Manhattan Self-Attention Diffusion Residual Networks with Dynamic Bias Rectification for BCI-based Few-Shot Learning

Hao Wang[1], Li Xu[1*], Yuntao Yu[2], Weiyue Ding[3], Yiming Xu[4]
[1]College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]China Electronics Standardization Institute, Beijing, China
[3]School of Mathematics, Harbin Institute of Technology, Harbin, China
[4]College of Engineering, Tokyo Institute of Technology, Tokyo, Japan
b223060004@hrbeu.edu.cn, xuli@hrbeu.edu.cn, yuyt@cesi.cn, wyding0501@hotmail.com, kyo@ac.sc.e.titech.ac.jp

Abstract

The distribution biases and scarcity of samples in multi-source data present significant challenges for few-shot learning (FSL) tasks based on brain-computer interface (BCI). Recent efforts have explored the application of diffusion mechanisms in FSL, typically utilizing labeled data to augment the support set. However, this approach has not effectively utilized unlabeled data nor addressed distribution biases. Inspired by the latest advancements in FSL, we propose the manhattan self-attention diffusion residual networks (MSADiff-Resnet) with dynamic bias rectification. This model explicitly adds the manhattan self-attention diffusion layer to resnet, using attention mechanisms and manhattan distance-based decay function to control local diffusion intensity, and adjusts the global diffusion strength through the parameter. This diffusion mechanism bridges labeled and unlabeled data, addressing the limitations associated with sample availability. Additionally, we effectively tackle the distribution biases of multi-source data through inter-class bias rectification and dynamic intra-class bias rectification. Moreover, this study presents for the first time a universal deep learning framework specifically designed for BCI-based FSL tasks. Extensive experiments on multi-source BCI task datasets have validated the effectiveness of proposed method.

Introduction

Brain-computer interface (BCI) (Chapin et al. 1999) control the environment by decoding neural activity. Currently, deep learning is the primary method for handling BCI tasks (Zhao, Yan, and Lu 2021; Song et al. 2020; Li et al. 2024), but the difficulty of data collection and labeling in real-world scenarios limits its generalization capabilities in low-data situations (Rong et al. 2023). Few-shot learning (FSL) (Fei-Fei, Fergus, and Perona 2006) addresses the issue of data scarcity in deep learning by learning transferable knowledge from data-rich base classes and using a small number of labeled samples to assist in predicting new classes.

However, in FSL, we primarily face two challenges. The first challenge is that the diversity of sample sources introduces distribution biases, which can impact the model's generalization capabilities (Li et al. 2020; Yang, Kim, and Yun 2024). The second challenge is that the limited number of training samples leads to model overfitting on the support set data (Wu et al. 2024).

In response to the challenge of distribution biases in multi-source samples, we drew inspiration from (Liu, Song, and Qin 2020) to adjust for both inter-class and intra-class biases, proposing a dynamic bias rectification method. Inter-class bias refers to the distance between the mean vectors of the support and query sets, while intra-class bias refers to the discrepancy between the actual prototype and the prototype computed from the support set. By dynamically adjusting the relationship between each sample and its class prototype, we rectify the prototypes to enhance the model's robustness when handling distributional biases in multi-source samples.

To address the challenge of overfitting caused by the limited number of samples in the support set, we propose the manhattan self-attention diffusion residual network (MSADiff-Resnet). This model enhances interactions between labeled and unlabeled samples through a diffusion mechanism, aggregating similar points. The manhattan self-attention (MSA) mechanism (Fan et al. 2024) controls the flow of information and optimizes diffusion paths by considering the decaying weights based on manhattan distances between sample points. This approach effectively transmits information between similar data points, significantly mitigating the overfitting problem caused by the limited number of support set samples.

Currently, deep learning models based on electroencephalography (EEG) are designed for specific datasets and tasks (Liang et al. 2023; Heinrichs, Heim, and Weber 2023; Wang, Zhang, and Tang 2024), lacking generalizability. To address this issue, we propose a universal deep learning framework for processing BCI-based FSL tasks. For multi-source EEG data, we first enhance task-related components using task-related component analysis (TRCA) (Tanaka, Katura, and Sato 2013). Subsequently, we convert EEG signals with differing statistical characteristics into a uniform two-dimensional image format using the gramian angular field (GAF) (Wang and Oates 2015). We divide the multi-source dataset into support and query sets, then perform distribution bias rectification. Finally, labels for the query set are obtained using the proposed MSADiff-Resnet model.

In summary, the contributions of this paper are as follows:

- A dynamic bias rectification method was proposed to address inter-class and intra-class distribution biases. For

*Corresponding author

inter-class bias, a shifting term is added to the normalized query set to bring it closer to the center of similar samples in the support set. For intra-class bias, the similarity between each sample and the class prototype is dynamically weighted. Additionally, a pseudo-labeling strategy is used to expand the support set, allowing for the rectification of prototypes.

- To combat overfitting caused by the limited sample size of the support set, the MSADiff-Resnet has been proposed. This model enhances the interaction between labeled and unlabeled samples through a diffusion mechanism and controls the diffusion paths of information using the MSA mechanism, optimizing the dynamics of information flow.
- A universal deep learning framework has been developed to address FSL tasks in BCI datasets. This framework systematically integrates key steps such as data preprocessing, optimization, format standardization, and distribution bias rectification. It employs the MSADiff-Resnet to predict labels for the query set. The effectiveness of the proposed method has been validated through extensive experiments on multi-source datasets.

Related Work

Few-Shot Learning

This section focuses on FSL methods based on metric learning and model learning, which are closely related to the proposed method .

Metric-based methods solve classification problems by learning a measure of similarity between samples (Guo et al. 2023; Chen et al. 2023). Prototype Networks (PN) (Snell, Swersky, and Zemel 2017) classify based on the distance between query points and the nearest class prototype, but their performance is limited in high-dimensional embedding spaces (Gong 2023). BD-CSPN (Liu, Song, and Qin 2020) builds on PN by using feature shift strategy and pseudo-labeling strategy to adjust biases. However, the pseudo-labeling strategy utilizes only a few high-confidence query points, leading to low data utilization rates, and reliance solely on cosine similarity fixes the relationship between features and prototypes. Model-based methods focus on designing architectures that can quickly adapt to new tasks (Lai et al. 2022; Zhu et al. 2024), enhancing classification performance of limited labeled samples through attention mechanisms, but they lack consideration for distribution biases in multi-source data and the potential of unlabeled data.

In response to the challenges mentioned above, this study introduces a novel integration strategy that combines the strengths of both metric-based and model-based learning methods. In terms of metrics, we dynamically rectify the sample prototypes to reduce differences between source samples and refine model parameters through a prototype loss function. In terms of the model, we employ a diffusion mechanism to enhance the impact of unlabeled data on the processing of labeled data.

Diffusion Mechanism

Diffusion is a fundamental mechanism describing interactions between different entities in physical processes, and its concept has been widely applied across various fields (Dhariwal and Nichol 2021; Atwood and Towsley 2016; Gasteiger, Weißenberger, and Günnemann 2019; Avdeyev et al. 2023). Currently, there are few applications of the diffusion mechanism in FSL. Meta-Diffusion Models (Hu et al. 2023) and DiffAlign (Roy et al. 2022) use diffusion to generate pseudo data for sharpening decision boundaries, but they overlook the rich information contained in the query set. Diff-ResNet (Wang et al. 2023) adds convection-diffusion layers between residual networks to promote information exchange between the query set and support set data. However, this method uses a gaussian kernel to pre-construct the weight matrix, making the diffusion intensity fixed and lacking dynamic adjustment, and it does not consider distribution biases in multi-source data.

Addressing these shortcomings, we introduce MSADiff-Resnet. This model employs the MSA mechanism (Fan et al. 2024) to achieve local dynamic diffusion based on data distance metrics and characteristics transferred by the model. It utilizes the diffusion parameter to regulate the global diffusion intensity, enabling more effective utilization of unlabeled data. Additionally, the proposed dynamic biases rectification method improves the distribution biases.

Few-Shot Learning in the BCI Field

Unlike the rapid development of FSL in the field of computer vision, currently, only a few studies have applied FSL techniques to the BCI field (An et al. 2023; Pan et al. 2023; Bhosale, Chakraborty, and Kopparapu 2022). These studies mainly focus on labeled data, struggling with overfitting issues due to sample limitations and are only applicable to specific tasks. Addressing these deficiencies, a universal framework for FSL based on EEG signals has been proposed. It systematically integrates key steps for handling FSL with multi-source EEG data, achieves distribution bias rectification, and effectively utilizes unlabeled data, addressing the challenges of data diversity and scarcity in BCI field.

Method

Figure 1 displays our methodological framework, which includes five essential modules: (a) Original Signal Input; (b) TRCA; (c) GAF; (d) Distribution Bias Rectification; and (e) MSADiff-Resnet. The process begins with the optimization of multisource EEG signals for feature reproducibility in the TRCA module. Subsequently, the GAF module utilizes polar coordinate and cosine transformation to effectively extract the EEG signals' features of periodic changes, trends, and energy distribution, while also converting EEG signals of varied data structures, sampling rates, and channel configurations into standardized two-dimensional feature maps. Following embedding training, the distribution bias rectification module performs dynamic rectifications on the biases present in the multi-source data. The MSADiff-Resnet module, leveraging resnet's inherent skip connections, enhances

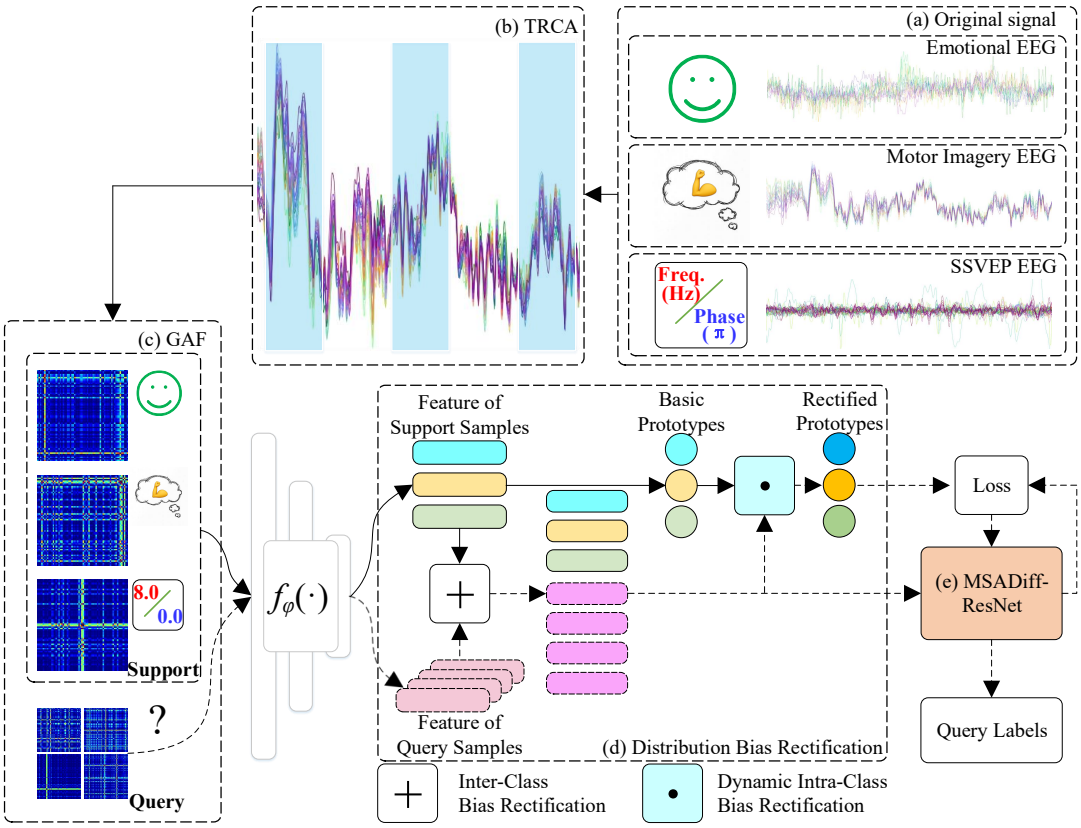

Figure 1: The overall framework of the method. The details of the main contribution modules (d) and (e) are displayed in Figure 2 and Figure 3, respectively.

the model's generalization abilities. It employs convection-MSADiff layers to exert forces on the data points, encouraging clustering of similar data, and concludes with a fully connected layer that predicts the labels of the query set. The hybrid loss function merges prototype loss with cross-entropy loss to refine the model's parameters.

Distribution Bias Rectification

The dataset $X = X_s \cup X_q$ consists of two subsets: $X_s = \{(x_i, y_i)\}_{i=1}^{N_1}$, the support set (labeled), and $X_q = \{x_j\}_{j=1}^{N_2}$, the query set (unlabeled), with N_1 representing the total number of samples in the support set and N_2 the total number of samples in the query set. The goal of FSL is to accurately determine the labels of the query set when the size of the support set, $|N_1|$, is very small.

In FSL task, samples are embedded into low-dimensional space, where the central point of each class is known as a prototype. These base prototypes are obtained by averaging the features of the support samples within each class.

$$P_n = \frac{1}{K} \sum_{i=1}^{K} \bar{x}_{i,n} \tag{1}$$

where, K represents the number of samples in each class; $\bar{x}_{i,n}$ denotes the normalized features of the i-th sample in class n within the support set.

However, there is a bias between the prototypes derived from the limited support set and the true prototypes, which can be categorized into inter-class and intra-class biases.

Inter-class bias refers to the distance between the average vectors of the support set and the query set. Minimizing this distance is a typical method for reducing distribution bias. The inter-class bias, denoted as B_{cross}:

$$B_{\text{cross}} = \mathbb{E}_{X_s \sim p_S}[X_s] - \mathbb{E}_{X_q \sim p_Q}[X_q] \tag{2}$$

where, p_S and p_Q represent the distributions of the support set and query set, respectively. Reducing B_{cross} is achieved by shifting the query set towards the support set. The prototype, after inter-class correction, is as follows:

$$\bar{x}'_q = \bar{x}_q + \xi \tag{3}$$

$$\xi = \frac{1}{|N_1|} \sum_{i=1}^{|N_1|} \bar{x}_{i,s} - \frac{1}{|N_2|} \sum_{j=1}^{|N_2|} \bar{x}_{j,q} \tag{4}$$

where, $\bar{x}_q$ represents the normalized query features; ξ is the shifting term; $\frac{1}{|N_1|} \sum_{i=1}^{|N_1|} \bar{x}_{i,s}$ is the average feature of the support set X_s; and $\frac{1}{|N_2|} \sum_{j=1}^{|N_2|} \bar{x}_{j,q}$ is the average feature of the query set X_q. By adjusting the feature vectors of the query set samples, the distribution bias of the multi-source samples is reduced, bringing them closer to the center of the

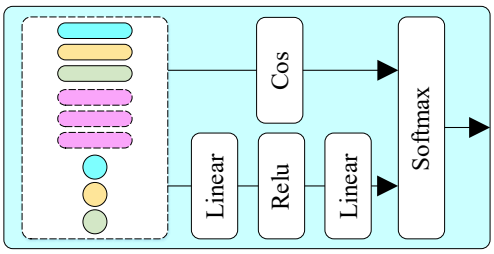

Figure 2: Dynamic intra-class bias rectification module.

corresponding class in the support set within the embedded space.

Intra-class bias refers to the discrepancy between the expected prototype and the actual prototype:

$$B_{\text{intra}} = \mathbb{E}_{X\sim p_X}[X] - \mathbb{E}_{X'\sim p_{X_X}}[X'] \tag{5}$$

where, p_X represents the distribution of all samples belonging to a specific class, while $p_{X'}$ denotes the distribution of the available labeled samples within that class. The expected prototype is represented by the average features of all samples within the class. The limitation of having only K samples per class in FSL introduces a bias in the prototype.

Next, a pseudo-labeling strategy is employed to reduce bias, where temporary labels are assigned to unlabeled data based on predicted confidence (Li et al. 2019). The top Z most confident query samples are selected to augment the support set X_s: $X'_s = X_s \cup X^Z_{q_{\text{pseudo}}}$. The rectified prototype is calculated using the normalized feature $\bar{x}'$:

$$P'_n = \sum\nolimits_{i=1}^{Z+K} w_{i,n} \cdot \bar{x}'_{i,n} \tag{6}$$

where, $w_{i,n}$ represents the dynamically adjusted weights that modulate the relationship between the support samples and the base prototype, calculated as follows:

$$w_{i,n} = \frac{\exp\left(\varepsilon \cdot \cos\cdot\left(W_2 \cdot \sigma\left(W_1 \cdot \left[x'_{i,n}, P_n\right]\right)\right)\right)}{\sum_{j=1}^{K+Z} \exp\left(\varepsilon \cdot \cos\cdot\left(W_2 \cdot \sigma\left(W_1 \cdot \left[x'_{j,n}, P_n\right]\right)\right)\right)} \tag{7}$$

where, ε is a scalar parameter; cos refers to the cosine similarity; P_n is the base prototype from Equation (1); the activation function $\sigma(\cdot)$ utilizes ReLU; W_1 and W_2 are parameters learned during training. The dynamic intra-class bias rectification module, shown in Figure 2, gives greater influence to samples that are more closely related to the prototype during the training process. This reduces the impact of outliers on the prototype, making the distribution of the rectified prototype $P'n$ more closely aligned with the expected prototype.

Then, we use the rectified prototype for the computation of the model's loss function, as shown in the Equation:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{P'_c} \tag{8}$$

$$\mathcal{L}_{CE} = -\sum\nolimits_{i=1}^{N_1}\sum\nolimits_{c=1}^{C} y_{i,c}\log\left(f\left(x_i\right)_c\right) \tag{9}$$

$$\mathcal{L}_{P'_c} = \sum\nolimits_{j=1}^{N_2}\sum\nolimits_{c=1}^{C} f\left(x_j\right)_c \cdot \left\|\bar{x}'_j - P'_c\right\|^2 \tag{10}$$

where, $\mathcal{L}_{CE}$ represents the cross-entropy loss; $\mathcal{L}_{P'_c}$ denotes the prototype loss; C is the total number of classes; P'_c is the rectified prototype for class c. $\mathcal{L}_{CE}$ directly influences the decision boundaries by maximizing the predicted probabilities, while $\mathcal{L}_{P'_c}$ minimizes the distance between the samples and the prototypes, allowing the model to adaptively fit the feature representations to the current data distribution.

MSADiff-Resnet

In resnet, x_i^k represents the feature of the data x_i after the k-th residual block. Residual connections refer to the addition of x_i^k to x_i^{k+1} through skip connections. Here, the convolutional layer, batch normalization layer, and other layers are combined and represented by the function f. Therefore, each residual block can be expressed as:

$$x_i^{k+1} = x_i^k + f\left(x_i^k, \theta^k\right) \tag{11}$$

where, θ^k represents the parameters of the k-th block. From the perspective of ordinary differential equations (ODEs), f is considered the velocity, x_i^k and x_i^{k+1} are viewed as the starting and ending positions of x_i, respectively. By introducing a time step Δt into f, resnet can be seen as the forward euler discretization of Equation (12), which describes the evolution of x_i:

$$\frac{dx_i(t)}{dt} = f\left(x_i(t), \theta(t)\right), \quad x_i(0) = x_i \tag{12}$$

The time step represents the independent movement of the convection term in the point iteration across layers. To enhance the interaction between the support set and the query set, a diffusion mechanism is introduced in Equation (13):

$$\frac{dx_i(t)}{dt} = f\left(x_i(t), \theta(t)\right) - \gamma\sum\nolimits_{j=1}^{N} w_{ij}\left(x_i(t) - x_j(t)\right) \tag{13}$$

where, N represents the number of data; $\gamma > 0$ is a parameter that adjustments the strength of diffusion; $w_{ij} \geq 0$ are the weights between x_i and x_j. The desired weight matrix facilitates the clustering of similar points. The convection term f is set as a two-layer network with a width w:

$$f(x(t), \theta(t)) = \sum\nolimits_{i=1}^{w} a_t^{(i)}\sigma\left(\mathbf{w}_t^{(i)} \cdot x(t) + b_t^{(i)}\right) \tag{14}$$

where, $x(t) \in \mathbb{R}^d$; $b_t^{(i)} \in \mathbb{R}$; $a_t^{(i)}$, $w_t^{(i)} \in \mathbb{R}^d$ and f: $\mathbb{R}^d \to \mathbb{R}^d$. The activation function $\sigma(\cdot)$ is ReLU. $\theta(t) = \left[\mathbf{w}_t^{(i)}, b_t^{(i)}, a_t^{(i)}\right]_{i=1}^{w}$ represents the set of network weights.

Using the classical Lie-Trotter splitting scheme (Geiser 2009), the convection-diffusion Equation (13) is discretized, and then a time step Δt is introduced into f and γ:

$$x_i^{k+1/2} = x_i^k + f\left(x_i^k, \theta^k\right) \tag{15}$$

$$x_i^{k+1} = x_i^{k+1/2} - \gamma\sum\nolimits_{j=1}^{N} w_{ij}\left(x_i^{k+1/2} - x_j^{k+1/2}\right) \tag{16}$$

The diffusion step Equation (16) serves as a stabilization for the convection step Equation (15). The weight matrix is

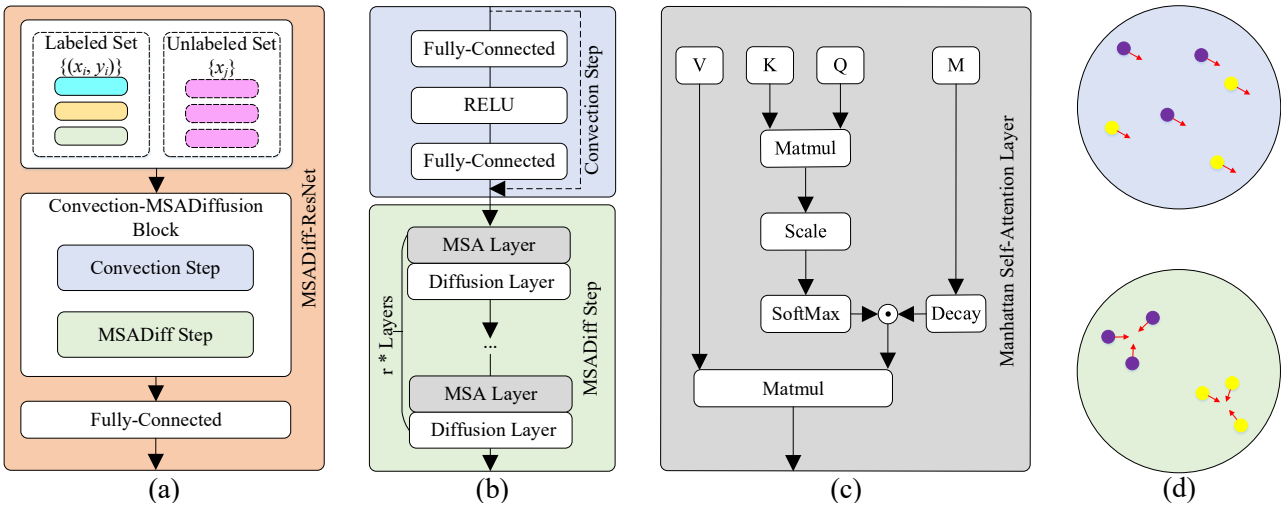

Figure 3: MSADiff-Resnet (a) Network structure; (b) Details of each block; (c) MSA mechanism (M represents Manhattan distance, Decay represents the decay function); (d) Changes in data points due to convection and diffusion processes.

measured by a gaussian kernel to assess the similarity between data points (Wang et al. 2023). However, the weight matrix w_{ij} constructed in this manner remains fixed after model initialization, meaning that the model cannot dynamically adjust the intensity of interactions between data points based on information learned during the training process. To address it, we use MSA to compute the weight matrix:

$$\mathrm{MSA}_{ij} = \left(\mathrm{softmax}\left(\frac{W_Q x_i^k W_K^T x_i^k}{\sqrt{d}}\right) \cdot e^{-\left|x_i^k - x_j^k\right|}\right) W_V x_i^k \quad (17)$$

where, W_Q, W_K, and W_V are the weight matrices for the query, key, and value respectively; $e^{-\left|x_i^{k+1/2} - x_j^{k+1/2}\right|}$ is a decay function based on the manhattan distance; $\sqrt{d}$ stabilizes the gradients. The decay function uses the manhattan distance to measure the similarity between data points and adjusts the range and intensity of the forces applied, reducing the interactions between distant nodes, similar to the physical process of diffusion. In Equation (16), the traditional gaussian kernel is replaced with a MSA mechanism:

$$x_i^{k+1} = x_i^{k+1/2} - \gamma \sum_{j=1}^{N} MSA_{ij}\left(x_i^{k+1/2} - x_j^{k+1/2}\right) \quad (18)$$

In Equation (18), the attention mechanism and the decay function work together to achieve precise control of local diffusion intensity. The global parameter γ regulates the overall diffusion intensity, balancing various local effects to maintain overall stability and facilitate rapid convergence. The method combines local and global diffusion regulation, allows for effective interaction between labeled and unlabeled data. r MSADiff layers are added after residual block. The structure of MSADiff-Resnet is shown in Figure 3.

Experiments

Datasets

BCICIV-2a (Brunner et al. 2008) dataset includes 4 different motor imagery tasks (left hand, right hand, both feet, and tongue) from 9 subjects. This experiment involves EEG data from 22 channels. Each run consists of 48 tests distributed evenly and randomized in sequence (12 tests per class), with 1 session containing 6 runs, totaling 288 tests.

Seed-V (Liu et al. 2021) dataset includes 5 emotional tasks (happy, sad, disgust, neutral and fear) from 16 subjects. This experiment involves EEG data from 62 channels. Each session includes 15 tests distributed evenly and randomized in sequence (3 tests per emotional class), with a total of 3 sessions, totaling 45 tests.

BETA (Liu et al. 2020) dataset involves steady-state visual evoked potentials tasks from 70 subjects, targeting 40 different areas of a virtual keyboard, each area flickering at a different frequency. This experiment involves EEG data from 64 channels. Each run includes 40 tests in random sequence, with 1 session containing 4 runs, totaling 160 tests.

The multi-source dataset is composed of 14,512 images from 49 classes across the above three datasets. It is divided into three parts: 31 base classes for training, 8 novel classes for validation, and the remaining 10 classes for testing.

Implementation Details

We used Conv4 and ResNet-18 as backbones, training all models with a label smoothing factor of 0.1 for cross-entropy loss and an SGD optimizer. batch size=64, epochs=100, the training procedure does not involve meta-learning. The original data points were mapped to an embedding space $\mathbb{R}^M$ with M=512. During evaluation, MSADiff layers $r = 6$, global diffusion intensity $\gamma = 0.5$. In each N-way K-shot task, N new classes were randomly sampled. Then, K samples were drawn from each of these N classes for training, followed by testing with 15 new samples from each class that were distinct from the initial K samples. N was 5, with K set to both 1 and 5. We sampled 300 such tasks. To ensure that all classes appeared in the test set, we conducted 10 rounds of experiments with dataset repartitioning, resulting in a total of 3000 FSL tasks.

Methods	Backbone	BCICIV-2a		SEED-V		BETA	
		k=1	k=5	k=1	k=5	k=1	k=5
PN	Conv4	33.16±0.32*	51.80±1.12*	21.74±0.26*	47.25±1.26*	28.04±0.18*	47.85±1.95*
MSCNN	Conv4	33.71±1.18*	56.65±0.29*	26.22±0.17*	50.25±1.61*	33.41±0.38*	48.18±0.33*
BD-CSPN	Conv4	37.94±0.87*	57.30±1.92*	24.15±1.53*	51.68±1.32*	31.67±1.24*	49.51±1.26*
Meta-DM	Conv4	38.99±2.52*	57.16±1.65*	27.66±0.27*	56.43±1.40*	34.66±0.96*	48.79±2.05*
AFR	Conv4	36.71±1.20*	59.45±0.84*	25.19±1.72*	54.85±2.27*	34.30±1.05*	49.65±1.44*
DA RelationNet	Conv4	38.94±2.04*	59.88±2.14*	26.59±1.35*	56.42±0.38*	35.56±1.23*	50.81±1.80*
Diff-ResNet	Conv4	39.23±0.76*	60.25±0.18*	27.82±1.66*	58.25±0.75*	37.83±1.53*	52.98±1.46*
DiffAlign	Conv4	41.05±1.58*	62.05±1.45*	29.75±1.55*	58.57±0.83*	38.04±0.87*	51.72±0.15*
Proposed	Conv4	**42.93±0.72**	**63.96±0.71**	**30.75±1.08**	**59.57±0.85**	**39.45±0.41**	**54.76±0.80**
PN	ResNet18	45.72±0.75*	64.14±2.09*	35.29±1.42*	58.42±1.62*	38.21±1.82*	56.64±1.83*
MSCNN	ResNet18	45.18±1.08*	68.48±0.80*	35.24±1.31*	67.43±1.32*	37.43±1.10*	58.37±1.84*
BD-CSPN	ResNet18	46.09±1.04*	69.10±1.54*	39.62±0.45*	69.05±1.93*	41.84±1.28*	59.47±1.45*
Meta-DM	ResNet18	45.32±1.87*	70.49±1.26*	39.70±1.37*	68.43±0.38*	44.14±0.55*	61.77±1.80*
AFR	ResNet18	47.69±0.90*	70.86±1.50*	41.64±1.60*	68.95±1.92*	47.53±1.48*	59.60±0.50*
DA RelationNet	ResNet18	51.39±0.78*	71.58±1.94*	39.79±2.15*	70.15±0.90*	44.78±0.59*	61.45±1.31*
Diff-ResNet	ResNet18	48.72±1.79*	71.68±0.76*	42.04±1.17*	70.98±1.91*	47.69±1.39*	60.95±0.91*
DiffAlign	ResNet18	53.40±0.40*	72.85±1.88*	41.75±0.66*	70.39±0.21*	**49.27±1.52***	63.75±0.73*
Proposed	ResNet18	**55.49±0.25**	**75.48±0.98**	**44.31±1.68**	**74.13±0.95**	48.51±1.57	**65.96±1.04**

Table 1: Average accuracies and standard deviations (Acc±Std %) of our model and baselines on the three datasets in the multi-source dataset. p-value between the method and proposed: * indicating ($p < 0.001$).

Experimental Results

The proposed MSADiff-Resnet is a hybrid structure of metric-based and model-based FSL. Therefore, we have selected the most recent methods from these two categories, along with the latest methods in diffusion-based FSL and the most recent techniques for FSL in the BCI field as our baselines. The experimental results are shown in Table 1. Among them, PN (Snell, Swersky, and Zemel 2017), BD-CSPN (Liu, Song, and Qin 2020), AFR (Zhu et al. 2024), Meta-Diffusion Models (Hu et al. 2023), DiffAlign (Roy et al. 2022), Diff-ResNet (Wang et al. 2023), DA RelationNet (An et al. 2023), MSCNN (Pan et al. 2023).

Table 1 displays the results of various SOTA FSL methods on the multi-source BCI dataset. Our MSADiff-Resnet achieved superior or comparable performance across all tasks, outperforming most cutting-edge methods by approximately 1% to 3%. Notably, compared to FSL methods specifically designed for the BCI field (MSCNN and DA RelationNet), our proposed method achieved an improvement of about 4% to 5% in peak performance across different tasks. This marks the exceptional capability of the proposed method, advancing the development of FSL in BCI field. Meanwhile, the proposed method achieves consistent improvements on Conv4 and ResNet18 backbones for most tasks, validating its generalisability.

Ablation Study

To validate the effectiveness of the distribution bias rectification mechanism and the MSADiff mechanism, we conducted an ablation study on the multi-source dataset through 10 experiments, randomly extracting 1000 instances of 5-way 1-shot and 5-way 5-shot tasks, as shown in Table 2 and Figure 4. Specifically, the variant models were categorized as follows: (i) the original Diff-Resnet, which removed both the distribution bias rectification mechanism and the MSADiff mechanism; (ii) Diff-Resnet with a static distribution bias rectification mechanism (Liu, Song, and Qin 2020) but without the MSADiff mechanism; (iii) Diff-Resnet equipped with a dynamic distribution bias rectification mechanism but without the MSADiff mechanism; (iv) Diff-Resnet that included the MSADiff mechanism but removed the distribution bias rectification mechanism.

Methods	Backbone	Multi-Source Dataset	
		k=1	k=5
(i)	Conv4	34.96±1.32*	57.16±0.80*
(ii)	Conv4	35.47±0.25*	57.82±1.51*
(iii)	Conv4	35.93±1.82*	58.25±1.04*
(iv)	Conv4	36.06±1.43*	58.95±2.15*
Proposed	Conv4	**37.71±0.74**	**59.43±0.79**
(i)	ResNet18	46.15±1.45*	67.87±1.19*
(ii)	ResNet18	46.86±1.72*	68.35±1.49*
(iii)	ResNet18	47.44±2.07*	69.38±0.84*
(iv)	ResNet18	48.65±0.71*	70.54±1.06*
Proposed	ResNet18	**49.44±1.17**	**71.86±0.99**

Table 2: Average accuracies and standard deviations (Acc±Std %) of ablation studies. p-value between the method and proposed: * indicating ($p < 0.001$).

The proposed method achieved the best results across all tasks, outperforming variants (iii) and (iv) by approximately 2% and 1%, respectively. This demonstrates the effectiveness of combining dynamic distribution bias rectification mechanism with MSADiff mechanism. Compared to variant (i), variants (ii), (iii), and (iv) also showed improvements

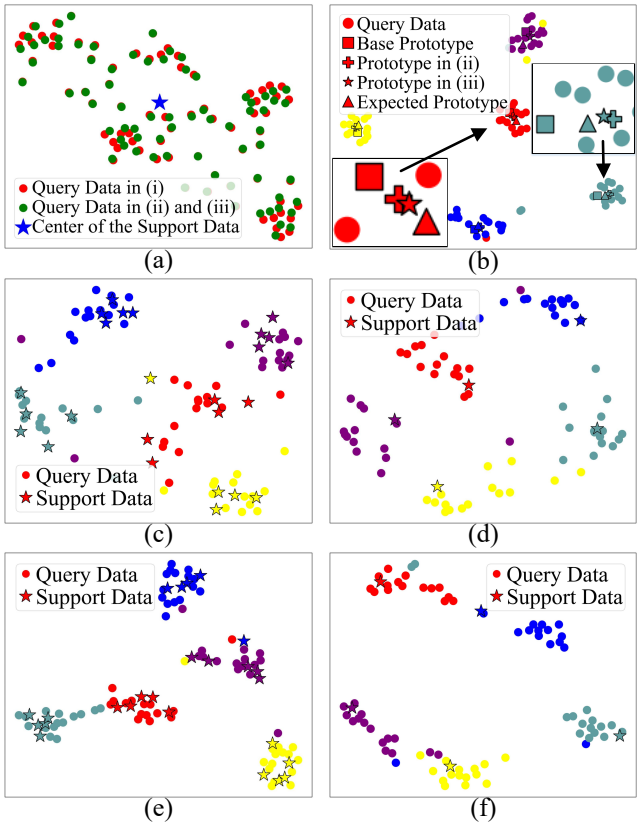

Figure 4: Ablation study (a) Inter-class bias rectification; (b) Intra-class bias rectification; (c) 5-shot, Diff; (d) 1-shot, Diff; (e) 5-shot, MSADiff; (f) 1-shot, MSADiff.

across all tasks by approximately 0.5%, 1%, and 2%, respectively, highlighting the roles of distribution bias rectification mechanism and MSADiff mechanism. The performance of variant (iii) exceeded that of (ii), proving the superior performance of dynamic distribution bias rectification mechanism. Additionally, the performance of variant (iv) surpassed that of both (ii) and (iii), indicating that using MSADiff mechanism to address sample limitations in practical scenarios results in more significant improvements.

Figure 4 (a) and (b) display 5-way 1-shot tasks randomly sampled from a multi-source dataset. Figure 4 (a) seeks to reduce inter-class bias by aligning query set samples closer to the center of the support set. Compared to the red circles, most green circles, which represent variants (ii) and (iii), are closer to the centroid of the support set (blue pentagon), demonstrating the effectiveness of the inter-class bias rectification mechanism. Figure 4 (b) focuses on prototype rectification through the augmentation of the support set, with different colors representing different classes. Key details are magnified to highlight subtleties. Basic prototypes are calculated using support set, while expected prototypes utilize all samples in the collection. Prototypes in variants (ii) and (iii) are derived through static and dynamic intra-class bias rectifications, respectively. Due to the scarcity of labeled samples, there is a noticeable deviation between the basic

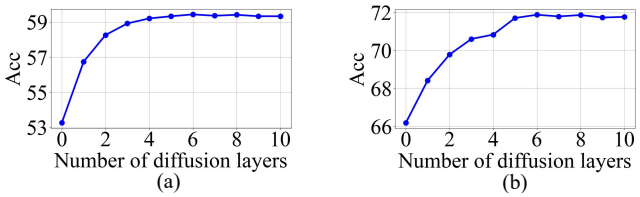

Figure 5: Discussion of diffusion (a) Conv4; (b) ResNet18.

and expected prototypes. The prototype in (iii) (pentagon) is closer to expected prototype (triangle), show the superiority of the dynamic intra-class bias rectification mechanism.

Figure 4 (c) and (d) respectively showcase random 5-way 5-shot and 5-way 1-shot tasks from the multi-source dataset using variant (i), with different classes marked in various colors, support sets indicated by pentagons, and query sets by circles. The boundaries between different classes are clearly delineated, and the clustering of query points within the same class is moderately effective, indicating that the diffusion mechanism provides good separability of the samples. Figure 4 (e) and (f) display the 5-way 5-shot and 5-way 1-shot tasks for variant (iv), respectively. Compared to variant (i), it further improves the clustering effect of similar query points and the separability of the data. This suggests that the MSADiff mechanism more effectively facilitates the exchange of information between labeled and unlabeled data, thereby better attracting similar data points.

Discussion of Diffusion Parameters

Next, we explored the configuration of diffusion layers and global diffusion intensity. Extensive experiments revealed that excessively high global diffusion intensity often leads to anomalous results. To maintain system stability, we fixed the global diffusion intensity at $\gamma = 0.5$ and incrementally increased the number of diffusion layers r. The results from multi-source dataset experiments are displayed in Figure 5.

As observed in Figure 5, across the two backbone networks, the accuracy steadily increases as the number of diffusion layers $r < 6$. At this point, there is a noticeable inflection in the trend, and additional layers do not significantly enhance accuracy further. Given the boundary effects within the model, it appears that information has been sufficiently propagated through the network by the time it reaches 6 layers. To avoid the issue of over-diffusion, it has been determined that 6 is the optimal number of diffusion layers.

Conclusion

To address the challenges of distribution bias and sample limitations in FSL, we introduced the novel MSADiff-Resnet model, which incorporates dynamic distribution bias rectification mechanism and local-global MSADiff mechanism. Additionally, to handle multi-source BCI-based FSL tasks, we proposed a universal deep learning framework. Extensive experiments verified the superiority of the proposed method. In future work, we plan to integrate more temporal information into our framework and aim to develop universal model for multimodal physiological signals.

Acknowledgments

This research was funded in part by STI 2030—Major Projects 2021ZD0200406, the National Natural Science Foundation of China under grant No. 62172122, and the Key Research and Development Program of Heilongjiang Province under grant No. 2022ZX01A19.

References

An, S.; Kim, S.; Chikontwe, P.; and Park, S. H. 2023. Dual attention relation network with fine-tuning for few-shot EEG motor imagery classification. *IEEE Transactions on Neural Networks and Learning Systems*.

Atwood, J.; and Towsley, D. 2016. Diffusion-convolutional neural networks. *Advances in neural information processing systems*, 29.

Avdeyev, P.; Shi, C.; Tan, Y.; Dudnyk, K.; and Zhou, J. 2023. Dirichlet diffusion score model for biological sequence generation. In *International Conference on Machine Learning*, 1276–1301. PMLR.

Bhosale, S.; Chakraborty, R.; and Kopparapu, S. K. 2022. Calibration free meta learning based approach for subject independent EEG emotion recognition. *Biomedical Signal Processing and Control*, 72: 103289.

Brunner, C.; Leeb, R.; Müller-Putz, G.; Schlögl, A.; and Pfurtscheller, G. 2008. Institute for Knowledge Discovery (Laboratory of Brain-Computer Interfaces). *Graz Univ Technol*, 16: 1–6.

Chapin, J. K.; Moxon, K. A.; Markowitz, R. S.; and Nicolelis, M. A. 1999. Real-time control of a robot arm using simultaneously recorded neurons in the motor cortex. *Nature neuroscience*, 2(7): 664–670.

Chen, X.; Ge, C.; Wang, M.; and Wang, J. 2023. Supervised contrastive few-shot learning for high-frequency time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7069–7077.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Fan, Q.; Huang, H.; Chen, M.; Liu, H.; and He, R. 2024. Rmt: Retentive networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5641–5651.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4): 594–611.

Gasteiger, J.; Weißenberger, S.; and Günnemann, S. 2019. Diffusion improves graph learning. *Advances in neural information processing systems*, 32.

Geiser, J. 2009. *Decomposition methods for differential equations: theory and applications*. CRC Press.

Gong, Y. 2023. Meta-learning with differentiable convex optimization. Technical report, EasyChair.

Guo, Q.; Haotong, G.; Wei, X.; Fu, Y.; Yu, Y.; Zhang, W.; and Ge, W. 2023. Rankdnn: Learning to rank for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 728–736.

Heinrichs, F.; Heim, M.; and Weber, C. 2023. Functional Neural Networks: Shift invariant models for functional data with applications to EEG classification. In *International Conference on Machine Learning*, 12866–12881. PMLR.

Hu, W.; Jiang, X.; Liu, J.; Yang, Y.; and Tian, H. 2023. Metadm: Applications of diffusion models on few-shot learning. *arXiv preprint arXiv:2305.08092*.

Lai, J.; Yang, S.; Liu, W.; Zeng, Y.; Huang, Z.; Wu, W.; Liu, J.; Gao, B.-B.; and Wang, C. 2022. tsf: Transformer-based semantic filter for few-shot learning. In *European Conference on Computer Vision*, 1–19. Springer.

Li, H.; Chaudhari, P.; Yang, H.; Lam, M.; Ravichandran, A.; Bhotika, R.; and Soatto, S. 2020. Rethinking the hyperparameters for fine-tuning. *arXiv preprint arXiv:2002.11770*.

Li, H.; Xu, G.; Du, C.; Li, Z.; Han, C.; Tian, P.; Li, B.; and Zhang, S. 2024. Facilitating applications of SSVEP-BCI by effective Cross-Subject knowledge transfer. *Expert Systems with Applications*, 249: 123492.

Li, X.; Sun, Q.; Liu, Y.; Zhou, Q.; Zheng, S.; Chua, T.-S.; and Schiele, B. 2019. Learning to self-train for semi-supervised few-shot classification. *Advances in neural information processing systems*, 32.

Liang, H.; Liu, Y.; Wang, H.; Jia, Z.; and Center, B. 2023. Teacher Assistant-Based Knowledge Distillation Extracting Multi-level Features on Single Channel Sleep EEG. In *IJCAI*, 3948–3956.

Liu, B.; Huang, X.; Wang, Y.; Chen, X.; and Gao, X. 2020. BETA: A large benchmark database toward SSVEP-BCI application. *Frontiers in neuroscience*, 14: 627.

Liu, J.; Song, L.; and Qin, Y. 2020. Prototype rectification for few-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 741–756. Springer.

Liu, W.; Qiu, J.-L.; Zheng, W.-L.; and Lu, B.-L. 2021. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2): 715–729.

Pan, J.; Cai, H.; Huang, H.; He, Y.; and Li, Y. 2023. Multiple scale convolutional few-shot learning networks for online P300-based brain–computer interface and its application to patients with disorder of consciousness. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–16.

Rong, Y.; Lu, X.; Sun, Z.; Chen, Y.; and Xiong, S. 2023. ESPT: a self-supervised episodic spatial pretext task for improving few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 9596–9605.

Roy, A.; Shah, A.; Shah, K.; Roy, A.; and Chellappa, R. 2022. Diffalign: Few-shot learning using diffusion based synthesis and alignment. *arXiv preprint arXiv:2212.05404*, 3.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Song, T.; Liu, S.; Zheng, W.; Zong, Y.; and Cui, Z. 2020. Instance-adaptive graph for EEG emotion recognition. In

Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2701–2708.

Tanaka, H.; Katura, T.; and Sato, H. 2013. Task-related component analysis for functional neuroimaging and application to near-infrared spectroscopy data. *NeuroImage*, 64: 308–327.

Wang, T.; Dou, Z.; Bao, C.; and Shi, Z. 2023. Diffusion mechanism in residual neural network: Theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2): 667–680.

Wang, Y.; Zhang, B.; and Tang, Y. 2024. DMMR: Cross-subject domain generalization for EEG-based emotion recognition via denoising mixed mutual reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 628–636.

Wang, Z.; and Oates, T. 2015. Imaging time-series to improve classification and imputation. *arXiv preprint arXiv:1506.00327*.

Wu, J.; Liu, X.; Yin, X.; Zhang, T.; and Zhang, Y. 2024. Task-Adaptive Prompted Transformer for Cross-Domain Few-Shot Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6012–6020.

Yang, Y.; Kim, T.; and Yun, S.-Y. 2024. Leveraging Normalization Layer in Adapters With Progressive Learning and Adaptive Distillation for Cross-Domain Few-Shot Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16370–16378.

Zhao, L.-M.; Yan, X.; and Lu, B.-L. 2021. Plug-and-play domain adaptation for cross-subject EEG-based emotion recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 863–870.

Zhu, X.; Wang, S.; Lu, J.; Hao, Y.; Liu, H.; and He, X. 2024. Boosting Few-Shot Learning via Attentive Feature Regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7793–7801.