Title
Cognitive Debiasing via Disentangled Pre-training for Cross-Subject EEG Emotion Recognition

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors
Chen, Qi
Liu, Zhaiyi
Wang, Lei
et al.

Publication Date
2026

Cognitive Debiasing via Disentangled Pre-training for Cross-Subject EEG Emotion Recognition

Qi Chen[1]**, Zhaiyi Liu**[2]**, Lei Wang**[3]**, Zihe Wang**[1]**, Jian Zheng*** **(zhengjian@cmu.edu.cn)**[4] **&**
Bo Jin* **(jinbo@dlut.edu.cn)**[5]

[1]School of Computer Science and Technology, Dalian University of Technology
[2]Faculty of Medicine, Dalian University of Technology
[3]Thrust of Artificial Intelligence, Hong Kong University of Science and Technology (Guangzhou)
[4]Department of Neurosurgery, Shengjing Hospital of China Medical University
[5]School of Innovation and Entrepreneurship, Dalian University of Technology

Abstract

Capturing shared cognitive processes across individuals is crucial for cross-subject EEG emotion recognition. Existing studies overlook the subjective experiential component introduced by individual cognitive biases, causing redundant information to be entangled with the extracted emotion-related features. To this end, we propose a cross-subject EEG emotion recognition method named CDDP (Cognitive Debiasing via Disentangled Pre-training). Specifically, during pre-training, cognitive disentanglement and contrastive learning are leveraged to extract subject-invariant intrinsic features (essentially emotion-relevant features) and subject-specific bias features (subjective experiential knowledge) from EEG signals. Meanwhile, a variational autoencoder (VAE) is introduced to generate diverse bias features in the latent space, alleviating overfitting caused by limited source domain EEG data. During fine-tuning, an emotion classifier is jointly trained with the pre-trained subject-invariant intrinsic feature encoder to capture discriminative emotion representations. CDDP achieves accuracies of 88.62% and 75.38% on the SEED and SEED-IV datasets, respectively, demonstrating state-of-the-art performance.

Keywords: Cognitive bias; Disentanglement; Contrastive learning; Cross-subject; EEG emotion recognition.

Introduction

EEG-based emotion recognition is a widely studied topic in the brain-computer interface (BCI) field (Samal & Hashmi, 2024). It has shown strong potential in applications such as mental health monitoring (Rahman et al., 2025) and intelligent human-computer interaction (Erat et al., 2024). In practice, the key challenge is recognizing emotions in unseen subjects. This task is known as cross-subject EEG emotion recognition. Inter-individual differences in brain structure and cognitive processes lead to substantial EEG heterogeneity (Genon et al., 2022). Such heterogeneity limits the generalization of conventional emotion recognition models (Suykens & Vandewalle, 1999) in cross-subject scenarios.

Existing studies mainly adopt transfer learning methods (Pan & Yang, 2009) to mitigate distribution shift caused by inter-individual variability. Depending on whether target domain data are accessible during training, transfer learning is commonly categorized into domain adaptation (DA) and domain generalization (DG). DA methods (Qiu et al., 2024) jointly leverage source domain data and a small amount of target domain data. They align features from different domains into a shared space. This alignment alleviates the mismatch between training and testing distributions. However, in practice, data from new users are often difficult to obtain. In some cases, such data are entirely unavailable. This limitation restricts the applicability of DA. In contrast, DG methods (Sun et al., 2025) do not rely on target domain data. They train models on data from multiple source domain subjects. The goal is to learn subject-invariant emotion features. Therefore, DG better matches deployment requirements in practical BCI applications. Despite their success, these methods often overlook the neural mechanisms underlying emotional information processing. This oversight limits the ability to capture features that are intrinsically emotion-relevant. Neuroscience studies indicate that, under the same stimulus conditions, neural activity can exhibit a certain degree of synchrony across different subjects (H. Liu et al., 2021). Several works (Hu et al., 2025; Shen et al., 2022) use this idea to construct positive and negative pairs for contrastive learning. They aim to learn subject-invariant emotion features. However, such methods ignore cognitive bias toward the same stimulus across individuals. For example, patients with depression tend to make negative interpretations, whereas patients with anxiety tend to overestimate threats (Moulds & McEvoy, 2025). Such cognitive bias can be regarded as a source of individual differences in emotion processing. If it is indiscriminately entangled into emotion representations, the model can learn redundant information coupled with individual bias. Moreover, limited source domain EEG data make it difficult for the training set to cover diverse bias patterns. This issue further increases the risk of overfitting. These factors together constrain the generalization ability of cross-subject models.

To address these issues, a cross-subject EEG emotion recognition method, termed CDDP (Cognitive Debiasing via Disentangled Pre-training), is proposed. This method examines inter-individual variability in EEG signals from a new perspective of cognitive debiasing. The core motivation is that, if individuals can psychologically separate subjective perceptions of reality from intrinsic thought representations, bias-induced interference with shared emotion representations can be suppressed. This separation facilitates the extraction of robust subject-invariant intrinsic features. As a result, per-

*Corresponding authors.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

formance on cross-subject EEG emotion recognition can be improved.

The main contributions of CDDP are summarized as follows:

- A feature disentanglement and reconstruction module is designed to disentangle subject-invariant intrinsic features from subject-specific bias features, thereby eliminating the interference of cognitive bias at the feature level.
- We introduce a VAE to generate diverse bias features in the latent space. These bias features are combined with subject-invariant intrinsic features to synthesize new EEG samples. This strategy alleviates overfitting caused by limited source domain samples.
- Extensive experiments are conducted on benchmark datasets, and the results demonstrate that the proposed CDDP effectively improves cross-subject EEG emotion recognition performance.

Related work

Cross-subject EEG emotion recognition

Inter-individual differences in brain cognitive processes have hindered the deployment of EEG-based affective brain-computer interfaces in real-world scenarios. Prior studies have mainly leveraged domain adaptation (Qiu et al., 2024) or meta-learning (Zhang, Yang, et al., 2024) to improve cross-subject performance. However, these approaches typically require target domain data that are difficult to obtain, which limits their practicality. In contrast, domain generalization does not require target domain data and is more consistent with real-world application needs. The work most closely related to ours is DMMR (Wang et al., 2024), which adopts a pre-training and fine-tuning paradigm to encourage the encoder to learn subject-invariant features. However, cognitive psychology research has shown that emotion perception is often shaped by prior experience and expectations. As a result, people tend to attend to cues that align with their beliefs, which leads to systematic cognitive biases. Such bias can lead to an overestimation or underestimation of emotional intensity (Genzer et al., 2025). These methods overlook this bias, leaving the learned features entangled with subjects' subjective experience and thus undermining generalization. To this end, our CDDP framework aims to eliminate the interference of inter-individual cognitive biases.

Feature disentanglement

Disentanglement-based domain generalization methods typically disentangle feature representations into domain-invariant and domain-specific features, and jointly optimize them using a reconstruction loss and regularization terms. The reconstruction loss preserves the original information, whereas the regularization terms promote effective disentanglement between domain-invariant and domain-specific features. Prior work has attempted to introduce disentanglement into cross-subject EEG emotion recognition. Zhao et al. (2021) disentangled EEG features into subject-specific private components and emotion components shared across subjects, and jointly trained multiple private classifiers. However, because cognitive bias can influence the private component, incorporating it into emotion classification decisions can reduce recognition accuracy. To this end, this work takes a cognitive debiasing perspective and removes redundant bias information through disentangled pre-training. Meanwhile, bias features are randomly generated to expand the diversity of cognitive patterns, thereby further improving generalization in cross-subject scenarios.

Methods

Overall architecture

The overall architecture of CDDP is illustrated in Figure 1. The model consists of two stages: pre-training and fine-tuning. The pre-training stage consists of four main components: Masked Channel Attention Module (MCA), Feature Disentanglement and Reconstruction Module (FDR), Dual Domain Classifier Module (DDC), and Bias Feature Random Generation Module (BiasVAE). It aims to maximally suppress subject-specific bias information and to pre-train a subject-invariant intrinsic feature encoder. To this end, CDDP jointly optimizes the domain classification loss, reconstruction loss, KL divergence loss, and contrastive loss to collaboratively constrain the learning process. During fine-tuning, the attention network and the subject-invariant intrinsic feature encoder are frozen, and an emotion classifier is trained to perform emotion recognition. The modules are described in detail below.

Masked channel attention module

The sliding-window technique is commonly used to segment a multi-channel EEG time series of length L into multiple segments. Given $X = (x_1, x_2, \dots, x_L)$, $x_l \in \mathbb{R}^C$ denotes the observation at timestamp l. The feature dimension satisfies $C = K * F$, where K denotes the number of channels and F denotes the number of frequency bands.

During pre-training, a fixed proportion of EEG channels is randomly selected and set to zero to reduce the impact of noise. Inspired by PPDA (Zhao et al., 2021), a feature weighting method based on the self-attention mechanism is applied to the masked input data x'_l. This method automatically assigns the weights $\alpha_l \in \mathbb{R}^C$ to the channels and frequency bands of EEG signals, thereby enhancing the ability to effectively capture essential information. The core idea involves projecting the input data x'_l onto a new feature space through a linear transformation. The projection results are then normalized using the softmax function and element-wise multiplied with the original features to obtain the weighted features $\widetilde{x} \in \mathbb{R}^{L \times C}$. The procedure can be formulated as:

$$\alpha_l = softmax\left(w_l x'_l + b_l\right), \tag{1}$$

$$\widetilde{x_l} = \alpha_l x'_l, \tag{2}$$

where $w_l \in \mathbb{R}^{C*C}$ and $b_l \in \mathbb{R}^C$ are the trainable weights and bias, respectively.

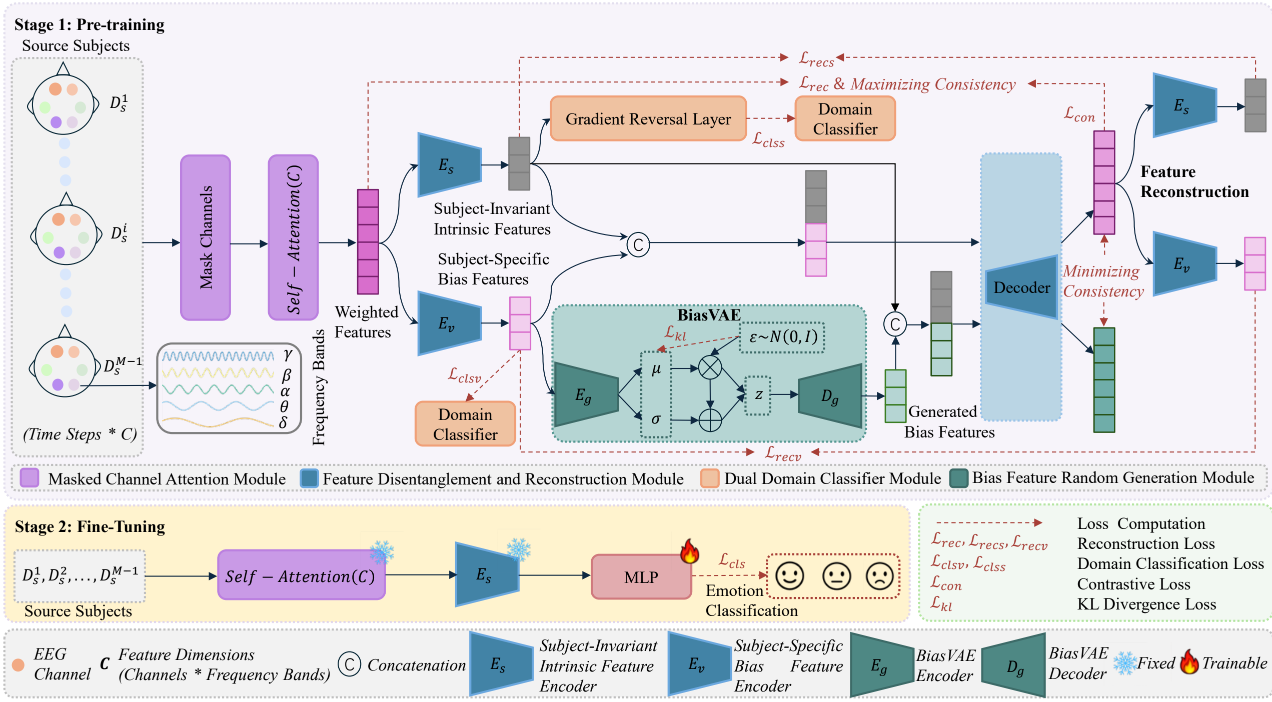


Figure 1: The workflow of the CDDP framework. CDDP follows a two-stage paradigm comprising pre-training and fine-tuning. The pre-training stage consists of four main components: Masked Channel Attention Module, Feature Disentanglement and Reconstruction Module, Dual Domain Classifier Module, and Bias Feature Random Generation Module. During fine-tuning, the attention network and the subject-invariant intrinsic feature encoder are frozen, and an emotion classifier is trained to perform emotion recognition.

Feature disentanglement and reconstruction module

Long Short-Term Memory (LSTM) networks are commonly used to model temporal dependencies in EEG signals (Wang et al., 2024; Zhao et al., 2021). Therefore, to achieve the disentanglement of EEG feature representations, lstm-based encoders are developed as a pair: subject-invariant intrinsic feature encoder E_s and subject-specific bias feature encoder E_v. E_s is employed to extract the subject-invariant intrinsic features x_s, while E_v is tasked with extracting the subject-specific bias features x_v. Subsequently, the extracted features x_s and x_v are concatenated and input into an lstm-based decoder to perform the reconstruction task. The global reconstruction process can be expressed as:

$$(x_s, x_v) = (E_s(\widetilde{x}), E_v(\widetilde{x})), \tag{3}$$

$$x' = Decoder(x_s \oplus x_v), \tag{4}$$

$$\mathcal{L}_{rec} = \mathrm{MSE}(\widetilde{x}, x'). \tag{5}$$

To further enhance the performance of the encoder and decoder in feature disentanglement and reconstruction tasks, the reconstructed feature x' is reintroduced into E_s and E_v to reconstruct the subject-invariant intrinsic feature x_s' and the subject-specific bias feature x_v', respectively. The local reconstruction (LREC) process can be expressed as:

$$\mathcal{L}_{recs} = \mathrm{MSE}(x_s, x_s'), \mathcal{L}_{recv} = \mathrm{MSE}(x_v, x_v'). \tag{6}$$

Therefore, the total reconstruction loss can be defined as:

$$\mathcal{L}_{recon} = \mathcal{L}_{rec} + \mathcal{L}_{recs} + \mathcal{L}_{recv}. \tag{7}$$

Dual domain classifier module

We design two independent domain classifiers, C_s and C_v, to constrain the learning of subject-invariant intrinsic features and subject-specific bias features, respectively, thereby enhancing the separability of the two feature types. Both classifiers are implemented using fully connected networks, but their parameters are not shared. In particular, a gradient reversal layer (GRL) is introduced before C_s. During backpropagation, the GRL multiplies the gradients by a negative weighting factor $-\eta$. This operation encourages the encoder E_s to extract features that cannot be distinguished by C_s. A unique identity label y^i is assigned to each subject to supervise the training of the two domain classifiers. The procedure can be represented as:

$$\begin{aligned}\mathcal{L}_{clsd} &= \mathcal{L}_{clss} + \mathcal{L}_{clsv} \\ &= -\eta\, y^i \log\big(C_s(x_s)\big) + y^i \log\big(C_v(x_v)\big), \\ &\quad i \in \{1, 2, \ldots, M-1\}.\end{aligned} \tag{8}$$

Bias feature random generation module

We introduce contrastive learning to better suppress subject-specific redundant information. However, the scarcity of source domain EEG samples limits the effectiveness of contrastive learning. To address this limitation, we propose BiasVAE to augment the diversity of inter-individual cognitive bias features. Structurally, this module is similar to a VAE (Kingma, 2013). It consists of an encoder E_g and a decoder D_g, both implemented using fully connected networks. The encoder maps subject-specific bias features into a latent space. Its posterior distribution is encouraged to match a Gaussian distribution with zero mean and an identity covariance matrix. The KL divergence loss is used for optimization, which can be formulated as:

$$(\mu, \log \sigma^2) = E_g(x_v), \tag{9}$$

$$\mathcal{L}_{gen} = KL\left(N\left(\mu, \sigma^2\right) \parallel N(0, \mathrm{I})\right). \tag{10}$$

A latent variable z is then randomly sampled from this distribution and passed to the decoder D_g to generate a new subject-specific bias feature x_g. This generated bias feature is fused with the subject-invariant intrinsic feature x_s and fed into the decoder to generate a new EEG sample x'_g. The procedure can be formulated as:

$$z = \mu + \sigma \odot \epsilon, \epsilon \sim \mathcal{N}(0, I), \tag{11}$$

$$x_g = D_g(z), {x_g}' = Decoder\left(x_s \oplus x_g\right). \tag{12}$$

Finally, a triplet loss is adopted to construct the contrastive constraint. It minimizes the feature distance between an input sample and its reconstruction, while maximizing the distance to a generated sample that carries different bias features. The triplet loss function is defined as:

$$\mathcal{L}_{con} = \max\left(d(\widetilde{x}, x') - d(\widetilde{x}, {x_g}') + margin, 0\right). \tag{13}$$

Loss computation

During pre-training, we optimize CDDP by minimizing the following loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \lambda_1 \mathcal{L}_{clsd} + \lambda_2 \mathcal{L}_{gen} + \lambda_3 \mathcal{L}_{con}, \tag{14}$$

where λ_1, λ_2 and λ_3 are weight coefficients that are adaptively adjusted via Dynamic Weight Average (DWA) (S. Liu et al., 2019), which assigns smaller weights to loss terms that change more rapidly.

Fine-tuning

During fine-tuning, we adopt labeled EEG data from source subjects for supervised learning. The CDDP combines a pre-trained attention network, a pre-trained subject-invariant intrinsic feature encoder, and an emotion classifier consisting of a fully connected network for the emotion recognition task. It is worth noting that the channels are not masked during the fine-tuning process. The emotion classification loss function is defined as:

$$\mathcal{L}_{cls} = y^j \log\left(CLS\left(E_s\left(\widetilde{x}_{un}\right)\right)\right), j \in \{1, 2, \ldots, N\}, \tag{15}$$

where y^j denotes the ground truth emotion label. $\widetilde{x}_{un}$ denotes the attention-weighted features. CLS denotes the emotion classifier.

Algorithms 1 summarize the pseudo-code for the training procedure.

Algorithm 1 The training procedure of the CDDP

Input: Source subjects EEG data $D_s = \left\{x_s^i, y_s^i\right\}_{i=1}^{m-1}$, epoch E_p, E_f, iteration N_p, N_f, channel mask ratio r.
Output: The trained model of CDDP.

The pre-training stage:
Initialize the model parameters w.
for $epoch = 1 : E_p$ **do**
  for $iteration = 1 : N_p$ **do**
    $\tilde{x} \leftarrow MCA(Mask(x; r))$.
    $x_s \leftarrow E_s(\tilde{x}),\ \ x_v \leftarrow E_v(\tilde{x})$. ▷ disentanglement
    $x' \leftarrow Decoder(x_s \oplus x_v)$. ▷ global recon
    $x'_s \leftarrow E_s(x'),\ \ x'_v \leftarrow E_v(x')$. ▷ local recon
    $x'_g \leftarrow Decoder(x_s \oplus BiasVAE(x_v))$. ▷ generation
    $\mathcal{L}_{con} \leftarrow \mathcal{L}_{triplet}(\tilde{x}, x', x'_g)$. ▷ contrastive learning
    $\mathcal{L}_{\text{total}} \leftarrow$ use Eq. (14).
    $w \leftarrow \text{Update}(w; \nabla_w \mathcal{L}_{\text{total}})$.
  end for
end for
return The pre-trained model of CDDP.
The fine-tuning stage:
Initialize the emotion classifier parameters θ_{cls}.
for $epoch = 1 : E_f$ **do**
  for $iteration = 1 : N_f$ **do**
    $\mathcal{L}_{cls} \leftarrow$ use Eq. (15).
    $\theta_{cls} \leftarrow \theta_{cls} - \eta \nabla_{\theta_{cls}} \mathcal{L}_{cls}$.
  end for
end for
return The trained model of CDDP.

Experiments

Benchmark datasets

We evaluate CDDP on two benchmark EEG emotion recognition datasets: SEED (Zheng & Lu, 2015) and SEED-IV (Zheng et al., 2018). The SEED dataset contains EEG data from 15 subjects and covers three emotion categories: neutral, negative, and positive. The SEED-IV dataset also includes EEG data from 15 subjects and extends the emotion categories to four classes: neutral, sad, fear, and happy. EEG data for each subject are collected across three independent sessions using 62 channels. For fair comparison, only data from the first session of all subjects are used in this study. After downsampling and filtering, the raw EEG signals are decomposed into five frequency bands: $\alpha, \beta, \delta, \theta$, and γ. A 1-second non-overlapping sliding window is then applied to extract differential entropy (DE) features from the preprocessed signals, forming the final dataset.

Baselines and evaluation criteria

We compare our model by comprehensively evaluating nine baseline models, including the traditional machine learning method: SVM (Suykens & Vandewalle, 1999); the contrastive learning methods: GMSS (Li et al., 2022) and CL-CS (Hu et al., 2025); the domain adaptation methods: PPDA (Zhao et al.,

2021) and IDDA (An et al., 2024); the domain generalization methods: DResNet (Ma et al., 2019), MoGE (X.-H. Liu et al., 2024), DMMR (Wang et al., 2024), and AC-CDCN (Sun et al., 2025).

To ensure the reliability of the evaluation, leave-one-subject-out cross-validation is used as the evaluation strategy. In each experiment, data from a single subject are used as the test set, while data from the remaining subjects are used to train the model, ensuring that each subject is tested independently. The results from all experiments are then combined, and the overall average accuracy (Avg.) and standard deviation (Std.) are calculated to evaluate the model's stability and ability to generalize.

Implementation details

The default hyperparameter settings are summarized as follows: The channel mask ratio is set to 0.2. The hidden layer dimensions of the subject-invariant intrinsic feature encoder, subject-specific bias feature encoder, and BiasVAE are all set to 64. The hidden dimension of the fully connected layer in the emotion recognition module is set to 128. The time step L is set to 30 for the SEED dataset and 10 for the SEED-IV datasets, while the batch size remains 256 across all datasets. The Adam optimizer is used in conjunction with the CosineAnnealingWarmRestarts strategy for optimization. All experiments are conducted using PyTorch on a single NVIDIA RTX4090 24GB GPU.

Main results

Comparison with baseline methods Table 1 shows the results of the CDDP evaluation on the two benchmark datasets. The results demonstrate that CDDP outperforms current baseline methods according to commonly used evaluation metrics and achieves state-of-the-art performance. Specifically, for the SEED dataset, the emotion recognition rate of CDDP increases to 88.62%, representing an improvement of approximately 1.92% compared to PPDA (Zhao et al., 2021), a domain adaptation method that also employs the principle of disentanglement. This result indicates that CDDP is more effective in extracting subject-invariant intrinsic features through the principle of disentangling. For the SEED-IV dataset, CDDP improves accuracy by 2.68% compared to the domain generalization method DMMR (Wang et al., 2024), which also adopts the pre-training and fine-tuning paradigm.These results support our claim that suppressing interference from diverse individual cognitive bias features helps improve model performance in cross-subject scenarios.

Stability analysis Figure 2 presents the detailed results under the leave-one-subject-out cross-validation setting, where each subject is used in turn as the target subject. Accuracy varies only slightly for most target subjects. In addition, Table 1 illustrates that CDDP exhibits the lowest or comparable standard deviation among all compared methods, demonstrating its superior stability and robustness in cross-subject experiments.

Table 1: Performance comparison of our proposed CDDP with baselines. The best results are in bold.

Methods	Year	SEED		SEED-IV	
		Avg.	Std.	Avg.	Std.
SVM	1991	56.73	16.29	37.99	12.52
DResNet	2019	85.30	7.97	72.10	12.70
PPDA	2021	86.70	7.10	71.20	6.20
GMSS	2022	86.52	6.22	73.48	7.41
IDDA	2024	85.75	8.11	72.36	9.43
MoGE	2024	88.00	4.50	74.30	**6.10**
DMMR	2024	88.27	5.62	72.70	8.01
CL-CS	2025	88.30	8.90	-	-
AC-CDCN	2025	87.14	5.60	71.77	12.92
CDDP	2026	**88.62**	**3.66**	**75.38**	6.58

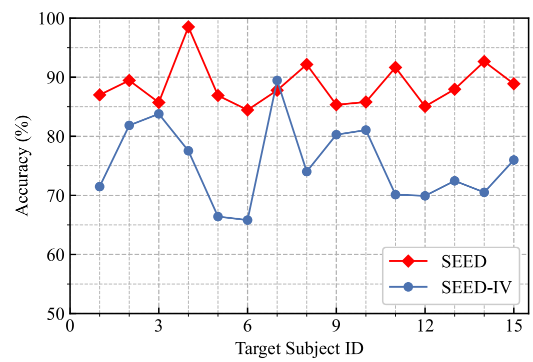


Figure 2: Stability analysis on the SEED and SEED-IV datasets.

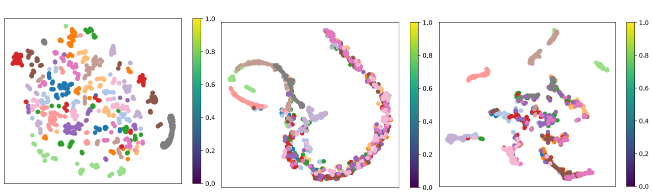
Ablation studies To validate the contribution of each module, ablation studies are conducted on two datasets. The specific settings are as follows: "w/o Mask Channels" denotes removing the channel masking operation in the MCA module; "w/o LREC" denotes removing local feature reconstruction in the reconstruction task and retaining only global feature reconstruction; "w/o C_v" denotes removing the domain classifier that constrains subject-specific bias feature learning and retaining only a single domain classifier to strengthen subject-invariant intrinsic feature learning; "w/o BiasVAE" denotes removing the bias feature random generation module and simultaneously discarding the contrastive learning constraint.

Table 2 illustrates the results of the ablation studies, where removing any modular component resulted in performance degradation (lower accuracy and higher standard deviation). Specifically, removing the channel masking operation weakens the model's ability to suppress noise and irrelevant channel information, making it difficult to capture robust EEG features. In contrast, introducing local feature reconstruction provides finer-grained information during disentanglement and reconstruction, thereby strengthening the joint optimization of the encoder and decoder. Furthermore, removing C_v makes subject-specific bias features difficult to be sufficiently separated and suppressed, which weakens the model's feature disentanglement capability. Finally, BiasVAE helps expand the diversity of bias patterns, which alleviates distribution shifts caused by inter-individual cognitive differences and improves the model's robustness and generalization. In summary, these components jointly improve the model's performance in recognizing emotions from noisy EEG signals in cross-subject scenarios.

Table 2: Ablation studies in CDDP. The best results are in bold.

Methods	SEED		SEED-IV	
	Avg.	Std.	Avg.	Std.
w/o Mask Channels	87.65	5.67	74.25	8.86
w/o LREC	84.63	5.50	73.12	9.29
w/o C_v	85.03	4.98	73.56	8.23
w/o BiasVAE	83.88	6.12	71.43	9.87
CDDP	**88.62**	**3.66**	**75.38**	**6.58**

The effect of feature disentanglement To evaluate the effectiveness of CDDP in EEG feature disentanglement, 80 samples are randomly drawn for each subject from the SEED dataset, and the representations are visualized using t-SNE (Van der Maaten & Hinton, 2008). Each subject is represented by a unique color. As shown in Figure 3(a), the raw features exhibit a disordered distribution in the embedding space. After pre-training, Figure 3(b) shows substantial overlap of subject-invariant intrinsic features across individuals, indicating that the model extracts shared cognitive representations related with emotion. In contrast, the subject-specific bias features in Figure 3(c) are partitioned into multiple clusters and show significant subject separability: features from the same subject are more clustered in the embedding space, whereas features from different subjects are more clearly separated. This observation suggests that personalized cognitive information is effectively aggregated into the bias component. Overall, these results validate the proposed method in separating individualized cognition from shared cognitive representations.



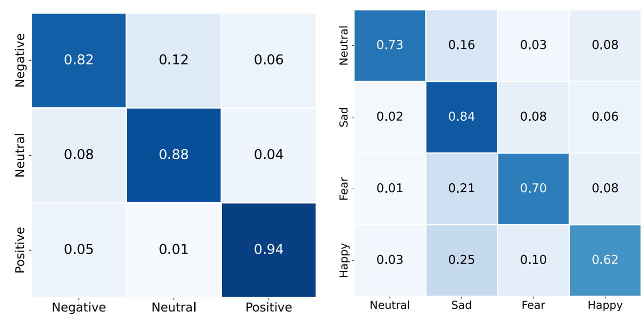
(a) Original features

(b) Pre-trained invariant features

(c) Pre-trained bias features

Figure 3: The effect of feature disentanglement on the SEED dataset.

Confusion matrices To thoroughly analyze the performance of CDDP in recognizing different emotion categories, Figure 4 presents its recognition results on two datasets, visualized using a confusion matrix. The results in Figure 4(a) indicate that for the SEED dataset, positive emotions are the easiest to recognize, whereas negative and neutral emotions may exhibit certain similarities, leading the model to often misidentify both emotions. This observation is consistent with the results of MoGE (X.-H. Liu et al., 2024). Conversely, the results shown in Figure 4(b) and Figure 4(c) indicate that, in the SEED-IV dataset, sad emotions are most easily distinguished, while happy emotions represent the most challenging category to differentiate. One possible reason is that the dataset contains relatively few samples in the happy class, leading to insufficient feature learning for this category and making it more likely to be misclassified as other emotion classes.



(a) SEED

(b) SEED-IV

Figure 4: Confusion matrices on the SEED and SEED-IV datasets.

Conclusion and future work

We propose CDDP (Cognitive Debiasing via Disentangled Pre-training), a cross-subject EEG emotion recognition method. Motivated by cognitive psychology and data-driven learning strategies, CDDP aims to capture emotion-related cognitive processes shared across individuals. Extensive ablation studies demonstrate the effectiveness of the CDDP components. Moreover, the experimental results further indicate that separating individualized cognition from shared cognitive representations plays an important role in improving cross-subject emotion recognition performance. Experiments on the SEED and SEED-IV datasets show strong performance on key metrics such as accuracy and standard deviation, which validates the advantage of the proposed method for emotion recognition. In future work, we will further extend CDDP to support additional modalities and explore new paradigms that integrate multimodal physiological signals with LLM (Large Language Model). This extension is expected to enable the model to output not only classification results but also interpretable evidence for its decisions, thereby building an agent that can more deeply understand users' cognitive processes.

Acknowledgments

This research is supported by the National Key R&D Program of China (2023YFC3604205), National Natural Science Foundation of China (No. 62172074), Program of Introducing Talents of Discipline to Universities (Plan 111) (No. B20070).

References

An, Y., Hu, S., Liu, S., Wang, Z., Wang, X., & Ma, X. (2024). Cross-subject eeg emotion recognition based on interconnected dynamic domain adaptation. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 12981–12985. https://doi.org/10.1109/ICASSP48485.2024.10446957

Erat, K., Şahin, E. B., Doğan, F., Merdanoğlu, N., Akcakaya, A., & Durdu, P. O. (2024). Emotion recognition with eeg-based brain-computer interfaces: A systematic literature review. *Multimedia tools and applications*, *83*(33), 79647–79694. https://doi.org/10.1007/s11042-024-18259-z

Genon, S., Eickhoff, S. B., & Kharabian, S. (2022). Linking interindividual variability in brain structure to behaviour. *Nature Reviews Neuroscience*, *23*(5), 307–318. https://doi.org/10.1038/s41583-022-00584-7

Genzer, S., Rubin, M., Sened, H., Rafaeli, E., Ochsner, K. N., Cohen, N., & Perry, A. (2025). Directional bias in interpersonal emotion perception. *Nature Communications*. https://doi.org/10.1038/s41467-025-66879-2

Hu, M., Xu, D., He, K., Zhao, K., & Zhang, H. (2025). Cross-subject emotion recognition with contrastive learning based on eeg signal correlations. *Biomedical Signal Processing and Control*, *104*, 107511. https://doi.org/10.1016/j.bspc.2025.107511

Kingma, D. P. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*. https://indico.math.cnrs.fr/event/11377/attachments/4589/6915/18012024_Kingma-and-Welling-2022%20Auto-Encoding%20Variational%20Bayes.pdf

Li, Y., Chen, J., Li, F., Fu, B., Wu, H., Ji, Y., Zhou, Y., Niu, Y., Shi, G., & Zheng, W. (2022). Gmss: Graph-based multi-task self-supervised learning for eeg emotion recognition. *IEEE Transactions on Affective Computing*, *14*(3), 2512–2525. https://doi.org/10.1109/TAFFC.2022.3170428

Liu, H., Zhao, C., Wang, F., & Zhang, D. (2021). Inter-brain amplitude correlation differentiates cooperation from competition in a motion-sensing sports game. *Social cognitive and affective neuroscience*, *16*(6), 552–564. https://doi.org/10.1093/scan/nsab031

Liu, S., Johns, E., & Davison, A. J. (2019). End-to-end multi-task learning with attention. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1871–1880. https://openaccess.thecvf.com/content_CVPR_2019/html/Liu_End-To-End_Multi-Task_Learning_With_Attention_CVPR_2019_paper.html

Liu, X.-H., Jiang, W.-B., Zheng, W.-L., & Lu, B.-L. (2024). Moge: Mixture of graph experts for cross-subject emotion recognition via decomposing eeg. *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3515–3520. https://doi.org/10.1109/BIBM62325.2024.10822354

Ma, B.-Q., Li, H., Zheng, W.-L., & Lu, B.-L. (2019). Reducing the subject variability of eeg signals with adversarial domain generalization. *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26*, 30–42. https://doi.org/10.1007/978-3-030-36708-4_3

Moulds, M. L., & McEvoy, P. M. (2025). Repetitive negative thinking as a transdiagnostic cognitive process. *Nature Reviews Psychology*, *4*(2), 127–141. https://doi.org/10.1038/s44159-024-00399-6

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Qiu, L., Ying, Z., Feng, W., & Pan, J. (2024). Cross-subject eeg emotion recognition based on multitask adversarial domain adaption. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*. https://escholarship.org/uc/item/4jt9f3k9

Rahman, A. U., Ali, S., Wason, R., Aggarwal, S., Abohashrh, M., Daradkeh, Y. I., & Ullah, I. (2025). Emotion-based mental state classification using eeg for brain-computer interface applications. *Computational Intelligence*, *41*(4), e70112. https://doi.org/10.1111/coin.70112

Samal, P., & Hashmi, M. F. (2024). Role of machine learning and deep learning techniques in eeg-based bci emotion recognition system: A review. *Artificial Intelligence Review*, *57*(3), 50. https://doi.org/10.1007/s10462-023-10690-2

Shen, X., Liu, X., Hu, X., Zhang, D., & Song, S. (2022). Contrastive learning of subject-invariant eeg representations for cross-subject emotion recognition. *IEEE Transactions on Affective Computing*, *14*(3), 2496–2511. https://doi.org/10.1109/TAFFC.2022.3164516

Sun, Y., Yang, L., He, H., & Du, J. (2025). Ac-cdcn: A cross-subject eeg emotion recognition model with anti-collapse domain generalization. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*. https://escholarship.org/uc/item/2772j8s6

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, *9*(3), 293–300. https://doi.org/10.1023/A:1018628609742

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11). https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

Wang, Y., Zhang, B., & Tang, Y. (2024). Dmmr: Cross-subject domain generalization for eeg-based emotion recognition via denoising mixed mutual reconstruction. *Proceedings of the AAAI conference on artificial intelligence*, *38*(1), 628–636. https://doi.org/10.1609/aaai.v38i1.27819

Zhang, Q., Yang, L., et al. (2024). Cross-subject emotion classification based on dual-attention mechanism and metatransfer learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*. https://escholarship.org/uc/item/5kb2r6rj

Zhao, L.-M., Yan, X., & Lu, B.-L. (2021). Plug-and-play domain adaptation for cross-subject eeg-based emotion recognition. *Proceedings of the AAAI conference on artificial intelligence*, *35*(1), 863–870. https://doi.org/10.1609/aaai.v35i1.16169

Zheng, W.-L., Liu, W., Lu, Y., Lu, B.-L., & Cichocki, A. (2018). Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE transactions on cybernetics*, *49*(3), 1110–1122. https://doi.org/10.1109/TCYB.2018.2797176

Zheng, W.-L., & 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. https://doi.org/10.1109/TAMD.2015.2431497