EmoVAE: Bridging Affective Neural Dynamics for Cross-Dataset Emotion Recognition

Zhuoyi Huang[1†], Rongtao Chen[1†], Xinhan Lin[1], Jiajun Chen[1] & Jiahui Pan (panjiahui@m.scnu.edu.cn)[1]
[1]School of Artificial Intelligence, South China Normal University, Foshan, China

Abstract

Emotion recognition from neural signals constitutes a cornerstone of cognitive science, providing objective access to affective dynamics. However, reliable decoding across datasets is hindered by domain shifts in recording protocols and neurophysiological variability. We propose EmoVAE, a domain adaptation framework that bridges emotional neural dynamics via variational latent distribution alignment. EmoVAE uses a multi-scale spatiotemporal aggregation encoder to capture spatiotemporal patterns of affect-related activity, and an EEG-Latent module to map these representations into a continuous probabilistic space. Within this latent space, we align class-conditional distributions using conditional maximum mean discrepancy, yielding stable, geometry-aware emotion clusters across domains. Experiments on SEED and SEED-VII show state-of-the-art performance for both cross-subject and cross-dataset protocols, demonstrating robust generalization under domain shift. These findings suggest that continuous latent alignment is an effective strategy for building reliable EEG-based emotion recognition systems for cognitive and clinical applications.

Keywords: Electroencephalogram (EEG), cognitive science, cross-dataset, domain generalization, latent space alignment.

Introduction

Emotion recognition serves as an important role for exploring efficient cognitive processes, driving the development of fields such as human-computer interaction and mental health monitoring (R. Chen et al., 2025; Cowie et al., 2001). However, traditional emotion measurement methods, such as subjective self-reporting and behavioral measurement, are inherently subjective and prone to deception (Khare et al., 2024). In contrast, electroencephalography (EEG) directly measures neural activities related to emotional states, providing a more reliable basis for emotional states.

In recent years, emotion recognition based on EEG has made progress both in individual-dependent and individual-independent settings (Bhosale et al., 2022; Nath et al., 2020). However, the generalization ability of the model in cross-individual scenarios still faces challenges, and its performance is usually significantly lower than that in individual-dependent settings. More seriously, cross-dataset recognition reveals a deeper problem. It not only involves cross-individual differences but also introduces systematic distribution shifts caused by multiple heterogeneities such as acquisition devices, experimental paradigms, and annotation standards (Joshi et al., 2022). This complex domain difference severely hinders the learning of robust representations. Therefore, improving cross-dataset generalization is a crucial step toward establishing EEG-based emotion recognition as a reliable cognitive tool, rather than a dataset-specific engineering solution.

To address the severe domain shift caused by cross-dataset scenarios, it is necessary to extract features that reflect the affective process while being insensitive to domain variation. Since emotion-related neural responses manifest as temporal dynamics and inter-channel interactions, prior studies typically extract spatiotemporal features (Klepl et al., 2024; F. Zheng et al., 2023). For example, LGGNet hierarchically extracts spatiotemporal features of the electroencephalogram signals through multi-scale time convolution layers and local-global graph neural networks (Ding et al., 2024). STFCGAT independently calculates differential entropy for each electrode to extract temporal features, then uses these features as nodes and performs spatial aggregation through an attention mechanism based on a predefined functional connection graph, thereby forming spatiotemporal features (Z. Li et al., 2023). Although these successes have been achieved, current methods typically separate the spatiotemporal dimensions in the feature extraction process and cannot capture the closely coupled spatiotemporal interaction patterns that are crucial for emotion decoding, resulting in the learned features being prone to be biased by the specific characteristics of the datasets.

Under cross-dataset protocols, shifts in acquisition and labeling can markedly widen the distribution gap between source and target. Domain adaptation methods tackle this issue by jointly learning emotional discriminative representations and reducing cross-domain discrepancy, encouraging features to transfer across datasets (Pan et al., 2025). For instance, MSGDA uses a graph domain adaptation network to model individual information and public information from multiple subjects to capture dynamic functional connectivity and regional brain states (Wang et al., 2024). PR-PL learns category prototypes from labeled source data and uses the interaction between prototypes and samples to ensure consistency in the category conditional distribution (Zhou et al., 2023). However, these methods align in a discrete feature space, which is difficult to carry continuous semantics, resulting in blurred boundaries between different emotion categories.

To address these challenges, we propose a novel domain adaptation framework. It continuously aligns conditional probability distributions for emotion recognition. Our con-

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

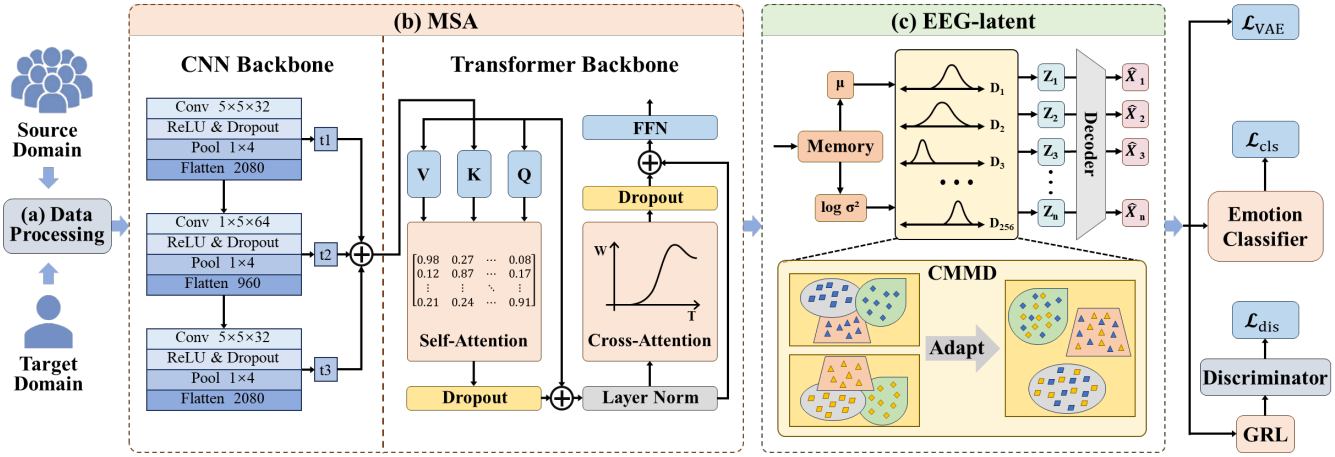


Figure 1: Proposed domain adaptation framework with two key components: (a) Data preprocessing: filtering, segmentation, and differential entropy feature extraction; (b) MSA module: multiscale spatiotemporal encoding with CNNs and channel attention; (c) EEG-latent module: variational latent alignment via conditional maximum mean discrepancy.

tributions can be summarized as follows:

- **Bridging Spatiotemporal Neural Dynamics:** We propose a multiscale spatiotemporal aggregation (MSA) architecture that captures temporal dynamics and inter-channel dependencies of affective activity. By treating channels as tokens in self-attention, it learns implicit functional connectivity without hand-crafted spatial priors, yielding robust emotion-discriminative representations.
- **Continuous Latent Alignment of Emotional Distributions:** We introduce an EEG-Latent module that maps representations into a variational latent space and aligns class-conditional distributions across domains via conditional maximum mean discrepancy (CMMD). This yields more compact and separable emotion clusters while preserving affective semantics.
- **Generalizable Emotion Recognition:** We benchmark cross-dataset transfer on SEED and SEED-VII, achieving substantial performance gains over state-of-the-art methods (55.00% for SEED→SEED-VII and 51.70% for the reverse). The results support latent alignment as a principled way to mitigate domain shift for cognitively grounded emotion recognition.

Methodology

We consider an unsupervised domain adaptation setting for EEG emotion recognition. The labeled source domain is $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, and the target domain is $\mathcal{D}_t = \{x_j^t\}_{j=1}^{N_t}$, where x denotes an EEG sample and y is its emotion label. During training, target labels are not accessible, so y_j^t is unknown. The complete architecture of our proposed model is shown in Figure 1. To enable our proposed method to be reproducible, we provide the full code at the link below: https://github.com/tableMHT/EmoVAE.git.

Data Preprocessing

To improve cross-domain consistency and suppress acquisition noise, raw EEG signals are first downsampled to 200 Hz for efficiency, followed by a 1–45 Hz band-pass filter to remove low-frequency drift and high-frequency artifacts. The continuous streams are then partitioned into non-overlapping 1-second windows.

For each window, we compute Differential Entropy (DE) features over five canonical bands: Delta (1–4 Hz), Theta (4–8 Hz), Alpha (8–14 Hz), Beta (14–30 Hz), and Gamma (30–45 Hz). Assuming the band-limited signal within a window approximately follows a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, its DE is defined as

$$\mathrm{DE} = -\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}\right) dx, \quad (1)$$

where σ^2 is the variance of the signal segment. Since trial durations differ across datasets, we enforce a fixed temporal length T_0 for all trials. Sequences longer than T_0 are truncated, while shorter ones are zero-padded. Accordingly, each EEG channel is represented by concatenating the 5-band DE features across T_0 time steps, yielding a dimension-consistent node feature that preserves the temporal structure.

To avoid target-domain leakage while alleviating inter-subject scale variations, we normalize features using source-only statistics. Let μ_j and σ_j be the mean and standard deviation of the j-th feature computed from $\{X_s\}$; then both source and target samples are standardized as

$$\hat{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j + \epsilon}, \quad \forall x_i \in \{X_s, X_t\}, \quad (2)$$

where ϵ is a small constant. The resulting $\{x_i^s\}_{i=1}^{N_s}$ and $\{x_i^t\}_{i=1}^{N_t}$ are used as the model inputs.

Multiscale Spatiotemporal Aggregation Module

MSA aims to achieve robust spatiotemporal representation learning. It uses a CNN backbone to encode local dynam-

ics and a channel-wise Transformer encoder to capture global inter-channel dependencies, effectively forming an implicit functional connectivity graph without hand-crafted spatial priors.

Hierarchical Convolution Encoder For the c-th EEG channel, the input map is $\hat{\mathrm{X}}_c \in \mathbb{R}^{B_f \times L}$, where B_f denotes the number of frequency bands and L denotes the aligned temporal length.

We employ a three-stage 1D CNN to capture patterns with progressively enlarged receptive fields. Specifically, each stage consists of a convolution operator followed by normalization, non-linearity and pooling, and its output is finally flattened into a vector representation:

$$\mathrm{u}_c^{(k)} = \mathrm{Flat}\Big(\mathrm{Pool}\big(\sigma(\mathrm{Norm}(\mathrm{Conv}_k(\mathrm{u}_c^{(k-1)}))\big)\Big), \quad k = 1, 2, 3, \tag{3}$$

$$\mathrm{f}_c = \mathrm{Concat}\big(\mathrm{u}_c^{(1)}, \mathrm{u}_c^{(2)}, \mathrm{u}_c^{(3)}\big). \tag{4}$$

where $\mathrm{u}_c^{(0)} = \hat{\mathrm{X}}_c$ is the input feature map of channel c; $\mathrm{Conv}_k(\cdot)$ denotes the convolution layer in the k-th stage, $\mathrm{Norm}(\cdot)$ is a normalization operator, $\sigma(\cdot)$ is a point-wise nonlinearity, and $\mathrm{Pool}(\cdot)$ performs temporal downsampling to improve robustness. $\mathrm{Flat}(\cdot)$ reshapes the stage output into a vector, and $\mathrm{Concat}(\cdot)$ concatenates multi-stage representations to form the final channel-wise feature f_c.

Spatial Feature Extraction In EEG, spatial information is primarily reflected by cross-channel dependencies within the same temporal window. Common attention design tokenizes time steps and performs attention along the temporal axis, which may under-emphasize the explicit channel-to-channel interactions.

In contrast, we treat each EEG channel as one token, whose embedding summarizes all time-frequency observations of that channel in the window. Such design enables attention to directly operate on C channel tokens, producing spatially-aware channel representations.

We compute self-attention across the C channel tokens. The attention weight is computed as:

$$\mathrm{W} = \mathrm{Softmax}\left(\frac{\mathrm{QK}^\top}{\sqrt{d_k}}\right) \in \mathbb{R}^{C \times C}, \tag{5}$$

where Q, K are linear projections of the channel-token embeddings. W is a row-stochastic affinity matrix whose (i, j)-th element measures how much channel i attends to channel j. The spatially-aware channel representations are then produced via weighted aggregation:

$$\mathrm{M} = \mathrm{WV} \in \mathbb{R}^{C \times d_k}, \tag{6}$$

where V represents the feature matrix obtained through a learnable linear mapping. Each row of V corresponds to the aggregated representation of an EEG channel. M aggregates value vectors from all channels into each channel, yielding M that encodes global inter-channel interactions, which is subsequently used for alignment and emotion classification.

Temporal Feature Extraction Emotion-evoked EEG patterns are typically transient and sparse: the most discriminative evidence often emerges shortly after the stimulus onset, while the pre-stimulus segment provides a compact baseline that helps characterize the transition into an affective state. Therefore, instead of uniformly pooling the entire trial, we explicitly focus on a short pre-stimulus interval together with the complete post-stimulus period, and use learnable queries to extract a small set of salient temporal summaries.

Given the stimulus onset index t_0, we keep a window Ω that covers a short pre-onset interval and all post-onset time steps:

$$\Omega = \{t_0 - \tau_{\mathrm{pre}}, \ldots, T\}, \tag{7}$$

$$T' = |\Omega|, \tag{8}$$

where τ_{pre} controls the length of the pre-stimulus context, and Ω specifies the selected time indices.

We introduce learnable query vectors Q_l to probe the temporal memory and extract multiple complementary summaries. Cross-attention is performed as:

$$\mathrm{Z} = \mathrm{Softmax}\left(\frac{\mathrm{Q}_l\mathrm{K}^\top}{\sqrt{d_k}}\right) \in \mathbb{R}^{N_q \times T'}, \tag{9}$$

$$\mathrm{Z}_{\mathrm{agg}} = \mathrm{AV} \in \mathbb{R}^{N_q \times d_k}, \tag{10}$$

where N_q represents the number of query slots used to extract multiple salient temporal summaries from the memory sequence. Z is a query-to-time affinity matrix whose element $\mathrm{Z}_{q,t}$ quantifies how much the q-th query attends to the t-th time step, thus $\mathrm{Z}_{\mathrm{agg}}$ becomes a set of N_q temporal summaries, each being a weighted aggregation over the selected timeline. By using multiple queries, the model can capture sparse emotion-related bursts at different moments while remaining robust to subject-specific temporal variability.

EEG Latent Distribution Alignment Module

Direct alignment on deterministic features may lead to irregular clusters and unstable decision boundaries, especially under cross-dataset settings. Therefore, we introduce an EEG-latent module that (i) maps aggregated summaries to a variational latent space with KL regularization, (ii) performs class-conditional distribution alignment in the latent space via conditional maximum mean discrepancy, and (iii) samples latent codes for decoding and reconstruction, ensuring information preservation.

Variational Mapping For each aggregated token $\mathrm{z} \in \mathrm{Z}_{\mathrm{agg}}$, we infer a gaussian posterior μ and $\log \sigma^2$. Then we sample latent tokens as $\mathrm{z}' = \mu + \sigma \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, \mathrm{I})$. $\mathcal{N}(0, \mathrm{I})$ is the isotropic gaussian prior, which serves as a distributional anchor to regularize the lat ent space to obtain a smooth and well-shaped latent geometry.

The KL divergence to the standard normal prior is calculated as:

$$\mathcal{L}_{\mathrm{KL}} = \mathbb{E}\big[\mathrm{KL}(q(\mathrm{z}'|\mathrm{Z}_{\mathrm{agg}}) \,\|\, \mathcal{N}(0, \mathrm{I}))\big], \tag{11}$$

where $q(\mathrm{z}' \mid \mathrm{Zagg})$ denotes the variational posterior inferred from the aggregated tokens Zagg, and z' is a latent token

sampled via the reparameterization trick. Minimizing $\mathcal{L}_{\mathrm{KL}}$ encourages the inferred posterior to stay close to the prior, yielding a smoother and more compact latent geometry and thereby improving training stability and cross-domain generalization.

Latent Conditional Alignment After obtaining the potential distribution, we explicitly reduced the conditional probability discrepancy by pulling the potential features of the source and target data of the same category closer together. This step is performed in the latent space to benefit from the KL-shaped geometry and to avoid aligning noisy high-dimensional features.

Let z_i^s and z_j^t denote latent codes of source and target samples, and $\hat{\mathrm{y}}_{\mathrm{j}}^{\mathrm{t}}$ be the pseudo-label of target sample. To align conditional distributions, we compute MMD within each emotion class and sum over classes:

$$\mathcal{L}_{\mathrm{CMMD}} = \sum_{c=1}^{C} \omega_c \, \mathrm{MMD}^2\left(\mathrm{Z}_s^{(c)}, \mathrm{Z}_t^{(c)}\right), \tag{12}$$

$$\begin{aligned}\mathrm{MMD}^2\left(\mathrm{Z}_s^{(c)}, \mathrm{Z}_t^{(c)}\right) =& \frac{1}{|\mathrm{Z}_s^{(c)}|^2} \sum_{\mathrm{z}_i^s, \mathrm{z}_{i'}^s \in \mathrm{Z}_s^{(c)}} k(\mathrm{z}_i^s, \mathrm{z}_{i'}^s) \\ &+ \frac{1}{|\mathrm{Z}_t^{(c)}|^2} \sum_{\mathrm{z}_j^t, \mathrm{z}_{j'}^t \in \mathrm{Z}_t^{(c)}} k(\mathrm{z}_j^t, \mathrm{z}_{j'}^t) \\ &- \frac{2}{|\mathrm{Z}_s^{(c)}||\mathrm{Z}_t^{(c)}|} \sum_{\mathrm{z}_i^s \in \mathrm{Z}_s^{(c)}} \sum_{\mathrm{z}_j^t \in \mathrm{Z}_t^{(c)}} k(\mathrm{z}_i^s, \mathrm{z}_j^t),\end{aligned} \tag{13}$$

where $\mathrm{Z}_s^{(c)} = \{\mathrm{z}_i^s \mid y_i^s = c\}$ and $\mathrm{Z}_t^{(c)} = \{\mathrm{z}_j^t \mid \hat{y}_j^t = c\}$ are class-conditioned latent sets in source and target domains, respectively. $k(\cdot, \cdot)$ is an RBF kernel, and ω_c balances classes. Minimizing $\mathcal{L}_{\mathrm{CMMD}}$ aligns source and target latents within each emotion class, leading to more compact and stable clusters.

Information reconstruction We then sample latent codes and decode them back to the feature space. A reconstruction objective is used to prevent information collapse and to retain discriminative cues for downstream classification and adaptation.

We stack the sampled tokens as $\mathrm{Z}' \in \mathbb{R}^{N_q \times d_z}$. We decode Z' by cross-attention to obtain channel-wise representation:

$$\mathrm{H} = \mathrm{Softmax}\left(\frac{(\mathrm{M}\mathrm{W}_Q)\,(\mathrm{Z}'\mathrm{W}_K)^\top}{\sqrt{d_k}}\right)(\mathrm{Z}'\mathrm{W}_V) \in \mathbb{R}^{C \times d_k}, \tag{14}$$

where $\mathrm{M} \in \mathbb{R}^{C \times d_k}$ is the channel memory, and $\mathrm{W}_Q, \mathrm{W}_K, \mathrm{W}_V$ are learnable projections.

Finally, a reconstruction head predicts the original CNN features:

$$\hat{\mathrm{T}} = \mathrm{H}\mathrm{W}_r + \mathrm{b}_r, \tag{15}$$

$$\mathcal{L}_{\mathrm{rec}} = \|\hat{\mathrm{T}} - \mathrm{T}\|_2^2, \tag{16}$$

where $\hat{\mathrm{T}}$ and T are the reconstructed and original CNN features, and Wr and br are learnable parameters. Minimizing $\mathcal{L}$rec prevents latent collapse and preserves discriminative cues.

We denote the weighted sum of $\mathcal{L}_{\mathrm{KL}}$, $\mathcal{L}_{\mathrm{CMMD}}$, and $\mathcal{L}_{\mathrm{rec}}$ as $\mathcal{L}_{\mathrm{VAE}}$, with λ_1, λ_2 and λ_3 balancing the contributions of KL regularization, conditional distribution matching, and reconstruction, respectively.

Overall Training Objective

Let r denote the vectorized feature representation obtained by concatenating all channel embeddings. On top of r, we attach an emotion classifier and a domain discriminator with GRL to encourage discriminative prediction on the source domain while reducing marginal domain discrepancy. This yields the emotion classification loss $\mathcal{L}_{\mathrm{cls}}$ and the adversarial domain discrimination loss $\mathcal{L}_{\mathrm{dis}}$.

The final loss aggregates the three core terms:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{dis}}\mathcal{L}_{\mathrm{dis}} + \lambda_{\mathrm{vae}}\mathcal{L}_{\mathrm{VAE}}, \tag{17}$$

where λ_{dis} and λ_{vae} balance adversarial alignment and VAE regularization in the overall training. EmoVAE enables efficient conditional distribution alignment, yielding compact intra-class clusters and stable decision boundaries.

Experiments

Emotional Datasets

The experiments were conducted on two widely used benchmark emotion datasets:

- **SEED** (W.-L. Zheng & Lu, 2015) contains EEG recordings from 15 subjects and targets three emotion classes: negative, neutral, and positive. Emotional responses were induced via 15 carefully selected 4-minute video clips. Each subject participated in three separate sessions, and each session included 15 EEG recording trials.
- **SEED-VII** (Jiang et al., 2024) covers 20 subjects across four sessions. A total of 80 video clips were used to evoke seven emotions: disgust, fear, sad, neutral, happy, anger, and surprise. EEG signals were recorded in 20 trials per session, and subjects provided continuous labels reflecting emotional intensity.

Experiment Settings

Experiment Protocols To comprehensively evaluate the robustness and generalization ability of EmoVAE, we adopted two different protocols:

- Cross-Subject Protocol: This protocol employs the leave-one-out (LOSO) cross-validation method. In each iteration, the data from one subject is used as the test set, while the remaining data is used as the training set.
- Cross-Dataset Protocol: This protocol assesses the model's ability to transfer knowledge to completely different target datasets. In each iteration, all labeled data from the source domain and all unlabeled data from the target subjects are used to train the model. Then, the model is

Table 1: Summary of Cross-subject Model Performance on All Datasets

Datasets	Methods	P_{acc}	F1 score	P value
SEED	UDDA	87.11±06.35	88.31±07.08	0.055
	MS-MDA	88.93±06.72	89.76±06.37	0.053
	PR-PL	93.06±05.12	93.04±06.17	0.034
	EEGMatch	91.35±07.03	91.30±04.90	0.042
	NSAL-DGAT	95.69±05.24	95.36±01.20	0.030
	Ours	**97.93±01.90**	**97.65±02.11**	—
SEED-VII	UDDA	45.17±07.71	45.66±07.13	0.037
	MS-MDA	46.71±08.27	45.96±07.98	0.036
	PR-PL	52.23±06.26	52.40±10.90	0.025
	EEGMatch	60.42±02.94	60.88±01.23	0.031
	NSAL-DGAT	63.40±09.20	64.96±06.21	0.033
	Ours	**83.37±03.52**	**84.09±02.77**	—

Note: P_{acc}: Overall accuracy; P value: The result of a paired t-test based on P_{acc} between the proposed method and each compared method.

tested only using the data from that specific target subject. Significant discrepancies in label spaces across benchmark datasets make label unification a critical prerequisite for cross-dataset experiments. We adopt a unified three-class scheme consisting of positive, negative, and neutral, which is widely used in emotion recognition research. Accordingly, we remap SEED-VII emotion labels into these three categories. Specifically, "happy" and "surprise" are grouped into the positive class, while "sadness", "fear", "disgust" , and "anger" are grouped into the negative class.

Environment and Parameters We trained the model for 200 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4}. During training, both the source and target loaders use a batch size of 16, while the test loader uses a batch size of 64 for evaluation. For the loss weighting, we set $\lambda_1 = 0.1$, $\lambda_2 = 0.001$, $\lambda_3 = 1.0$, $\lambda_{\text{dis}} = 0.1$ and $\lambda_{\text{vae}} = 1.0$.

Cross-Subject Results

We test the model's ability to generalize to unseen individuals within a single domain through a cross-subject protocol on two datasets. We compared it with five other domain adaptation methods: UDDA (Z. Li et al., 2022), MS-MDA (H. Chen et al., 2021), PR-PL (Zhou et al., 2023), EEGMatch (Zhou et al., 2025) and NSAL-DGAT (Yang et al., 2025). Table 1 summarized the comprehensive results and provided a comparison with existing SOTA methods. Our proposed model achieved SOTA across all benchmark datasets, attaining an average accuracy of 97.93% on the SEED and 83.37% on SEED-VII datasets.

Cross-Dataset Results

We also conducted experiments in the challenging cross-dataset scenario. Table 2 compares the performance of existing methods in the cross-dataset setting. Our model achieved an accuracy of 55.00% in the SEED→SEED-VII task, and 51.70% for the transfer from SEED-VII→SEED, consistently yielding the best performance across the evaluated cross-dataset tasks.

Table 2: Cross-Dataset Performance Comparison

Dataset	Method	P_{acc}	F1 score	P value
SEED ↓ SEED-VII	UDDA	38.41±07.34	39.15±07.88	0.025
	MS-MDA	39.81±08.53	39.95±08.48	0.025
	PR-PL	42.52±09.07	43.66±09.02	0.017
	EEGMatch	45.04±08.02	46.18±07.98	0.023
	NSAL-DGAT	47.51±07.03	48.65±06.99	0.035
	Ours	**55.00±06.85**	**54.21±06.10**	—
SEED-VII ↓ SEED	UDDA	40.07±09.15	40.95±08.79	0.024
	MS-MDA	40.81±09.01	43.95±08.96	0.023
	PR-PL	44.62±09.43	45.76±09.38	0.013
	EEGMatch	47.71±08.62	48.85±08.58	0.019
	NSAL-DGAT	50.81±07.83	51.95±07.79	0.040
	Ours	**51.70±05.51**	**52.18±05.71**	—

Note: P_{acc}: Overall correctness; P value: The result of a paired t-tests based on P_{acc} between the proposed method and each compared method.

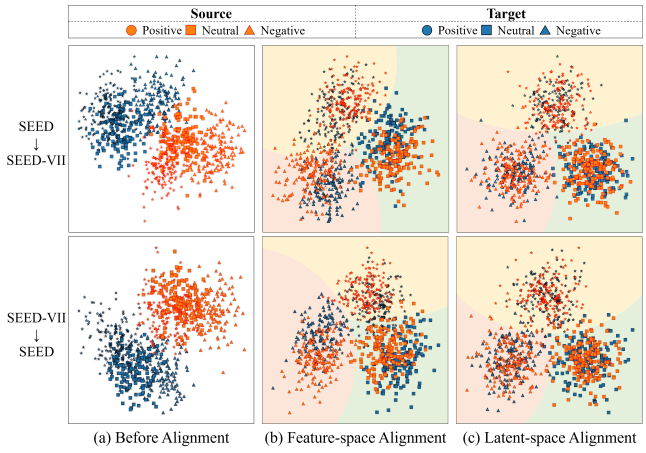


Figure 2: The t-SNE visualizations of source and target samples in cross-dataset experiments. (a) Before alignment, where source and target clusters show clear domain shift. (b) After alignment in the feature space, source and target samples become more overlapped. (c) After alignment in the latent space, the class clusters are further compact and better matched across domains.

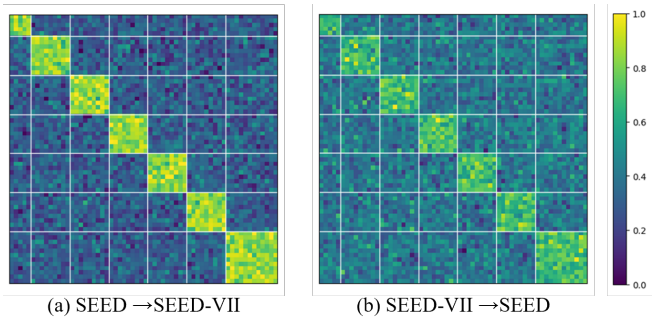
Figure 2 visualizes the feature distributions of different emotion classes in two domains, as well as the representations after feature-space alignment and latent-space alignment. Feature-space alignment often yields looser clusters with irregular geometry and occasional class-wise sticking. In contrast, latent-space alignment produces more structured embeddings: KL regularization anchors representations to a shared prior and curbs uncontrolled dispersion. Meanwhile, stochastic sampling encourages perturbation-robust representations. As a

Table 3: Ablation Experiments of Different Model Combinations

Datasets	Methods	P_{acc}	F1 score
SEED → SEED-VII	w/o L_a	53.23±06.21	53.11±06.36
	w/o L_b	54.24±05.84	53.88±06.19
	w/o L_c	51.77±05.80	51.12±06.07
	w/o L_d	50.51±05.97	50.66±06.39
	Ours	**55.00±06.85**	**54.21±06.10**
SEED-VII → SEED	w/o L_a	49.72±04.12	50.25±04.06
	w/o L_b	48.63±05.04	49.24±05.26
	w/o L_c	46.36±04.61	46.56±04.99
	w/o L_d	44.71±05.97	46.23±06.37
	Ours	**51.70±05.51**	**52.18±05.71**

Note: P_{acc} represents Classification Accuracy.

result, latent-space alignment yields a stable cross-domain alignment between source and target categories.



(a) SEED →SEED-VII (b) SEED-VII →SEED

Figure 3: The visualization of brain functional connectivity maps for two cross-dataset tasks. The channels of this map are arranged in the SEED order, from left to right being the frontal lobe, parietal lobe, frontotemporal region, central region, parietal lobe, temporal lobe, and occipital lobe.

Furthermore, Figure 3 presents the brain functional connection maps obtained through averaging in cross-dataset experiments. The matrix indicates that during emotional stimulation, adjacent brain regions often have stronger and clearer functional connections. Despite the domain shift between datasets, the overall connectivity pattern remains consistent, suggesting that local functional coordination constitutes a relatively stable neurophysiological signature that our model can exploit for cross-dataset generalization.

Cognitive Interpretability Results

To provide a neurocognitive interpretation, we estimated channel importance for neutral, positive and negative emotions using monte carlo random subset masking: randomly masking channel subsets and measuring the resulting performance drop. Figure 4 shows the topographies, where warmer colors indicate higher contributions.

Positive emotion exhibits stronger left-lateral contributions, mainly over frontal areas. Neutral shows a more diffuse and

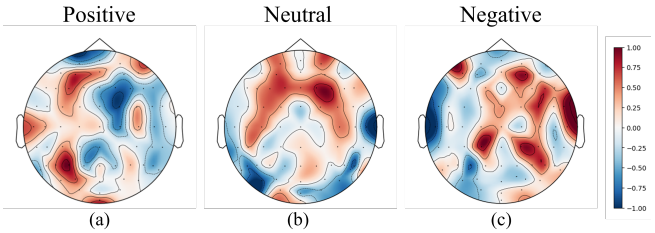


Figure 4: The neural activity distribution map on the SEED dataset: positive samples (left), neutral samples (middle), and negative samples (right).

balanced pattern with fewer focal peaks. Negative emotion reveals more right-lateral hotspots spanning frontal and centroparietal regions. Such valence-dependent hemispheric dissociation is consistent with the valence–arousal model and the frontal asymmetry hypothesis, which links affective valence to asymmetric prefrontal engagement. Overall, these various topographical features indicate that the original data contains spatial features that can distinguish emotions, which supports the modeling of spatiotemporal coupling dynamics and the alignment of representations in the latent space.

Ablation Study

To comprehensively validate the contribution of each key component in our proposed architecture, we performed an ablation study by systematically removing individual modules. Table 3 reports the ablation results.

We designed four ablation settings. The "w/o L_a" variant removes the temporal feature extraction module and relies only on the remaining representations for downstream learning. The "w/o L_b" variant discards the spatial/channel-interaction modeling component. The "L_c" variant disables the joint spatio-temporal feature extraction, preventing the model from capturing coupled temporal dynamics and global channel dependencies. The "L_d" variant removes the variational latent space construction and performs adaptation directly in the feature space. These results collectively demonstrate that each component of our proposed framework provides a distinct and significant contribution to the model's final performance.

Conclusion

In this paper, we propose EmoVAE, a domain adaptation framework for EEG-based emotion recognition under cross-subject and cross-dataset settings. EmoVAE integrates a multiscale spatiotemporal encoder to extract emotion-discriminative representations and an EEG-latent module that performs continuous conditional alignment in a variational latent space, while reconstruction preserves discriminative cues. Extensive evaluations on public EEG benchmarks under cross-subject and cross-dataset settings demonstrate that EmoVAE consistently outperforms SOTA methods. These results highlight the robustness and generalizability of our approach, and suggest its potential for affective computing.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant 62576142, the Guangdong Talent Plan under grant 2024TQ08X162 and the Guangdong Basic and Applied Basic Research Foundation under grant 2024A1515010524.

References

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

Chen, H., Jin, M., Li, Z., Fan, C., Li, J., & He, H. (2021). Ms-mda: Multisource marginal distribution adaptation for cross-subject and cross-session EEG emotion recognition. *Frontiers in Neuroscience*, *Volume 15 - 2021*. https://doi.org/10.3389/fnins.2021.778488

Chen, R., Xie, C., Zhang, J., You, Q., & Pan, J. (2025). A progressive multi-domain adaptation network with reinforced self-constructed graphs for cross-subject EEG-based emotion and consciousness recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *33*, 3498–3510. https://doi.org/10.1109/TNSRE.2025.3603190

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, *18*(1), 32–80.

Ding, Y., Robinson, N., Tong, C., Zeng, Q., & Guan, C. (2024). Lggnet: Learning from local-global-graph representations for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(7), 9773–9786. https://doi.org/10.1109/TNNLS.2023.3236635

Jiang, W.-B., Liu, X.-H., Zheng, W.-L., & Lu, B.-L. (2024). SEED-VII: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. *IEEE Transactions on Affective Computing*, *16*, 969–985.

Joshi, V. M., Ghongade, R. B., Joshi, A. M., & Kulkarni, R. V. (2022). Deep bilstm neural network model for emotion detection using cross-dataset approach. *Biomedical Signal Processing and Control*, *73*, 103407.

Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, *102*, 102019. https://doi.org/10.1016/j.inffus.2023.102019

Klepl, D., Wu, M., & He, F. (2024). Graph neural network-based EEG classification: A survey. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *32*, 493–503. https://doi.org/10.1109/TNSRE.2024.3355750

Li, Z., Zhang, G., Wang, L., Wei, J., & Dang, J. (2023). Emotion recognition using spatial-temporal EEG features through convolutional graph attention network. *Journal of Neural Engineering*, *20*(1), 016046.

Li, Z., Zhu, E., Jin, M., Fan, C., He, H., Cai, T., & Li, J. (2022). Dynamic domain adaptation for class-aware cross-subject and cross-session EEG emotion recognition. *IEEE Journal of Biomedical and Health Informatics*, *26*(12), 5964–5973. https://doi.org/10.1109/JBHI.2022.3210158

Nath, D., Anubhav, Singh, M., Sethia, D., Kalra, D., & Indu, S. (2020). A comparative study of subject-dependent and subject-independent strategies for EEG-based emotion recognition using lstm network. *Proceedings of the 2020 4th International Conference on Compute and Data Analysis*, 142–147.

Pan, T., Su, N., Shan, J., Tang, Y., Zhong, G., Jiang, T., & Zuo, N. (2025). GLADA: Global and local associative domain adaptation for EEG-based emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, *17*(1), 167–178. https://doi.org/10.1109/TCDS.2024.3432752

Wang, J., Ning, X., Xu, W., Li, Y., Jia, Z., & Lin, Y. (2024). Multi-source selective graph domain adaptation network for cross-subject EEG emotion recognition. *Neural Networks*, *180*, 106742.

Yang, Y., Wang, Z., Song, Y., Jia, Z., Wang, B., Jung, T.-P., & Wan, F. (2025). Exploiting the intrinsic neighborhood semantic structure for domain adaptation in EEG-based emotion recognition. *IEEE Transactions on Affective Computing*, *16*(3), 2466–2478. https://doi.org/10.1109/TAFFC.2025.3564272

Zheng, F., Hu, B., Zheng, X., & Zhang, Y. (2023). Spatial–temporal features-based EEG emotion recognition using graph convolution network and long short-term memory. *Physiological Measurement*, *44*(6), 065002.

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.

Zhou, R., Ye, W., Zhang, Z., Luo, Y., Zhang, L., Li, L., Huang, G., Dong, Y., Zhang, Y.-T., & Liang, Z. (2025). EEGMatch: Learning with incomplete labels for semisupervised EEG-based cross-subject emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(7), 12991–13005. https://doi.org/10.1109/TNNLS.2024.3493425

Zhou, R., Zhang, Z., Fu, H., Zhang, L., Li, L., Huang, G., Li, F., Yang, X., Dong, Y., Zhang, Y.-T., et al. (2023). PR-PL: A novel prototypical representation based pairwise learning framework for emotion recognition using EEG signals. *IEEE Transactions on Affective Computing*, *15*(2), 657–670.