DMMR: Cross-Subject Domain Generalization for EEG-Based Emotion Recognition via Denoising Mixed Mutual Reconstruction

Yiming Wang, Bin Zhang*, Yujiao Tang

School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China

{yimingwang, 3121358009}@stu.xjtu.edu.cn, bzhang82@xjtu.edu.cn

Abstract

Electroencephalography (EEG) has proven to be effective in emotion analysis. However, current methods struggle with individual variations, complicating the generalization of models trained on data from source subjects to unseen target subjects. To tackle this issue, we propose the Denoising Mixed Mutual Reconstruction (DMMR) model, employing a two-stage pre-training followed by fine-tuning approach. During the pre-training phase, DMMR leverages self-supervised learning through a multi-decoder autoencoder, which encodes and reconstructs features of one subject, aiming to generate features resembling those from other subjects within the same category, thereby encouraging the encoder to learn subject-invariant features. We introduce a hidden-layer mixed data augmentation approach to mitigate the limitations posed by the scarcity of source data, thereby extending the method to a two-stage process. To bolster stability against noise, we incorporate a noise injection method, named "Time Steps Shuffling", into the input data. During the fine-tuning phase, an emotion classifier is integrated to extract emotion-related features. Experimental accuracy on the SEED and SEED-IV datasets reached 88.27% (±5.62) and 72.70% (±8.01), respectively, demonstrating state-of-the-art and comparable performance, thereby showcasing the superiority of DMMR. The proposed data augmentation and noise injection methods were observed to complementarily enhance accuracy and stability, thus alleviating the aforementioned issues.

Introduction

Human emotions are closely related to health conditions and behavioral patterns, such as Autism Spectrum Disorder (Mayor-Torres et al. 2021) and depression (Bocharov, Knyazev, and Savostyanov 2017), as well as malicious behaviors resulting from the accumulation of negative emotions. Real-time monitoring of individuals' emotional states can contribute to objective health assessments and early warning of malicious behaviors. Due to the difficulty in disguising physiological signals, wearable devices are commonly used to monitor emotion-related physiological signals, such as EEG, eye movements (Lu et al. 2015), facial electromyography (Chen et al. 2015), etc. Among them, EEG-based emotion recognition has become a crucial means of emotion identification due to its high temporal resolution and accuracy.

EEG signals from specific channels and frequency bands demonstrate different responses to various emotional stimuli (Zheng and Lu 2015), making it possible to detect fine-grained emotional tendencies. However, individuals differ in their cranial structure and emotional experiences, which result in varying sensitivity to the same emotion. Consequently, there are significant distributional differences among data from different subjects, making it challenging for a model trained on source subject data to generalize effectively to target subjects. To solve this problem, transfer learning and other methods are employed to extract subject-invariant emotion features, with the expectation that emotional knowledge can be effectively transferred to target subjects from data of source subjects. Early works assumed that a large amount of unlabeled data or a small amount of labeled data from the target subjects is available, explicitly narrowing the distribution gap between source and target subjects. These methods are known as unsupervised domain adaptation (Luo and Lu 2021; Li et al. 2018a) and semi-supervised domain adaptation approaches (Li et al. 2020a). However, these approaches rely on the target subject's data to train the model, making them less user-friendly. A more challenging task is domain generalization (DG), which assumes that the target subjects are entirely unseen during the training process, so the model is trained solely on the source subject's data to create a subject-invariant and robust model.

There are two main types of cross-subject DG methods for EEG-based emotion recognition: Sample-intrinsic subject-invariant feature extraction methods (SI-SIFE) and Cross-subject subject-invariant feature extraction methods (CS-SIFE). The two methods are related to two non-inclusive categories of features: internally-invariant and mutually-invariant (Lu et al. 2022). The former extracts features

*Corresponding author

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

from a single subject, the later takes into account the differences and traits among subjects, treating each subject as a separate domain, thereby incorporating traditional DG techniques to extract subject-invariant features. Despite related methods' great achievements, several issues remain:

- Involve joint training (simultaneously training subject-independent feature extraction and emotion recognition tasks), where task interference may impede the extraction of subject-independent features.
- Tend to overfit to the scarce source data, resulting in poor generalization performance on unseen subject data.
- Not emphasize the model's robustness against potential noise in the data collection process.

Inspired by DiMAE (Yang et al. 2022), this paper proposes a new CS-SIFE model named DMMR to address the aforementioned issues. We introduce the "mutual reconstruction" method for EEG-based cross-subject emotion recognition. It reconstructs one subject's features into another's, which is a novel approach in the field. Unlike the Joint training mode, DMMR combines two-stage pretraining with mutual reconstruction self-supervised learning for the first time, addressing individual variability in EEG signals. To tackle the problem of scarce source data, we propose a mixed data augmentation strategy extending the mixup technique (Zhang et al. 2018), boosting data generalization without the need for extra parameters by creating new subject features in the hidden layer, improving recognition accuracy. To enhance the model's stability against noise, we incorporate the denoising task proposed by (Vincent et al. 2008). This involves relearning clean samples from noise-distorted features. We propose an EEG-tailored noise injection method enhancing denoising while preserving essential information, improving stability.

The main contributions of this paper are as follows:

- Proposing the DMMR model with a pre-training-fine-tuning paradigm, which extends the mutual reconstruction method by proposing a novel mix data augmentation approach in the hidden layer and a noise injection method named Time Steps Shuffling.
- The experimental accuracy on the SEED (Zheng and Lu 2015) and SEED-IV (Zheng et al. 2019) datasets reached 88.27% (±5.62) and 72.70% (±8.01), respectively, achieving state-of-the-art and comparable performance, demonstrating the superiority of DMMR.
- The proposed data augmentation and noise injection methods are observed to complementarily enhance accuracy and stability, alleviating the issues of the scarcity of source data and potential noise interference.

Related Work

This section describes the two subject-invariant feature extraction methods (SIFE): SI-SIFE and CS-SIFE. And analyze the differences between our approach and related work.

SI-SIFE methods always consider the correlations between different channels from a single subject. The method assumes cross-subject invariance in channel correlations. One strategy constructs inter-channel connections into a sequence manually, utilizing LSTM for high-dimensional emotional features (Li et al. 2020b). Others employ distance metrics or trainable parameters to create adjacency matrices, employing graph neural networks for high-dimensional emotional semantics (Song et al. 2018; Zhong, Wang, and Miao 2020; Zhang et al. 2021; Priyasad et al. 2022). Building upon this, GMSS (Li et al. 2022) enhances data using self-supervised learning, employing jigsaw tasks for robust intrinsic feature extraction and utilizes unsupervised contrastive learning methods for distance manipulation.

CS-SIFE methods treat each subject as a different domain. DG-DANN (Ma et al. 2019) employs gradient reversal (Ganin and Lempitsky 2015) to confuse subject discriminators and yielding subject-invariant features. Notably, it aligns the Jensen-Shannon divergence, implicitly aligns the marginal probability distributions across multiple subjects (Li et al. 2018b). DResNet (Ma et al. 2019) further combines residuals from subject-specific and subject-shared encoders for emotion classification. In the case of shared features in same-label samples across subjects, methods like contrastive learning manipulate sample distances. Clisa (Shen et al. 2022) uses contrastive learning to minimize distances between samples from the same emotional stimulus and maximize distances between different stimuli, ensuring consistent representations for identical emotional stimuli.

However, the aforementioned methods did not address three critical issues: constrained joint training, overfitting to scarce source data, and practical noise robustness needs. In image processing, Ghifary et al. (Ghifary et al. 2015) proposed MTAE using multiple decoders for mutual reconstruction. During the pre-training process, they employed a single domain's feature as input and assigned specific decoders for each source domain, aiding domain-invariant feature generation. Yang et al. (Yang et al. 2022) extended this with DiMAE, using CP-styleMix for data augmentation and a mask mechanism for visible parts. Differing from these methods, this paper further proposes hidden layer output mixing for data augmentation and a custom noise injection method for EEG signals to address the mentioned challenges.

Methods

We let $X_s = \{X_s^i, Y_s^i\}_{i=1}^n$ to be the data and labels of n source subjects, the model trained with the data of source subjects is used to predict the emotion class for the unseen target subject X_t. Similar to the data preprocessing method of (Zhao, Yan, and Lu 2021), to fully exploit the EEG data's temporal features, we employ overlapping sliding windows along the time axis with time steps T, thus a sample to be fitted into

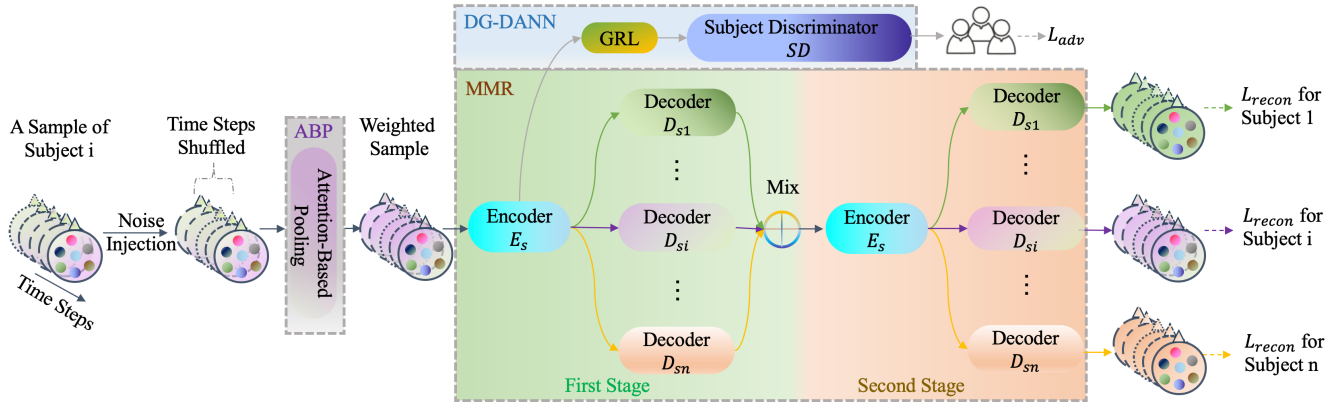

Figure 1. The pre-training phase of DMMR, it aimed at extracting subject-invariant features from multiple source subjects' data in a self-supervised mode, which consists of a noise injection process and three different modules, namely the ABP module, the MMR module and the DG-DANN module

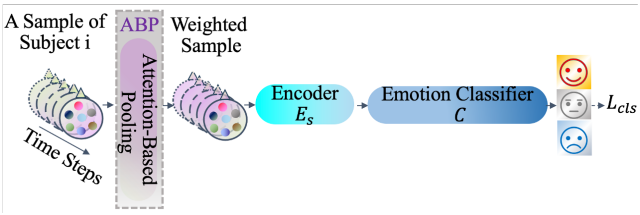

Figure 2. The fine-tuning phase of DMMR. Subject-invariant emotional features are further extracted, which are classified into different emotions by a new emotion classifier.

the model is represented as $x = (x_1, x_2 \cdots x_T) \in R^{T*D}$, among them, $D = C * B$ represents the number of feature dimensions, C represents the number of channels, and B represents the number of frequency bands.

Figure 1 and Figure 2 show the pre-training and fine-tuning phases of the proposed framework DMMR, respectively. The solid arrows indicate the flow of data, while the dashed arrows represent the losses to be computed. In the pre-training phase, to solve the potential noise problem, a noise injection process is used to the input, followed by three different modules, namely the ABP module, the DG-DANN module and the Mixed Mutual Reconstruction (MMR) module. In the ABP module, we introduce the way of (Zhao, Yan, and Lu 2021), use the self-attention method to add weights for important channels and frequent bands of a sample. In the MMR module, an encoder E_s is used to extract shared features among subjects. Like the way of (Yang et al. 2022), we set n special decoder $D_s = \{D_s^i\}_{i=1}^n$ to reconstruct features of different subjects within the same category. To solve the problem of the scarcity of source data, we further propose a two-stage mixed mutual reconstruction approach. In the first stage, the outputs of multiple decoders are mixed to generate a new mix-subject data. In the second stage, the mixed output is fed back into the multi-decoder autoencoder. To supervise the decoders with self-supervised learning, we utilize sampled samples of the same category taken from different subjects. These samples are then processed through the ABP module to generate weighted features, which can be considered as representations of the same category features in different subjects. The DG-DANN module (Ma et al. 2019) is an application of the DANN model (Li et al. 2018a) to DG problems. It leverages a multi-domain discriminator (since we treat each subject as an individual domain, we rename it as the Subject Discriminator SD) and domain adversarial techniques to extract subject-invariant features. In the fine-tuning phase, only the ABP module and encoder E_s is preserved, and a new emotion classifier C is added to achieve emotion classification. It further leverages emotion category labels for supervised learning, enabling the model to extract emotion-related but subject-invariant features. Accordingly, algorithm 1 summarizes the pseudo-code of DMMR. The testing process directly employs the fine-tuned model for evaluation, so we won't provide additional details in the following text and pseudo-code.

Noise Injection

In order to enhance the stability of the model against noise, we propose a method called "Time Steps Shuffling" to inject noise into the input samples. This method shuffles only the order of the temporal dimension, leaving the feature dimensions within individual time steps unaffected. Considering that the encoder structure utilizes a unidirectional LSTM, the final time step is retained more than other time steps and represents the current moment's emotional state. Hence, we fix the final time step and only shuffle others to preserve the essential characteristics of the input samples. To distinguish it from the original input feature x, we use x_{noised} to represent the noise injected feature.

The ABP Module

The ABP module is a bottom-layer feature weighting method proposed by (Zhao, Yan, and Lu 2021). It assigns

Algorithm 1: DMMR method

Input:
Iteration T1, T2.
Source data $X_s = \{X_s^i, Y_s^i\}_{i=1}^n$.
Output: optimal DMMR model
The Pre-Training Phase:
Randomly initialize ABP, E_s, $D_s^{1\sim n}$ and SD.
for t=1: T1 **do**
  Inject noise to source data.
  Optimize ABP, E_s, $D_s^{1\sim n}$ and SD by minimizing Equation (10).
end for
return ABP, E_s.
The Fine-Tuning Phase:
Randomly initialize C.
Obtain the pre-trained ABP, E_s.
for t=1: T2 **do**
  Optimize ABP, E_s and C by minimizing Equation (12).
end for
return ABP, E_s and C

weights to channels and frequency bands of samples automatically using a self-attention weighting method. This method introduces a linear layer to map the original features to a new feature space. The weights are obtained by normalizing the mapped features using the Softmax function. As shown in Formula 1, in which $W_0 \in R^{D*D}$ and $b_0 \in R^{D*1}$ represent trainable weights and bias in the linear layer, respectively, $\alpha \in R^{D*1}$ is a one-dimensional attention weight. The weighted features are obtained through element-wise multiplication of the normalized weights with the original features, as shown in Formula 2. $\tilde{x} \in R^{T*D}$ represents the weighted feature.

$$\alpha = softmax(W_0 x_{noised} + b_0) \quad (1)$$

$$\tilde{x} = \alpha \cdot x_{noised} \quad (2)$$

The MMR Module

The MMR module defines a shared encoder for all subjects and different decoders for each source subject. After extracting high-dimensional features with the encoder, these features are reconstructed into features of different subjects within the same category using different decoders. This forces the encoder to learn subject-invariant representations for specific emotions. As the reconstruction loss requires the use of emotional labels from the source subjects, this method is a self-supervised learning approach.

Both the encoder and multiple decoders take the weighted features from the ABP module's output corresponding to the subject as their input and supervision, respectively, creating a symmetric and mutually reconstructive autoencoder structure. The encoder and decoders in this paper are single-layer LSTM models. The construction of the decoder is consistent with that in (Zhao, Yan, and Lu 2021), generating features for each time step in reverse order and then using a linear layer to map the features to the same dimension as the input of the encoder. To simplify the description, we formally define the encoder and decoders in Equation 3. $o_1^i \in R^{T*D}$ represents the output representation of the i-th decoder in the first stage.

$$o_1^i = D_{si}(E_s(\tilde{x})),\ i \in \{1,2,\cdots n\} \quad (3)$$

To further address the scarcity of source data, we draw inspiration from the mixup technique (Zhang et al. 2018), which linearly combines different samples and labels to create new samples for data augmentation, we propose a two-stage mixed mutual reconstruction method. In the first stage, the outputs of each decoder are summed directly to obtain new mix-subject features $x_{mix} \in R^{T*D}$, as shown in Equation 4. These features are linear combinations of same-category features from different subjects, creating new subject features unseen by the model while preserving their label information.

$$x_{mix} = \sum_{i=1}^n o_1^i \quad (4)$$

In the second stage, we reconstruct these mix-subject features into features of different subjects within the same category using the encoder and multi-decoder structure defined above, as shown in Equation 5. $o_2^i \in R^{T*D}$ represents the output representation of the i-th decoder in the second stage.

$$o_2^i = D_{si}(E_s(x_{mix})),\ i \in \{1,2,\cdots n\} \quad (5)$$

Since the encoder and decoders between the two stages are parameter-sharing, we only need to calculate the reconstruction loss for each subject after the second stage. The Mean Squared Error (MSE) loss is employed to quantify the differences between the generated features and the corresponding subject features, as shown in Equation 6. In which $r^i \in R^{T*D}$ represents the representation of the i-th subject's features (without noise injection) after being weighted by the ABP module. These representations share the same labels as x and are used as supervision for the corresponding decoder. The final reconstruction loss l_{recon} is the sum of n individual reconstruction losses, as shown in Equation 7.

$$l_{recon}^i = MSE(o_2^i, r^i),\ i \in \{1,2,\cdots n\} \quad (6)$$

$$l_{recon} = \sum_{i=1}^n l_{recon}^i \quad (7)$$

The DG-DANN Module

We utilize the DG-DANN method from (Ma et al. 2019) to establish a multi-class subject discriminator SD (a single-layer fully connected network) for discerning feature ownership. Each subject receives a distinct ID for supervision during training. A gradient reversal layer (GRL) is added before the discriminator, multiplying gradients by $-\lambda$ in the backpropagation process. This confounds the discriminator and encourages subject-insensitive feature extraction by the encoder. This adversarial interplay between encoder and discriminator strives for a Nash equilibrium, enabling the encoder to generate subject-invariant features. It's important

to note that the DG-DANN module is limited to the first stage to avoid interfering with decoder feature generation. The process for feature ownership prediction is shown in Equation 8. $\hat{d}^i$ represents the Softmax predictions of SD. The computation of the adversarial loss l_{adv} is presented in Equation 9, in which d^i is the ID of the i-th subject.

$$\hat{d}^i = SD(E_s(\tilde{x})),\ i \in \{1,2,\cdots n\} \quad (8)$$

$$l_{adv} = -\lambda d^i log(\hat{d}^i),\ i \in \{1,2,\cdots n\} \quad (9)$$

Learning Loss in The Pretraining Phase

The final pre-training loss combines the reconstruction loss and adversarial loss using a balancing hyperparameter β, as shown in Equation 10.

$$l_{pre-training} = l_{recon} + \beta * l_{adv} \quad (10)$$

The Fine-Tuning Phase

The fine-tuning stage aims to further extract subject-invariant emotion features. The input data is the same as in the pre-training phase, and noise injection is no longer needed. We take the ABP module and encoder from the pre-trained model and add an emotion classifier C (a single-layer fully connected network) on top of the encoder's output to get the emotion prediction result. All parameters need to be fine-tuned in order to obtain distinct emotion category boundaries in the extracted features. We use the original emotion labels for supervised learning. As shown in Formula 11. $\hat{y}^j \in R^{c*1}$ is the Softmax predictions of emotion and c is the number of emotion classes. l_{cls} is the cross-entropy loss for emotion classification, as shown in Formula 12. Where $y^i \in R^{c*1}$ is the ground truth emotion label.

$$\hat{y}^j = C(E_s(ABP(x))),\ j \in \{1,2,\cdots c\} \quad (11)$$

$$l_{cls} = y^j log(\hat{y}^j)\ \ j \in \{1,2,\cdots c\} \quad (12)$$

Experiments

Datasets

We evaluated the performance of the DMMR model on two publicly available datasets, SEED and SEED-IV. Both datasets involve inducing EEG signals by presenting specific emotional videos. The SEED dataset consists of 15 different emotional videos, covering three emotion categories (positive, negative, and neutral). The SEED-IV dataset includes 24 different emotional videos, covering four emotion categories (happy, sad, neutral, and fearful). The experiments were conducted across three separate sessions. For data acquisition, both datasets utilized the ESI NeuroScan system, following the international 10-20 standard, to collect 62-channel EEG signals. The EEG signals were downsampled of 0-75Hz. These filtered signals were further divided into five frequency bands: δ: 1-3 Hz, θ: 4-7 Hz, α: 8-13 Hz, β: 14-30 Hz, and γ: 31-50 Hz. To extract frequency domain features from the EEG signals, a non-overlapping sliding window of 1 second size was applied in the raw signals, and in each window, the Differential Entropy (DE) feature was extracted. This process mapped every 1-second data to a 310-dimensional feature space (62 channels * 5 frequency bands), the data sizes for a single experiment are approximately 3400 and 830 for the two datasets, respectively.

Method	SEED		SEED-IV	
	Avg.	Std.	Avg.	Std.
DGCNN	79.95	9.02	-	-
BiHDM*	81.55	9.74	67.47	8.22
RGNN*	81.92	9.35	71.65	9.34
GMSS	86.52	6.22	**73.48**	**7.41**
DG-DANN	84.30	8.32	-	-
DResNet	85.30	7.97	-	-
Clisa	86.40	6.40	-	-
PPDA-NC	85.40	7.10	-	-
PPDA	86.70	7.10	-	-
DMMR	**88.27**	**5.62**	72.70	8.01

Table 1. Performance comparison of our proposed DMMR with baselines (%)

Implementation Details

We utilized the DE features from the first session of all subjects in both datasets. For evaluation, we adopted a leave-one-subject-out cross-validation approach, where we used one subject as the test set while using the remaining subjects as the training set. The average accuracy (avg.) and standard deviation (std.) were calculated across all subjects after each subject served as the target subject once.

Regarding hyperparameters, for both datasets, the input data has a feature dimension of 310, and the hidden size of the encoder is 64. The balancing hyperparameter β is set to 0.05. We utilize the Adam optimizer with a learning rate of 1e-3 and a weight decay rate of 5e-4. Due to the different sizes of the SEED and SEED-IV dataset, the batch sizes are set to 512 and 256, while time steps T are set to 30 and 10, respectively. To ensure reproducibility, the random seed for all experiments is fixed at 3. All experiments were implemented using PyTorch on a Nvidia Tesla V-100 GPU. Code of DMMR is at https://github.com/CodeBreathing/DMMR.

Experiment Results

Comparison with baseline methods The performance comparison between DMMR and baseline models is shown in Table 1. In order to comprehensively compare with relevant baseline methods, we selected the best-performing models in two categories of SIFE methods. Among them, DGCNN (Song et al. 2018), BiHDM (Li et al. 2020b), RGNN (Zhong, Wang, and Miao 2020), and GMSS (Li et al. 2022) are SI-SIFE methods. In which BiHDM and RGNN are unsupervised domain adaptation models, which

Method	SEED		SEED-IV	
	Avg.	Std.	Avg.	Std.
DMMR	**88.27**	5.62	72.70	**8.01**
w/o noise	87.15	6.02	**72.93**	12.16
w/o mix	85.42	5.89	71.12	10.06
w/o both	85.25	5.29	71.12	11.63
Mask Time Steps	85.46	5.61	70.83	10.79
Channels Shuffling	86.83	5.79	71.65	9.49
Mask Channels	86.54	**4.79**	72.41	9.69
Dropout	86.37	5.43	71.21	11.54

Table 2. Ablation Studies and Performance Comparison of Different Noise Injection Methods (%)

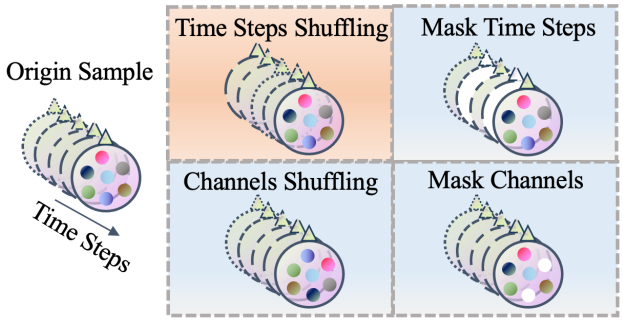

Figure 3. Different Noise injection methods, with the Time Steps Shuffling being the method used in DMMR.

can be used for DG tasks when the gradient reversal layer is removed. We use an asterisk "*" to indicate their performance in DG tasks. DG-DANN (Ma et al. 2019), DResNet (Ma et al. 2019), and Clisa (Shen et al. 2022) are CS-SIFE methods, where the first two are classical joint learning DG methods, and Clisa is a special method that generates more robust DE features through contrastive learning. Additionally, Some basic modules of DMMR and PPDA (Zhao, Yan, and Lu 2021) are shared for their excellent performance shown in PPDA, like the ABP module and DG-DANN module. Specifically, the two models both extract temporal features from the DE features using sliding windows and employ LSTM autoencoders. Therefore, PPDA is included as one of the baselines for comparison. However, unlike DG methods, PPDA calibrates the pre-trained model using a small amount of unlabeled data from the target subject. When the calibration part is eliminated, the resulting PPDA_NC model can achieve DG tasks. Differently, DMMR utilizes inter-subject mutual reconstruction to extract subject-invariant features, rather than PPDA's self-reconstruction approach, which only ensures the robustness of the generated features.

By comparing with the optimal results of two SIFE methods, the proposed DMMR model achieves the highest accuracy of 88.27% and the lowest standard deviation of 5.62 on the SEED dataset. Particularly, the performance of the DMMR model outperforms the PPDA model. However, there have been almost no reports on the performance of CS-SIFE methods on the SEED-IV dataset. By comparing with the current state-of-the-art SI-SIFE methods, we find that our method's performance of 72.70% (±8.01) is only slightly inferior to the GMSS model, which could be due to the smaller dataset, which limits the efficiency of the tailored data augmentation method. The performance comparison on both datasets demonstrates the effectiveness of our approach.

Ablation studies To analyze the effectiveness of our proposed noise injection method and data augmentation technique, we conducted ablation experiments on both datasets, as shown in the upper side of the double solid line in Table 2. "w/o noise" refers to the ablation of the noise injection module; "w/o mix" refers to the ablation of the mixed data augmentation module. In this case, the two-stage mixed mutual reconstruction reduces to a one-stage mutual reconstruction, where the reconstruction loss is directly applied to the first output of multiple decoders; "w/o both" means that both the two modules are ablated.

It is observed that both modules contribute to the overall performance improvement. The data augmentation method based on multi-decoder fusion shows a more significant effect on accuracy enhancement, while the noise injection method significantly reduces the standard deviation, indicating that both modules complement each other. Particularly, although ablating the noise injection module leads to slightly higher accuracy on the SEED-IV dataset compared to the DMMR model, its standard deviation is considerably higher, indicating that this method lacks stability.

Noise injection methods comparison To validate that our proposed noise injection method is more suitable for subject-invariant feature extraction in the EEG based emotion recognition task, Figure 3 illustrates the Time Steps Shuffling method proposed in this paper, along with three other random noise injection methods, including "Mask Time Steps" (excluding the last time step), "Channels Shuffling", "Mask Channels". It represents different time steps with various dashed and solid lines, and different colored circles indicate different channels, with white color indicating a situation where a time step or channel is masked out. In this case, the features of each time step are composed of C-dimensional channels and B-dimensional frequency bands. In order to manipulate the channel dimension, it is necessary to reshape the shape into (T, C, B), and then apply noise injection methods only along the C dimension. Additionally, Dropout is also a commonly used noise augmentation technique, which randomly drops a corresponding rate of channels or frequency bands at each time step. Therefore, we also include this method for comparison. Both the masking and dropout ratios are set to 20%. From the results in the lower side of the double solid line in Table 2, we found that our

proposed noise injection method performs the best, while the other four noise injection methods even perform worse than the "w/o noise" baseline, suggesting that for EEG data, a careful selection of the noise injection method is essential.

In particular, the mask-based and dropout methods both lead to a certain degree of information loss. We speculate that due to the relatively small size of the dataset, the negative impact of information loss is more pronounced. Moreover, the "Channels Shuffling" method confuses the boundaries among dimensions, making the data boundaries within the same dimension less distinct. The "Time Steps Shuffling" used in DMMR does not cause any information loss and has no impact on the features at individual time steps. As a result, it achieves better performance compared to other methods.

Visualization of extracted features To assess whether our model can extract subject-invariant features and clearly delineate the boundaries of emotion categories, we employed the T-SNE algorithm (Van der Maaten and Hinton 2008) on the randomly selected 50 samples from data of each source subject of the SEED dataset. We visualized the subject distribution and emotion distribution of the original features and the features extracted by our pre-trained model and fine-tuned model. The resulting plots are shown in Figure 4. In the subject distribution, each subject is represented by a unique color, notably, the target subject is represented in red; in the emotion distribution, positive, neutral, and negative correspond to orange, light blue, and dark blue, respectively.

From the perspective of subject distributions, the original distribution of subjects is relatively scattered with little overlap. However, following self-supervised pretraining, there is a high degree of feature overlap among subjects, and their distributions tend to become consistent. After fine-tuning, influenced by emotional categories, three clusters emerge, yet the distributions of subjects within each category still exhibit significant overlap. Regarding emotional distributions, the original emotional distribution shows considerable overlap. Since the pretraining process only focuses on extracting subject-invariant features, the issue of overlapping emotional distributions remains unaddressed. Nevertheless, through the fine-tuning process, features corresponding to the distinct emotional categories cluster separately with clear boundaries.

The aforementioned observations show that the pre-trained model captures subject-invariant features. The fine-tuning process further delineates category boundaries and captures subject-invariant features within each category. Particularly, based on the post-fine-tuning subject distribution graph (bottom left corner), the distribution of target subject data highly overlaps with the distribution of source subjects, indicating that the DMMR is effective in aligning the distributions between source subjects and unseen target subjects. From the post-fine-tuning emotional distribution plot (bottom right corner), it is observed that the features corresponding to positive-neutral and negative-neutral emotional pairs are more prone to confusion, aligning with the process of continuous emotional transitions.

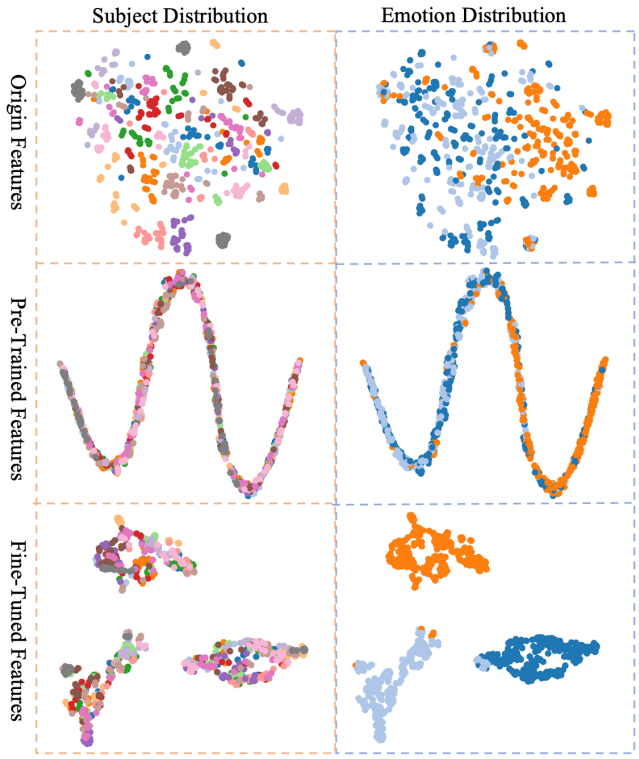

Figure 4. Features visualization. From left to right, we have the subject distribution and emotional distribution. From top to bottom, we show the original DE features, features after pre-training, and features after fine-tuning.

Conclusion and Future Work

In conclusion, this paper introduces the DMMR model, a novel approach for EEG-based emotion recognition that addresses critical challenges in DG. The model leverages a two-stage pre-training followed by fine-tuning approach, incorporating self-supervised learning through mutual reconstruction to extract subject-invariant features. To address the problem of the scarcity of source data and potential noise in the data, this paper proposes a method of mixed data augmentation at the hidden layer and a noise injection method called Time Steps Shuffling. The experimental accuracy on the SEED and SEED-IV datasets reached 88.27% (±5.62) and 72.70% (±8.01), respectively, achieving state-of-the-art and comparable performance. The methods for data augmentation and noise injection have been observed to effectively enhance both accuracy and stability, providing complementary solutions to the aforementioned issues.

In the future, we plan to explore DG methods under the condition of limited annotated EEG data for source subjects, accelerating the practical application of relevant techniques.

Acknowledgments

This work is supported by the Key Research and Development Program of Shaanxi (ProgramNo.2022GY-075), and the National Key Projects of China (ProgramNo. 2021XJTU0016).

References

Bocharov, A. V.;Knyazev, G. G.; and Savostyanov, A. N. 2017. Depression and implicit emotion processing: An EEG study. *Neurophysiologie Clinique/Clinical Neurophysiology* 47(3):225-230.

Chen, J.;Hu, B.;Xu, L.;Moore, P.; and Su, Y. 2015. Feature-level fusion of multimodal physiological signals for emotion recognition. In 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): IEEE, 395-399.

Ganin, Y., and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In International conference on machine learning: PMLR, 1180-1189.

Ghifary, M.;Kleijn, W. B.;Zhang, M.; and Balduzzi, D. 2015. Domain generalization for object recognition with multi-task autoencoders. In Proceedings of the IEEE international conference on computer vision, 2551-2559.

Li, H.;Jin, Y.-M.;Zheng, W.-L.; and Lu, B.-L. 2018a. Cross-Subject Emotion Recognition Using Deep Adaptation Networks. In Neural Information Processing: 25th International Conference, ICONIP 2018, 403-413. doi.org/10.1007/978-3-030-04221-9_36.

Li, J.;Qiu, S.;Shen, Y. Y.;Liu, C. L.; and He, H. 2020a. Multisource Transfer Learning for Cross-Subject EEG Emotion Recognition. *IEEE Transactions on Cybernetics* 50(7):3281-3293. doi.org/10.1109/TCYB.2019.2904052.

Li, Y.;Chen, J.;Li, F.;Fu, B.;Wu, H.;Ji, Y.;Zhou, Y.;Niu, Y.;Shi, G.; and Zheng, W. 2022. GMSS: Graph-Based Multi-Task Self-Supervised Learning for EEG Emotion Recognition. *IEEE Transactions on Affective Computing*.

Li, Y.;Tian, X.;Gong, M.;Liu, Y.;Liu, T.;Zhang, K.; and Tao, D. 2018b. Deep domain generalization via conditional invariant adversarial networks. In Proceedings of the European conference on computer vision (ECCV), 624-639.

Li, Y.;Wang, L.;Zheng, W.;Zong, Y.;Qi, L.;Cui, Z.;Zhang, T.; and Song, T. 2020b. A novel bi-hemispheric discrepancy model for EEG emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems* 13(2):354-367.

Lu, W.;Wang, J.;Li, H.;Chen, Y.; and Xie, X. 2022. Domain-invariant Feature Exploration for Domain Generalization. *Transactions on Machine Learning Research*.

Lu, Y.;Zheng, W.-L.;Li, B.; and Lu, B.-L. 2015. Combining eye movements and EEG to enhance emotion recognition. In Twenty-Fourth International Joint Conference on Artificial Intelligence, 1170-1176.

Luo, Y., and Lu, B.-L. 2021. Wasserstein-Distance-Based Multi-Source Adversarial Domain Adaptation for Emotion Recognition and Vigilance Estimation. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1424-1428. doi.org/10.1109/bibm52615.2021.9669383.

Ma, B.-Q.;Li, H.;Zheng, W.-L.; and Lu, B.-L. 2019. Reducing the subject variability of eeg signals with adversarial domain generalization. In International Conference on Neural Information Processing: Springer, 30-42.

Mayor-Torres, J. M.;Ravanelli, M.;Medina-DeVilliers, S. E.;Lerner, M. D.; and Riccardi, G. 2021. Interpretable sincnet-based deep learning for emotion recognition from eeg brain activity. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC): IEEE, 412-415.

Priyasad, D.;Fernando, T.;Denman, S.;Sridharan, S.; and Fookes, C. 2022. Affect recognition from scalp-EEG using channel-wise encoder networks coupled with geometric deep learning and multi-channel feature fusion. *Knowledge-Based Systems* 250:109038.

Shen, X.;Liu, X.;Hu, X.;Zhang, D.; and Song, S. 2022. Contrastive learning of subject-invariant EEG representations for cross-subject emotion recognition. *IEEE Transactions on Affective Computing*.

Song, T.;Zheng, W.;Song, P.; and Cui, Z. 2018. EEG emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing* 11(3):532-541.

Van der Maaten, L., and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9(11):2579-2605.

Vincent, P.;Larochelle, H.;Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning, 1096-1103.

Yang, H.;Tang, S.;Chen, M.;Wang, Y.;Zhu, F.;Bai, L.;Zhao, R.; and Ouyang, W. 2022. Domain Invariant Masked Autoencoders for Self-supervised Learning from Multi-domains. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI: Springer, 151-168.

Zhang, G.;Yu, M.;Liu, Y.-J.;Zhao, G.;Zhang, D.; and Zheng, W. 2021. SparseDGCNN: Recognizing Emotion from Multichannel EEG Signals. *IEEE Transactions on Affective Computing*:537–548. doi.org/10.1109/taffc.2021.3051332.

Zhang, H.;Cisse, M.;Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In International Conference on Learning Representations, 1-13.

Zhao, L. M.;Yan, X.; and Lu, B. L. 2021. Plug-and-play domain adaptation for cross-subject EEG-based emotion recognition. In the AAAI Conference on Artificial Intelligence, 863-870.

Zheng, W. L.;Liu, W.;Lu, Y.;Lu, B. L.; and Cichocki, A. 2019. EmotionMeter: A Multimodal Framework for

Recognizing Human Emotions. *IEEE Trans Cybern* 49(3):1110-1122. doi.org/10.1109/TCYB.2018.2797176.

Zheng, W. L., and Lu, B. L. 2015. Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. *IEEE Transactions on Autonomous Mental Development* 7(3):162-175. doi.org/10.1109/tamd.2015.2431497.

Zhong, P.;Wang, D.; and Miao, C. 2020. EEG-Based Emotion Recognition Using Regularized Graph Neural Networks. *IEEE Transactions on Affective Computing* 13(3):1290-1301. doi.org/10.1109/taffc.2020.2994159.