.88	88.19	69.79	72.92	60.42	82.29	78.47	84.72	77.16	0.6955
α = $3e^{-1}$	84.72	72.92	88.19	69.79	73.26	61.11	82.29	83.33	84.38	77.78	0.7032
$\boldsymbol{\alpha}$ **=** $\mathbf{4e^{-1}}$	86.81	75.00	91.67	73.61	76.39	61.46	85.76	75.69	87.15	**79.28**	**0.7259**
α = $5e^{-1}$	86.81	69.10	90.63	72.92	73.96	58.33	83.68	78.47	86.46	77.82	0.7042
α = $6e^{-1}$	83.68	62.15	92.36	74.65	71.88	61.11	79.86	78.82	86.81	76.81	0.6908
α = $7e^{-1}$	85.76	62.85	92.36	72.57	72.57	59.03	79.86	75.35	87.50	76.43	0.6857

TABLE V
THE CLASSIFICATION ACCURACY(%) OF M-FANET WITH DIFFERENT α ON DATASET II

	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	Average	Kappa
α = 0	69.00	96.33	81.00	63.33	84.67	69.00	98.00	75.67	66.33	69.33	80.62	53.33	75.55	0.6312
α = 1	69.67	93.00	77.67	63.33	83.33	72.33	98.00	73.67	72.00	74.33	82.62	53.33	76.11	0.6325
$\boldsymbol{\alpha}$ **= 5**	75.33	91.00	82.00	65.67	84.33	72.33	98.33	80.67	71.33	65.00	82.61	65.67	**77.86**	**0.6650**
α = 10	72.67	93.67	79.67	65.67	85.67	72.33	97.67	81.00	67.67	67.00	82.28	63.67	77.41	0.6639
α = 15	68.00	91.33	82.67	67.67	79.67	71.67	96.33	79.00	66.33	62.67	82.64	57.33	75.44	0.6350

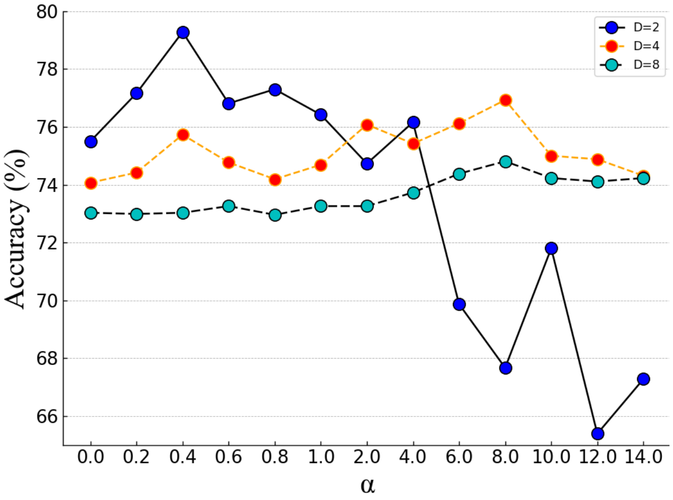


Fig. 7. Different values of D are chosen to adjust the number of trainable model parameters, and the influence of model size on the selection of α is observed on Dataset I. The results indicated that for larger models, a larger α is necessary for regularization.

function, and we can manipulate the impact through an adjustable parameter α. We experiment with varying the α in $\{0, 1e^{-1}, 2e^{-1}, 3e^{-1}, 4e^{-1}, 5e^{-1}, 6e^{-1}, 7e^{-1}\}$ and evaluate the effect of regularization term on Dataset I. When $\alpha = 0$, it implies that there is no constraint term between sub-models. As shown in Table IV, a small α (e.g., $1e^{-1}$) does not outperform a large α (e.g., $4e^{-1}$), indicating the necessity for enhancing regularization. However, an excessively large α (e.g., $7e^{-1}$) signifies excessive regularization, causing the classification accuracy to decrease. In this work, the best-balanced choice is $\alpha = 4e^{-1}$ on Dataset I.

We also vary the α in $\{0, 1, 5, 10, 15, 20\}$ and conduct experiments on Dataset II. As shown in Table V, when the α increases, the performance of our method gradually improves. Nonetheless, an excessively large α may lead to a decline in classification performance, as it might obstruct model convergence during training. The optimal balance is reached when $\alpha = 5$ on Dataset II.

Due to the difference in spatial information between Dataset I with 22 channels and Dataset II with 64 channels, we utilize different spatial filter multipliers (D). For Dataset I, we use $D = 2$, while for Dataset II, we use $D = 4$. During the experiments, we observed that changing the value of D results in variations in the distribution of α. For models with a larger D, indicating more parameters, a larger α is required to achieve optimal classification accuracy. Consequently, we set D in $\{2, 4, 8\}$ on Dataset I, where the corresponding numbers of parameters are 4080, 7696, and 16656 respectively, and examine the impact of different α values on average accuracy. Fig. 7 shows that smaller models exhibit greater sensitivity as α increases by the same amount. With the continuous increase of α, the small-size model shows a trend of an initial increase and then a decrease in accuracy, quickly reaching the optimal point ($\alpha = 0.4$). The medium-sized model demonstrates a similar trend but changes more slowly, with the optimal point occurring at a larger α value ($\alpha = 8$). In contrast, the big-size model displays a consistently slow increase in accuracy and has not yet reached their optimal point.

A larger value of D increases the number of parameters in M-FANet, therefore requiring a larger value of α, i.e., a more substantial regularization term, to constrain the training process. Therefore we recommend larger values of D and α for datasets containing more raw EEG information.

G. Visualization

1) Global Representation: The frequency band attention module extracts features directly from the raw EEG signals, critically influencing the feature extraction processes in all subsequent modules. Therefore, it is essential to ascertain whether the features learned by the frequency band attention module are exclusively derived from specific MI tasks. We extract the output from the frequency band attention module and represent it as a brain topography map for comparison

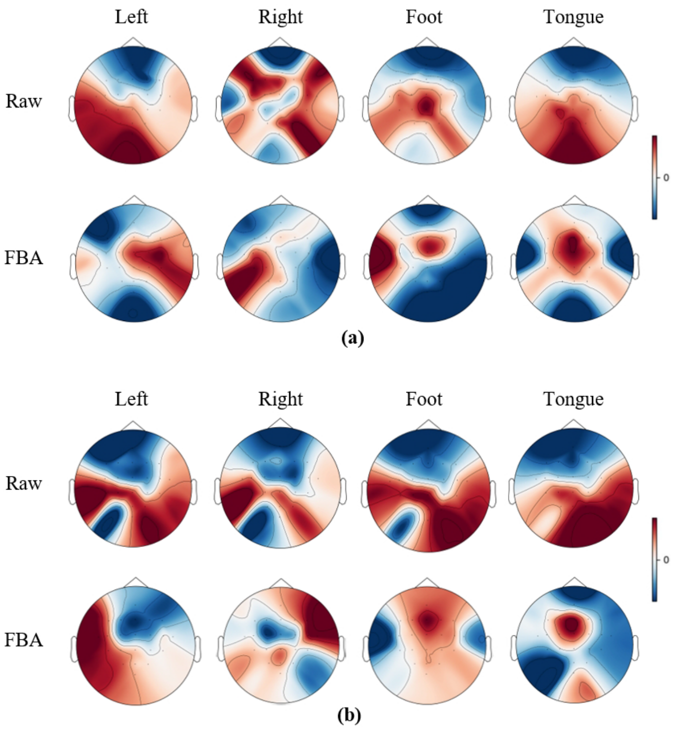


Fig. 8. Raw EEG topography averages over all trials for each MI task, the attribution patterns after the frequency band attention module (FBA) on the input EEG show that the features become more focused on the MI-relevant regions. (a) Attribution patterns for S07. (b) Attribution patterns average across all subjects.

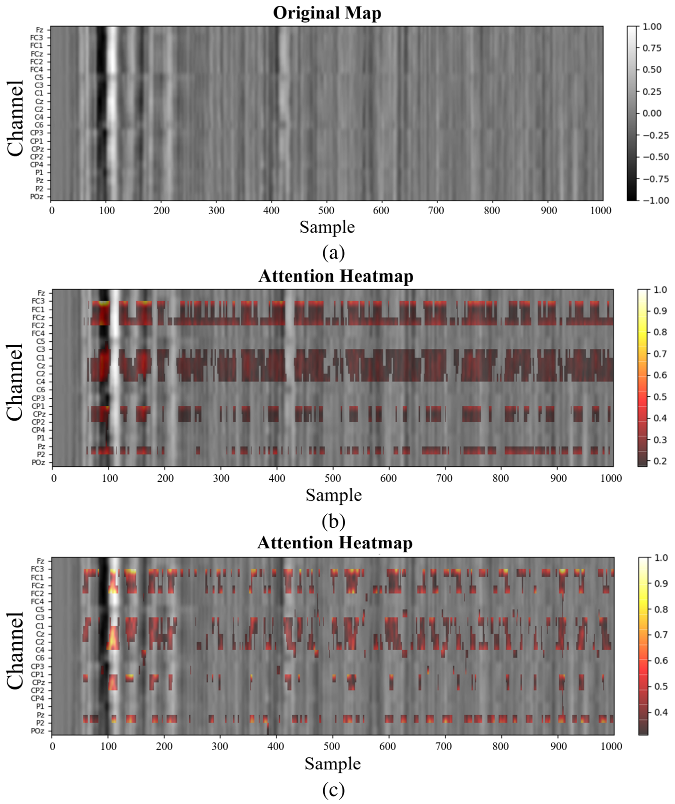


Fig. 9. The heatmap of S07 from Dataset I, where the gray color represents the original EEG image, and the red color represents the output of the local spatial attention module. (a) The origin EEG signals. (b), (c) The heatmap after 10 and 100 iterations, respectively.

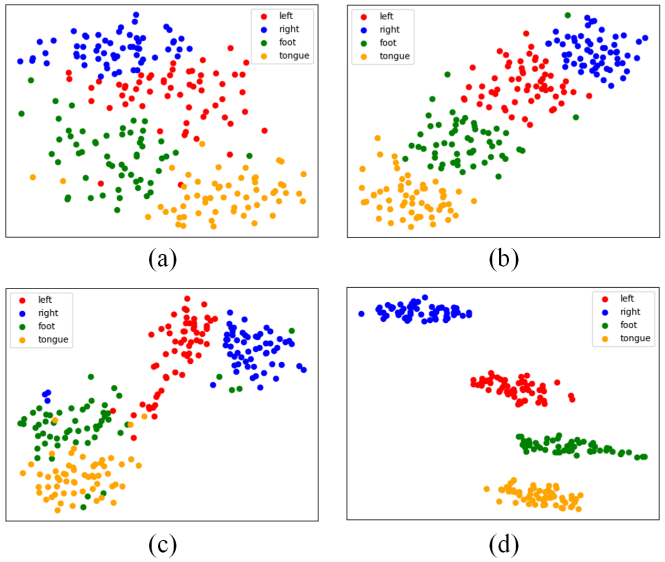


Fig. 10. The distribution of feature vectors for S07 based from Dataset I. All feature vectors are mapped to the 2D space using the t-SNE method. (a) DeepConvNet. (b) EEGNet. (c) FBCNet. (d) M-FANet.

with the raw EEG signals. Attribution patterns for S07 and all subjects are shown in Fig. 8 (a) and (b).

The raw EEG signals from subject S07 already exhibit well-known MI-related brain activation patterns. After processing through the frequency band attention module, the activation features under different MI tasks become more pronounced. When averaging the EEG under specific tasks of all subjects, the overall activation patterns appear irregular due to the variability in brain activation modes among individuals. However, the frequency band attention module still distinctly differentiates the brain activation patterns associated with different tasks.

2) Region of Interest: To evaluate the effectiveness of the local spatial attention module, we employ heatmaps to visualize both the input and output of this layer, thereby scrutinizing the tangible impact of local spatial attention. As shown in Fig. 9 (b), during the early stages of model training, the local spatial attention module chiefly concentrates on local channel groups associated with motion regions where attention is distributed over an expansive range, and the response values are comparatively low. As the model matures into a more stable state, shown in Fig. 9 (c), the attention becomes more refined, with an increment in response values. This phenomenon suggests that the local spatial attention module persistently focuses on local spatial information, intensifying effectiveness as the model evolves through training. The significance of this module lies in enabling subsequent large convolutional kernel, which extracts global spatial features, to pay more attention to spatial information relevant to MI that is manifested in the figure by higher response values in local channel groups activated by MI.

3) Feature Distribution: T-distributed stochastic neighbor embedding (t-SNE) [38] is a commonly used statistical dimension reduction and feature visualization method. After separate training with DeepConvNet, EEGNet, FBCNet, and M-FANet,

TABLE VI
COMPARISONS WITH STATE-OF-THE-ART METHODS IN THE RESOURCE CONSUMPTION ON DATASET I

Method	Parameters	Size(KB)	Time(ms)	MFLOPs
FBCNet [23]	4180	**20.48**	1486	**0.576**
ConvNet [21]	282889	1117	**1047**	3.534
Conformer [28]	789816	3148	1775	63.86
M-FANet	**4080**	28.10	1536	23.39

the feature distributions of S07 from Dataset I are shown in Fig. 10. The visual representation in Fig. 10 (d) highlights the superior feature extraction capability of M-FANet and provides a plausible explanation for its outstanding classification performance. The extracted features exhibit more explicit boundaries and more distinct clusters, suggesting that M-FANet effectively captures discriminative information and enhances the separability of different classes. These findings further support that the improved feature extracted by M-FANet contributes to its superior classification performance compared to DeepConvNet, EEGNet, and FBCNet.

V. DISCUSSION

The key to the research and application of MI-BCI lies in the accuracy of MI decoding methods. We propose a lightweight yet highly effective method called M-FANet. The distinctive feature of its network architecture is its reliance on attention mechanisms and the utilization of multiple attention modules based on the frequency and spatial characteristics of MI. In the frequency domain, we divide the original signals into multiple sub-bands and apply adaptive weights to non-overlapping sub-bands. These sub-bands are then linearly combined to remove redundant information and amplify relevant frequency-domain information. Inspired by the neurophysiological mechanisms of MI in the spatial domain, we employ a local spatial attention module for detailed feature extraction. When passing through MI-related brain areas, the local spatial attention module will output a significant activation, as demonstrated in Fig. 9, showcasing its sensitivity to the regions associated with MI. The local spatial attention module allows our network to prioritize MI-related responses in these areas. Additionally, we incorporate attention mechanisms from deep learning and utilize SEBlock to calibrate the feature maps. To address the challenges of small sample size, high feature dimensionality, and the risk of overfitting in EEG signals, we introduce R-Drop, which mitigates the inconsistencies between sub-models and enhances the generalization ability of the complete model during the inference stage.

The experimental results substantiate that M-FANet outperforms the current state-of-the-art methods in terms of performance. Ablation experiments illustrate that each attention module contributes significantly to the overall model. To assess the effectiveness of feature extraction, we visualize the features. We also discuss the impact of the α parameter on the performance of R-Drop, which indicates that as α increases, the model's classification performance gradually improves. However, exceedingly large values of α can lead to a deterioration in performance. We also note that the optimal value of α varies for deep learning networks of different parameter sizes. Networks with more parameters necessitate a larger α, likely due to the amplified inconsistency between sub-models requiring more potent regularization.

Table VI displays the resource consumption of four state-of-the-art methods. The M-FANet model completes a single forward pass in 1,536 milliseconds, requiring 23.39 Million Floating Point Operations (MFLOPs), which denotes the computational load of the network for each inference. However, M-FANet boasts a compact architecture, requiring only 4,080 parameters and a minimal memory footprint of 28.10KB, significantly less than ConvNet [21] and Conformer [28]. In summary, M-FANet strikes a balance between superior classification performance and minimal memory requirements.

Despite these advancements, several limitations call for further investigation. Firstly, we only conduct experiments on MI datasets. However, given that our proposed M-FANet exhibits high sensitivity to local regions, exploring its effects on other paradigms, such as event-related potentials, may prove valuable. Secondly, the adjustable parameter α in R-Drop currently depends on empirical settings. While extensive experimentation allows us to determine the optimal value of α, specific to a particular network architecture, developing a computationally efficient formula for calculating α, which holds significance across different network structures, could be beneficial. Thirdly, we fix the stride for sub-band partitioning currently. Although we can use some neurophysiological priors as references, a fixed stride may result in valuable information being submerged in sub-bands containing more redundant information. In our future work, we plan to explore the usage of a non-fixed stride for sub-band partitioning, where the number of sub-bands can be specified, and boundary points can be automatically determined.

Transfer learning is also a noteworthy method, enabling the utilization of pre-trained models on extensive datasets, followed by fine-tuning with a limited amount of specific subject data to improve classification performance [39], [40]. The combination of lightweight models and transfer learning facilitates high-performance decoding while maintaining low resource consumption. This approach is particularly beneficial for the further application and popularization of BCIs, especially in scenarios with limited resources, such as in portable and embedded devices.

VI. CONCLUSION

This study proposes a lightweight M-FANet that employs convolution for feature extraction and effectively selects frequency band features, local spatial features, and informative feature maps using attention mechanisms. Moreover, we introduce a training method, R-Drop, to tackle the training-inference inconsistency of the model prompted by dropout. We conduct extensive experiments on two datasets, and the results corroborate that our method outperforms other state-of-the-art methods. Ablation experiments and visualization further validate the effectiveness of the individual attention modules and R-Drop. The proposed M-FANet holds promising potential for MI-based BCI research and applications due to its high performance in MI-EEG decoding.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, Jun. 2002.

[2] J. W. Choi, B. H. Kim, S. Huh, and S. Jo, "Observing actions through immersive virtual reality enhances motor imagery training," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 7, pp. 1614–1622, Jul. 2020.

[3] A. Singh, A. A. Hussain, S. Lal, and H. W. Guesgen, "A comprehensive review on critical issues and possible solutions of motor imagery based electroencephalography brain–computer interface," *Sensors*, vol. 21, no. 6, p. 2173, Mar. 2021.

[4] D. Yadav, S. Yadav, and K. Veer, "A comprehensive assessment of brain computer interfaces: Recent trends and challenges," *J. Neurosci. Methods*, vol. 346, Dec. 2020, Art. no. 108918.

[5] F. Fahimi, Z. Zhang, W. B. Goh, T.-S. Lee, K. K. Ang, and C. Guan, "Inter-subject transfer learning with an end-to-end deep convolutional neural network for EEG-based BCI," *J. Neural Eng.*, vol. 16, no. 2, Apr. 2019, Art. no. 026007.

[6] K. Sakai et al., "Effects of visual-motor illusion on functional connectivity during motor imagery," *Exp. Brain Res.*, vol. 239, no. 7, pp. 2261–2271, Jun. 2021.

[7] P. D. E. Baniqued et al., "Brain–computer interface robotics for hand rehabilitation after stroke: A systematic review," *J. Neuroeng. Rehabil.*, vol. 18, no. 1, pp. 1–25, Jan. 2021.

[8] J. Choi, K. T. Kim, J. H. Jeong, L. Kim, S. J. Lee, and H. Kim, "Developing a motor imagery-based real-time asynchronous hybrid BCI controller for a lower-limb exoskeleton," *Sensors*, vol. 20, no. 24, p. 7309, Dec. 2020.

[9] S. Chen et al., "Longitudinal electroencephalography analysis in subacute stroke patients during intervention of brain–computer interface with exoskeleton feedback," *Frontiers Neurosci.*, vol. 14, p. 809, Aug. 2020.

[10] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of EEG-based brain–computer interface paradigms," *J. Neural Eng.*, vol. 16, no. 1, Jan. 2019, Art. no. 011001.

[11] S. B. I. Badia, A. G. Morgade, H. Samaha, and P. F. M. J. Verschure, "Using a hybrid brain computer interface and virtual reality system to monitor and promote cortical reorganization through motor activity and motor imagery training," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 21, no. 2, pp. 174–181, Mar. 2013.

[12] K. K. Ang and C. Guan, "EEG-based strategies to detect motor imagery for control and rehabilitation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 4, pp. 392–401, Apr. 2017.

[13] P. Arpaia, A. Esposito, A. Natalizio, and M. Parvis, "How to successfully classify EEG in motor imagery BCI: A metrological analysis of the state of the art," *J. Neural Eng.*, vol. 19, no. 3, Jun. 2022, Art. no. 031002.

[14] S. Sur and V. Sinha, "Event-related potential: An overview," *Ind. Psychiatry J.*, vol. 18, no. 1, p. 70, 2009.

[15] M. Middendorf, G. Mcmillan, G. Calhoun, and K. S. Jones, "Brain–computer interfaces based on the steady-state visual-evoked response," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 2, pp. 211–214, Jun. 2000.

[16] X. Zhang, L. Yao, X. Wang, J. Monaghan, D. McAlpine, and Y. Zhang, "A survey on deep learning-based non-invasive brain signals: Recent advances and new frontiers," *J. Neural Eng.*, vol. 18, no. 3, Jun. 2021, Art. no. 031002.

[17] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 41–56, Jan. 2008.

[18] K. K. Ang, Z. Y. Chin, C. Wang, C. Guan, and H. Zhang, "Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b," *Frontiers Neurosci.*, vol. 6, p. 39, Mar. 2012.

[19] A. Al-Saegh, S. A. Dawwd, and J. M. Abdul-Jabbar, "Deep learning for motor imagery EEG-based classification: A review," *Biomed. Signal Process. Control*, vol. 63, Jan. 2021, Art. no. 102172.

[20] Y. Li, X.-R. Zhang, B. Zhang, M.-Y. Lei, W.-G. Cui, and Y.-Z. Guo, "A channel-projection mixed-scale convolutional neural network for motor imagery EEG decoding," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 6, pp. 1170–1180, Jun. 2019.

[21] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017.

[22] S. Sakhavi, C. Guan, and S. Yan, "Learning temporal information for brain–computer interface using convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5619–5629, Nov. 2018.

[23] R. Mane, N. Robinson, A. P. Vinod, S.-W. Lee, and C. Guan, "A multi-view CNN with novel variance layer for motor imagery brain computer interface," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 2950–2953.

[24] 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," *J. Neural Eng.*, vol. 15, no. 5, Oct. 2018, Art. no. 056013.

[25] T. Hanakawa, I. Immisch, K. Toma, M. A. Dimyan, P. Van Gelderen, and M. Hallett, "Functional properties of brain areas associated with motor execution and imagery," *J. Neurophysiol.*, vol. 89, no. 2, pp. 989–1002, Feb. 2003.

[26] A. Guillot, C. Collet, V. A. Nguyen, F. Malouin, C. Richards, and J. Doyon, "Brain activity during visual versus kinesthetic imagery: An fMRI study," *Hum. Brain Mapping*, vol. 30, no. 7, pp. 2157–2172, Jul. 2009.

[27] H. Li, D. Zhang, and J. Xie, "MI-DABAN: A dual-attention-based adversarial network for motor imagery classification," *Comput. Biol. Med.*, vol. 152, Jan. 2023, Art. no. 106420.

[28] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG conformer: Convolutional transformer for EEG decoding and visualization," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 710–719, 2023.

[29] L. Wu et al., "R-Drop: Regularized dropout for neural networks," in *Proc. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 10890–10905.

[30] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass brain–computer interface classification by Riemannian geometry," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 4, pp. 920–928, Apr. 2012.

[31] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Mutual information-based selection of optimal spatial–temporal patterns for single-trial EEG-based BCIs," *Pattern Recognit.*, vol. 45, no. 6, pp. 2137–2144, Jun. 2012.

[32] O. Özdenizci and D. Erdogmus, "Information theoretic feature transformation learning for brain interfaces," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 1, pp. 69–78, Jan. 2020.

[33] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[35] M. Tangermann et al., "Review of the BCI competition IV," *Frontiers Neurosci.*, vol. 6, p. 55, 2012.

[36] A. Gramfort et al., "MNE software for processing MEG and EEG data," *NeuroImage*, vol. 86, pp. 446–460, Feb. 2014.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[38] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, Nov. 2008.

[39] K. Zhang, N. Robinson, S.-W. Lee, and C. Guan, "Adaptive transfer learning for EEG motor imagery classification with deep convolutional neural network," *Neural Netw.*, vol. 136, pp. 1–10, Apr. 2021.

[40] D. Wu, X. Jiang, and R. Peng, "Transfer learning for motor imagery based brain–computer interfaces: A tutorial," *Neural Netw.*, vol. 153, pp. 235–253, Sep. 2022.