Multi-Modal Domain Adaptation Variational Auto-encoder for EEG-Based Emotion Recognition

Yixin Wang, Shuang Qiu, Dan Li, Changde Du, Bao-Liang Lu, and Huiguang He

***Abstract*—Traditional electroencephalograph (EEG)-based emotion recognition requires a large number of calibration samples to build a model for a specific subject, which restricts the application of the affective brain computer interface (BCI) in practice. We attempt to use the multi-modal data from the past session to realize emotion recognition in the case of a small amount of calibration samples. To solve this problem, we propose a multi-modal domain adaptive variational autoencoder (MMDA-VAE) method, which learns shared cross-domain latent representations of the multi-modal data. Our method builds a multi-modal variational autoencoder (MVAE) to project the data of multiple modalities into a common space. Through adversarial learning and cycle-consistency regularization, our method can reduce the distribution difference of each domain on the shared latent representation layer and realize the transfer of knowledge. Extensive experiments are conducted on two public datasets, SEED and SEED-IV, and the results show the superiority of our proposed method. Our work can effectively improve the performance of emotion recognition with a small amount of labelled multi-modal data.**

***Index Terms*—Cycle-consistency, domain adaptation, electroencephalograph (EEG), multi modality, variational autoencoder.**

I. Introduction

EMOTION is a psychophysiological process triggered by the perception of stimulus, which plays a vital role in human behaviour, action and decision making [1]. With the development of human and machine communication, emotion recognition has increasingly become important for an advanced human-computer interaction system [2]. Since emotions are accompanied by a variety of external manifestations, the range of indicators of emotional state is wide, including facial expressions [3], voice [4], body language [5] to physiological signals [6]. Compared with other signals, physiological signals can capture participants' underlying responses. Electroencephalogram (EEG), as a high-resolution and effective physiological signal, has been widely used in the field of emotion recognition [7].

In conventional EEG-based emotion recognition, it is necessary to collect a large amount of calibration data from one person to train the effective model of this person to get good performance [8]. However, the collection of calibration data is time-consuming, which severely hinders the application of BCI in practice. A major challenge is using a small amount of data to quickly build a model. To address this issue, we design a multi-modal domain adaptation method to use the data from more modalities and previously collected data to improve classification performance under a small amount of samples.

It is difficult to precisely discriminate complex emotions using only one signal [9], while multi-modal fusion can exploit the complementarity of different signals. Recent studies indeed show that fusion of multiple modalities can improve emotion recognition performance significantly [10]–[12]. EEG has been reported to be a promising indicator to reflect emotion states since EEG signals can directly reflect brain activity [13]. Also, eye movement signals have become widely used for emotion recognition. It is because that they are important cues for context-aware environment, which convey important information for emotion recognition [14]. These two modalities combining central nervous signal and external behaviour have been considered to be a promising way to describe emotional states [11], [15]. Thus, we conduct this research on multi-modal signals (EEG signals and eye movement signals). Furthermore, we consider using a large amount of data collected by subjects on different days to improve the performance of emotion recognition, especially under a small amount of calibration samples. However, there is remarkable variability between the data collected in different days from one subject (which is also called a session). It is difficult to acquire multi-modal models that can work across sessions. Domain adaptation can be used to make the distribution of the source domain close to that of the target domain to improve the target domain's performance. The large amount of data collected in the past session from one

Manuscript received November 13, 2021; revised January 2, 2022; accepted January 23, 2022. This work was supported in part by National Natural Science Foundation of China (61976209, 62020106015, U21A20388); in part by the CAS International Collaboration Key Project (173211KYSB20190024); and in part by the Strategic Priority Research Program of CAS (XDB32040000). Recommended by Associate Editor Xin Luo. *(Corresponding author: Huiguang He.)*

Citation: Y. X. Wang, S. Qiu, D. Li, C. D. Du, B.-L. Lu, and H. G. He, "Multi-modal domain adaptation variational autoencoder for EEG-based emotion recognition," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 9, pp. 1612–1626, Sept. 2022.

Y. X. Wang, S. Qiu, D. Li, C. D. Du, and H. G. He are with the Research Center for Brain-inspired Intelligence, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing 100190, and the University of Chinese Academy of Sciences, Beijing 100049; Y. X. Wang is also with the Beijing Institute of Control and Electronic Technology, Beijing 100038, China; D. Li is also with the School of Mathematics and Information Sciences, Yantai University, Yantai 264003, China; H. G. He is also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Science, Beijing, China (e-mail: wangyxai@hotmail.com; shuang.qiu@ia.ac.cn; danliai@hotmail.com; duchangde@gmail.com; huiguang.he@ia.ac.cn).

B.-L. Lu is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: bllu@sjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2022.105515

subject is regarded as the source domain, and the small amount of data collected in a new day from the same subject is regarded as the target domain. In this paper, we try to use domain adaptation to build a cross-session multi-modal model for the improvement of emotion recognition. This has not been studied in previous studies [16].

There are three main challenges for multi-modal domain adaptation. Firstly, due to the addition of multi-modal information, we need to jointly model the heterogeneous features of different modalities to achieve semantic alignment. Secondly, there exists a domain gap between the source domain and the target domain, we need to reduce the distribution differences of each domain to effectively utilize the knowledge of the source domain. Thirdly, there are some samples that have incomplete modal representations, and it is desirable to solve the missing modality problem.

To resolve the problems mentioned above, we propose a multi-modal domain adaptive variational autoencoder (MMDA-VAE) method. Firstly, we build a multi-modal variational autoencoder (MVAE) to model the relationship of multi-modal emotional data, which can map multi-modal data in different domains into the same latent representation in a shared-latent space, and train a cross-session classifier on the shared-latent layer. On the one hand, the data from different domains (the source domain and the target domain) share the encoder paths of the MVAE, ensuring similarities between multi-modal data from different domains. On the other hand, we set up independent decoder paths of MVAE for each domain, which retains the characteristic of each domain's modality on the reconstruction layer. Secondly, in order to use the characteristic information of each modality, our method constrains the reconstructed data through adversarial learning loss and cycle-consistency loss rather than performing the transfer operation on the shared-latent layer. By performing domain confusion of each modality and multi-modal generation across domains in the reconstruction layer, the distance between the source domain and the target domain in the latent representation space can be implicitly shortened. In addition, we use the product of experts (PoE) rule to train the joint inference network for the joint posterior of the MVAE, which can efficiently learn the combined variational parameters missing modalities.

Our contributions are as follows. 1) We introduce the MMDA-VAE model that learns shared cross-domain latent representations of the EEG and eye movement data. 2) We propose two constraints: both adversarial learning loss and cycle-consistency loss to solve the multi-modal domain adaptation problem. 3) We extensively evaluate our model using two benchmark datasets, i.e., SEED and SEED-IV. The results show the superiority of our proposed method over traditional transfer learning methods and state-of-the-art deep domain adaptation methods.

II. Related Work

Multi-Modal Fusion: Multi-modal fusion is the concept to join information from two or more modalities to perform in some tasks [17], [18], which has been widely implemented for emotion recognition [11], [12], [19]–[22]. Lu *et al.* [11] used a fuzzy integral strategy to achieve modality fusion on EEG and eye movement signals. Liu *et al.* [20] used a bimodal deep autoencoder (BDAE) to extract shared representations of EEG and eye movement for the prediction of emotion states. Ranganathan *et al.* [12] exploited a multi-modal deep Boltzmann machine (DBM) to model feature distributions from face, body gesture, voice and physiological signals jointly for emotion classification. Multi-modal fusion can use more information provided by multi-modal data compared to single modal data, which improves the performance of emotion recognition. But there are two problems: When multi-modal data from different domains are simply input into the emotion recognition model. It becomes difficult to deal with missing modality [23].

Single-Modal Domain Adaptation: There are two categories of existing, shallow domain adaptation methods and deep domain adaptation methods [16]. Many traditional shallow domain adaptation methods have been applied in the emotion recognition field. Zheng and Lu [24] used four algorithms: transductive component analysis (TCA) [25], kernel principal component analysis (KPCA) [26], transductive support vector machine (TSVM) [27], and transductive parameter transfer (TPT) [28] based on SVM to build a general model for new target subjects. Chai *et al.* [29] proposed an adaptive subspace feature matching (ASFM) method, developed a linear transformation function to match the marginal distributions of the two domains' subspaces. In recent years, with the rapid development of deep learning, deep domain adaptation methods have become a popular research topic in the emotion recognition field. Li *et al.* [30] proposed a bi-hemisphere domain adversarial neural network (BiDANN) method to improve the generality of the EEG-based emotion recognition model. Li *et al.* [31] used association reinforcement loss on deep neural network (DNN) to adapt the joint distribution of the source and target domains. Single-modal domain adaptation has been widely studied and successfully applied to classification problems with a small number of samples. However, few studies explored the domain adaptation of multiple modalities. These methods simply used cascade features as input to deal with multi-modal problems without mining the relationships between modalities. By our model, the source and the target domain can make full use of the information from more than one modality to get better performance.

Cross-Modal Generative Models: The variational autoencoder (VAE) [32] is one of the deep generative models, which can be used to reconstruct data across domains in the field of domain adaptation. The VAE-based models use cross-domain reconstruction to capture the common information contained in the two domains in the shared latent-space. Shen *et al.* [33] proposed cross-aligned VAE to ensure that the latent text space of different domains have similar representations. Liu *et al.* [34] modelled each image domain by a VAE-GAN architecture and matched the latent representations from different domains. The VAE-based model has the structure of an encoder and decoder. While reducing the dimension, it ensures that the hidden layer can extract cross-modal and

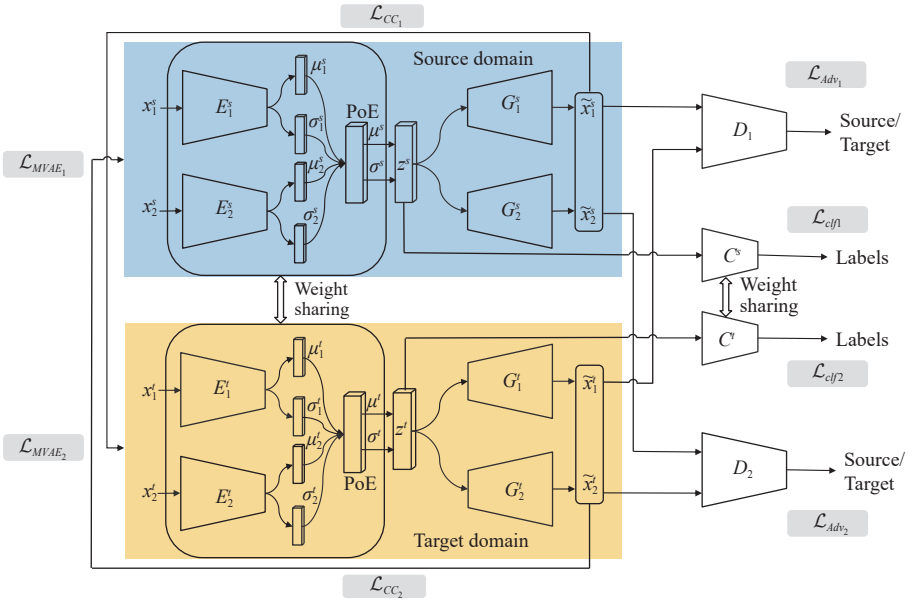


Fig. 1. The proposed MMDA-VAE framework. Two dashed boxes connected by the double arrow illustrate that the encoders of the target domain E_1^t and E_2^t share the weights of the source domain E_1^s and E_2^s. We represent the encoders and decoders using DNNs and use the PoE network to combine all the encoders and solve the modality missing problem. Here, $\widetilde{x}_1^s$ and $\widetilde{x}_2^s$ are reconstructed from the source domain. $\widetilde{x}_1^t$ and $\widetilde{x}_2^t$ are reconstructed from the target domain. D_1 and D_2 are adversarial discriminators for the respective modalities. In addition, we feed the reconstructed multi-modal data into the paths of the other domain and constrain the output to be same as the original data to achieve the VAE-like cycle consistency loss.

cross-domain information [35]–[37]. This reconstructed structure is naturally suitable for solving the problem of multi-modal domain adaptation. These methods have been used previously for text-style transfer [33] and image-to-image translation [34], which were rarely applied to multi-modal electrophysiological data. Therefore, they have the potential to solve similar problems in the field of emotion recognition.

III. Methodology

A. Background

Problem Formulation: This paper focuses on the scenarios where the source domain and the target domain both have two emotional modal types (i.e., EEG and eye movement). We denote the source domain as $\mathcal{D}_s = \{(x_{1,i}^s, x_{2,i}^s, y_i^s)\}_{i=1}^{n^s}$, where $(x_{1,i}^s, x_{2,i}^s)$ is the i-th EEG/eye movement data with label y_i^s. In the target domain, we are given a limited number of the labelled target data $\mathcal{D}_{tl} = \{(x_{1,i}^{tl}, x_{2,i}^{tl}, y_i^{tl})\}_{i=1}^{n^{tl}}$, and unlabelled target data $\mathcal{D}_{tu} = \{(x_{1,i}^{tu}, x_{2,i}^{tu})\}_{i=1}^{n^{tu}}$. The aim of MMDA-VAE is to train the model on $\mathcal{D}_s$ and $\mathcal{D}_{tl}$, and then evaluate on $\mathcal{D}_{tu}$.

VAE: The basic building block of our model is one VAE [32]. VAE is a latent variable generative model of the form $p_\theta(x,z) = p(z)p_\theta(x|z)$, where $p(z)$ is a prior probability, usually Gaussian. The decoder $p_\theta(x|z)$ consists of a deep neural network with the parameter θ and has a simple likelihood (e.g., Bernoulli or Gaussian). Finding the true conditional distribution on the latent variables $p_\theta(z|x)$ is the aim of variational inference. Since this distribution is interactive, it can be approximated by finding its closest proxy, $q_\phi(z|x)$, and then using the lower bound of variation to minimize their distance. The objective function or evidence lower bound (ELBO), can be defined as

$$\mathcal{L} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x)\|p_\theta(z)) \tag{1}$$

where the first term is the reconstruction error and the second term is the Kullback-Leibler divergence between distributions p and q. Encoder predictions μ and Σ where $q_\phi(z|x) = \mathcal{N}(\mu, \Sigma)$, generates a latent vector z by re-parameterization techniques.

B. MMDA-VAE

In this section, we first introduce a MVAE model [38], which learns the joint latent representation from multi-modal data. And then, we extend it to the multi-modal domain adaptation by adding discriminators for adversarial learning and cycle-consistency constraints. The whole structure of our method is shown in Fig. 1. For the sake of readability, we list frequently used symbols and their definitions in Table I.

MVAE: Different from single modal VAE, MVAE uses a generative model of the form $p_\theta(x_1, x_2, \ldots, x_N) = p(z)p_\theta(x_1|z)p_\theta(x_2|z)\ldots p_\theta(x_N|z)$ where $x_1, x_2, \ldots, x_N$ are N different modalities and z is a common latent variable. The ELBO becomes

$$\mathcal{L} = \mathbb{E}_{q_\phi(z|x_{1\ldots N})}[\sum_{i=1}^{N} \log p_\theta(x_i|z)] - KL(q_\phi(z|\boldsymbol{X})\|p_\theta(z)). \tag{2}$$

In order to solve the missing modality problem, we exploit a PoE structure [39]. If the individual distributions are uni- or multivariate Gaussians, their product will also be a multivariate Gaussian. The individual models can be called "experts"; multiplying the complicated expert distributions together and renormalizing can be very powerful.

For any subset $X \subseteq \{x_1, \ldots, x_N\}$, we define the joint inference network $q(z|\boldsymbol{X})$ as

$$q(z|\boldsymbol{X}) \propto p(z) \prod_{x_i \in X} q(z|x_i) \tag{3}$$

TABLE I
DEFINITION OF FREQUENTLY USED SYMBOLS

Symbol	Definition
$(\mathcal{X}_1^s, \mathcal{X}_2^s)$	The source domain
$(\mathcal{X}_1^t, \mathcal{X}_2^t)$	The target domain
(x_1^s, x_2^s)	A sample of the source domain
(x_1^t, x_2^t)	A sample of the target domain
$(\widetilde{x}_1^s, \widetilde{x}_2^s)$	A reconstructed sample of the source domain
$(\widetilde{x}_1^t, \widetilde{x}_2^t)$	A reconstructed sample of the target domain
E_1^s, E_2^s	Encoders of the source domain
E_1^t, E_2^t	Encoders of the target domain
G_1^s, G_2^s	Decoders of the source domain
G_1^t, G_2^t	Decoders of the target domain
z^s	Latent code of the source domain
z^t	Latent code of the target domain
D_1	Discriminator of the first modality
D_2	Discriminator of the second modality
C^s	Classifier of the source domain
C^t	Classifier of the target domain

where a prior expert $p(z)$ is a form of regularization. We assume that μ_i and σ_i are the i-th variational parameters of the uni-modal inference network $q(z|x_i)$, which is the expert. We can use a PoE, including a "prior expert", as the approximating distribution for the joint-posterior. The mean and the covariance of the multi-modality distribution are given as follows:

$$\mu = (\sum_i \mu_i \Sigma_i^{-1})(\sum_i \Sigma_i^{-1})^{-1}$$
$$\Sigma = (\sum_i \Sigma_i^{-1})^{-1}. \tag{4}$$

Thus, we can compute all multi-modal inference networks required for MVAE efficiently in terms of the N uni-modal components. If a modality is missing during training, MVAE can drop the corresponding inference network and use the existing modality sufficiently.

The encoder-decoder set $\{E_1^s, E_2^s, G_1^s, G_2^s\}$ constitutes a MVAE of the source domain for two modalities $\mathcal{X}_1^s$ and $\mathcal{X}_2^s$. The encoder-decoder pair $\{E_1^t, E_2^t, G_1^t, G_2^t\}$ constitutes the other MVAE of the target domain for two modalities $\mathcal{X}_1^t$ and $\mathcal{X}_2^t$. We assume that the multi-modal data (x_1^s, x_2^s) of the source domain in two modalities $\mathcal{X}_1^s$ and $\mathcal{X}_2^s$ can be mapped to a latent code z by E_1^s and E_2^s. The multi-modal data (x_1^t, x_2^t) of the target domain in $\mathcal{X}_1^t$ and $\mathcal{X}_2^t$ modalities can also be mapped to the same latent code z by E_1^t and E_2^t. G_1^s, G_2^s, G_1^t and G_2^t are decoding functions, which can map latent codes to multi-modal data. The encoders of the target domain E_1^t and E_2^t share the weights of the source domain E_1^s and E_2^s to catch the shared information between two domains. We set up independent decoder paths G_1^s, G_2^s, G_1^t and G_2^t, so that each modality of each domain retains its own characteristics in the reconstruction layer. Thus, we have two losses: $\mathcal{L}_{MVAE_1}$ and $\mathcal{L}_{MVAE_2}$ for the source and target data as (5) and (6) shown.

$$\begin{aligned}\mathcal{L}_{MVAE_1}(E_1^s, E_2^s, G_1^s, G_2^s) = {} & \lambda_1 \mathbb{E}_{z^s \sim q^s(z^s|x_1^s, x_2^s)}[\log p_{G_1^s}(x_1^s|z^s)] \\ & + \lambda_1 \mathbb{E}_{z^s \sim q^s(z^s|x_1^s, x_2^s)}[\log p_{G_2^s}(x_2^s|z^s)] \\ & - \lambda_2 KL(q^s(z^s|x_1^s, x_2^s)\|p_\theta(z))\end{aligned} \tag{5}$$

$$\begin{aligned}\mathcal{L}_{MVAE_2}(E_1^t, E_2^t, G_1^t, G_2^t) = {} & \lambda_1 \mathbb{E}_{z^t \sim q^t(z^t|x_1^t, x_2^t)}[\log p_{G_1^t}(x_1^t|z^t)] \\ & + \lambda_1 \mathbb{E}_{z^t \sim q^t(z^t|x_1^t, x_2^t)}[\log p_{G_2^t}(x_2^t|z^t)] \\ & - \lambda_2 KL(q^t(z^t|x_1^t, x_2^t)\|p_\theta(z)).\end{aligned} \tag{6}$$

Adversarial Learning: Our framework has two discriminators D_1 and D_2 to implement adversarial learning of the respective modalities. G_1^s can generate the first modality data generated from the multi-modal source data and G_2^s can generate the second modality data from the multi-modal source data. Similarly, G_1^t and G_2^t can generate the first and second modalities from the multi-modal target data. We used two adversarial discriminators D_1 and D_2 for the respective modalities, in charge of evaluating whether the reconstructed data are generated from the source domain or the target domain. D_1 is trained to confuse the first modality data reconstructed from the source domain G_1^s and the target domain G_1^t. In a similar way, D_2 can deal with the second modality data and make the second modality of the source and target data closer. The GAN objective functions are given by (7) and (8).

$$\begin{aligned}\mathcal{L}_{Adv_1}(E_1^s, E_2^s, E_1^t, E_2^t, G_1^s, G_1^t, D_1 = {} & \\ & \lambda_3 \mathbb{E}_{z^s \sim q^s(z^s|x_1^s, x_2^s)}[\log D_1(G_1^s(z^s))] \\ & + \lambda_3 \mathbb{E}_{z^t \sim q^t(z^t|x_1^t, x_2^t)}[\log(1 - D_1(G_1^t(z^t)))]\end{aligned} \tag{7}$$

$$\begin{aligned}\mathcal{L}_{Adv_2}(E_1^s, E_2^s, E_1^t, E_2^t, G_2^s, G_2^t, D_2) = {} & \\ & \lambda_3 \mathbb{E}_{z^s \sim q^s(z^s|x_1^s, x_2^s)}[\log D_2(G_2^s(z^s))] \\ & + \lambda_3 \mathbb{E}_{z^t \sim q^t(z^t|x_1^t, x_2^t)}[\log(1 - D_2(G_2^t(z^t)))].\end{aligned} \tag{8}$$

Cycle-Consistency: There exists a cycle-consistency mapping so that we can reconstruct the original input data from transferring back the reconstructed input data. Based on this principle, the multi-modal reconstruction data $(\widetilde{x}_1^s, \widetilde{x}_2^s)$ of the source domain is sent to the shared encoder and decoder of the target domain. The output data should be close to the original source data (x_1^s, x_2^s). In the same way, the multi-modal reconstructed data $(\widetilde{x}_1^t, \widetilde{x}_2^t)$ of the target domain through the MVAE path of the source domain can be returned to the original target data (x_1^t, x_2^t). By this cross-domain generation, it can further ensure that the data in the two domains have the same shared representation. We use a VAE-like objective function to model the cycle-consistency constraint, which are given by (9) and (10).

$$\begin{aligned}\mathcal{L}_{CC_1}(E_1^t, E_2^t, G_1^t, G_2^t = {} & \lambda_4 \mathbb{E}_{z^t \sim q^t(z^t|\widetilde{x}_1^s, \widetilde{x}_2^s)}[\log p_{G_1^t}(x_1^s|z^t)] \\ & + \lambda_4 \mathbb{E}_{z^t \sim q^t(z^t|\widetilde{x}_1^s, \widetilde{x}_2^s)}[\log p_{G_2^t}(x_2^s|z^t)] \\ & - \lambda_5 KL(q^t(z^t|\widetilde{x}_1^s, \widetilde{x}_2^s)\|p_\eta(z)) \\ & - \lambda_5 KL(q^s(z^s|x_1^s, x_2^s)\|p_\eta(z))\end{aligned} \tag{9}$$

$$\mathcal{L}_{CC_2}(E_1^s,E_2^s,G_1^s,G_2^s)=\lambda_4\mathbb{E}_{z^s\sim q^s(z^s|\widetilde{x}_1^t,\widetilde{x}_2^t)}[\log p_{G_1^s}(x_1^t|z^s)]+\lambda_4\mathbb{E}_{z^s\sim q^s(z^s|\widetilde{x}_1^t,\widetilde{x}_2^t)}[\log p_{G_2^s}(x_2^t|z^s)]-\lambda_5 KL(q^s(z^s|\widetilde{x}_1^t,\widetilde{x}_2^t)\|p_\eta(z))-\lambda_5 KL(q^t(z^t|x_1^t,x_2^t)\|p_\eta(z)) \quad (10)$$

where there are two KL terms to penalize the latent codes deviating from the prior distribution, since z^s and z^t are not exactly the same. Log-likelihood objective terms ensure twice transferred data resembles the original.

Classification Loss: We add one softmax layer after the latent representation layer to classify the source samples and a small fraction of target samples. The classification losses are composed of two cross-entropy losses from the source domain and the target domain, which are shown as follows:

$$\mathcal{L}_{clf1}(E_1^s,E_2^s,C^s)=\lambda_6\mathbb{E}_{(x_1^s,x_2^s,y^s)\sim(\mathcal{X}_1^s,\mathcal{X}_2^s,\mathcal{Y}^s)}[\mathcal{L}_{ce}(M(E_1^s(x_1^s),E_2^s(x_2^s)),y^s)] \quad (11)$$

$$\mathcal{L}_{clf2}(E_1^t,E_2^t,C^t)=\lambda_7\mathbb{E}_{(x_1^{tl},x_2^{tl},y^{tl})\sim(\mathcal{X}_1^{tl},\mathcal{X}_2^{tl},\mathcal{Y}^{tl})}[\mathcal{L}_{ce}(M(E_1^{tl}(x_1^{tl}),E_2^{tl}(x_2^{tl})),y^{tl})] \quad (12)$$

where $M(\cdot)$ represents the multi-modal fusion function.

Overall Objective:

$$\begin{aligned}\min_{\boldsymbol{E},\boldsymbol{G},\boldsymbol{C}}\max_{\boldsymbol{D}}\ &\mathcal{L}_{MVAE_1}(E_1^s,E_2^s,G_1^s,G_2^s)\\&+\mathcal{L}_{Adv_1}(E_1^s,E_2^s,E_1^t,E_2^t,G_1^s,G_1^t,D_1)\\&+\mathcal{L}_{CC_1}(E_1^t,E_2^t,G_1^t,G_2^t)\\&+\mathcal{L}_{MVAE_2}(E_1^t,E_2^t,G_1^t,G_2^t)\\&+\mathcal{L}_{Adv_2}(E_1^s,E_2^s,E_1^t,E_2^t,G_2^s,G_2^t,D_2)\\&+\mathcal{L}_{CC_2}(E_1^s,E_2^s,G_1^s,G_2^s)\\&+\mathcal{L}_{clf1}(E_1^s,E_2^s,C^s)+\mathcal{L}_{clf2}(E_1^t,E_2^t,C^t)\end{aligned} \quad (13)$$

where $\boldsymbol{E}=\{E_1^s,E_2^s,E_1^t,E_2^t\}$, $\boldsymbol{G}=\{G_1^s,G_2^s,G_1^t,G_2^t\}$, $\boldsymbol{C}=\{C^s,C^t\}$ and $\boldsymbol{D}=\{D_1,D_2\}$. We first apply a gradient ascent step to update $\boldsymbol{D}$ with $\boldsymbol{E},\boldsymbol{G},\boldsymbol{C}$ fixed and then apply a gradient descent step to update $\boldsymbol{E},\boldsymbol{G},\boldsymbol{C}$ with $\boldsymbol{D}$ fixed. The inference procedure of our method is shown in Fig. 2.

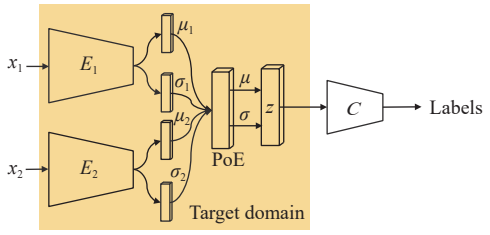


Fig. 2. The inference procedure of MMDA-VAE. The test sample is sent to the shared encoder path to get the latent representation z, and the final classification label is obtained through the classifier with shared parameters.

IV. Experiments

A. Datasets

SEED [11]: There were 15 film clips chosen to evoke three target emotions (positive, negative and neutral). The duration of each film clip was around four minutes. Fifteen volunteers were asked to watch these films three times, at an interval of one week or longer. There were 15 trials (five trials per emotion) for each session and the film clips for these three sessions were repeated. The raw EEG data was simultaneously recorded at a 1000 Hz sampling rate with 62 channels using the ESI NeuroScan System. Nine of the volunteers also simultaneously recorded eye movements. Since only nine volunteers collected eye movement signals, we used the data of these nine volunteers in our paper.

SEED-IV [40]: There were 72 film clips chosen to evoke four target emotions (happy, sad, fear, and neutral). The duration of each film clip was approximately two minutes. Fifteen volunteers were asked to watch these films in three days as three different sessions. Each session consisted of 24 trials (six trials per emotion), and the stimuli for these three sessions were completely different. The raw EEG data was simultaneously recorded at a 1000 Hz sampling rate with 62 channels using the ESI NeuroScan System. Eye movement signals were also simultaneously recorded using SMI ETG eye-tracking glasses.

For both two datasets, we performed similar data preprocessing and feature extraction. To further filter the noise and remove the artefacts, the EEG signals were processed with a band-pass filter between 1 and 75 Hz. Then, the EEG and eye movement data were re-sampled to reduce the computational complexity and align these two modalities.

After data preprocessing, we extracted the differential entropy (DE) [41] feature from EEG data. The DE feature is defined as follows:

$$h(X)=-\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{(x-\mu)^2}{2\sigma^2}\log\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{(x-\mu)^2}{2\sigma^2}dx=\frac{1}{2}\log 2\pi\sigma^2. \quad (14)$$

For SEED, short-term Fourier transforms with a 1 s time window without overlapping was used, for SEED-IV, we used a 4 s time window. The length of the EEG segment is strictly based on the setting in the originally published paper [11], [40]. The DE features can be calculated in five frequency bands: delta (1−4 Hz), theta (4−8 Hz), alpha (8−14 Hz), beta (14−31 Hz), and gamma (31−50 Hz), where we used all bands features for these two datasets. As for eye movements, the parameters collected by the eye tracker include pupil diameter, fixation dispersion, saccade amplitude, saccade duration, and blink. We extracted features such as mean, standard deviation, DE and so on, the details of which were consistent with the original literature [40].

B. Setup

We regarded three sessions of the same subject as three domains, where the target domain was one session data of the subject, and his/her existing sessions too turn as the source domain. Thus, we could create six transfer tasks: 1→2, 1→3, 2→1, 2→3, 3→1, 3→2 on both two datasets.

The training set consisted of the source data and the labelled target data, i.e., all the samples from the source session and

samples from the first three or four trials (one trial per class) in the other target session. We used samples from the second three or four trials in the target session as the validation set. The samples from the rest of the twelve or sixteen trials in the target session were used to evaluate classification accuracy. The average accuracies of all the subjects in the dataset were reported. The details of two datasets used in our experiments were summarized in Table II.

TABLE II
PROPERTIES OF THE DATA USED IN EXPERIMENTS

Dataset	Modality	Class	Training set	Validation set	Testing set
SEED	EEG, eye movement	3	1012	148	521
SEED-IV		4	957	110	606

All the encoders and decoders in VAE architecture were DNNs with three hidden layers. The hidden units of the encoders were 256-256-50, and the decoders were 50-256-256, where the latent embedding size was 50. To implement the adversarial learning, we used two three-layer DNNs as the discriminators, the hidden units of which are 256-256-2. The model was trained for 30 epochs by stochastic gradient descent using the Adam optimizer with a learning rate of 0.001. Every iteration, we prepared two mini-batches, one consisting of the source samples and the other of the labelled target samples. The batch sizes were 40 and 10, respectively. There were eight hyper-parameters in our model. We selected λ_1, λ_2, λ_3, λ_4 from $10^{[-4:-2]}$ and λ_5 from $10^{[-5:-3]}$. λ_6 was set to 10 and λ_7 was set to 0.1. Besides, we used L2 regularization and selected the parameter λ_{L2} in the range of $10^{[-4:-2]}$. The validation set was used for early stopping and determined these hyper-parameters.

C. Compared Methods

Our method was mainly compared with two types of methods, namely, the multi-modal fusion method and the single-modal domain adaptation method.

Multi-Modal Fusion Method: Since modality fusion based on multi-modal electrophysiological signals has been studied in the field of emotion recognition, we compared several baseline methods of modality fusion often used in this field, which included FLF, DLF based on SVM [11], discriminant correlation analysis (DCA)+LDA [42] and BDAE based on DBM [40].

- **FLF [11]** concatenates the EEG feature vector and the eye movement feature vector into a larger feature vector.
- **DLF [11]** combines the classification results of two classifiers to obtain the final decision, where the maximal (sum) rule was to calculate the maximal (sum) values of all the probabilities.
- **DCA + LDA [42]** removes the inter-class correlation and limits the correlation to that found within classes, which is often used in feature fusion for biometric recognition.
- **BDAE [40]** trains two individual restricted Boltzmann machines (RBMs) to extract the shared representations of both two modalities.

FLF is a conventional modality fusion method based on SVM, while BDAE is a common modality fusion method in deep learning. Both of them have been widely used in multi-modal emotion recognition. In order to verify the effect of different input conditions, we additionally designed two types of baseline based on FLF and BDAE.

- **Source only (SO)** uses the source samples to classify the unlabelled target samples.
- **Target only (TO)** uses the labelled target samples to train models, without the help of the source domain.

Single-Modal Domain Adaptation Method: SVM was the most common method implemented for EEG emotion recognition, and three conventional methods (KPCA, TSVM, and TCA) based on the SVM classifier that was often used as baselines for the domain adaptation problem. In recent years, with the development of deep learning, many deep domain adaptation methods have been applied in EEG emotion recognition. We compared the proposed method with both conventional and deep domain adaptation methods, with details were as follows:

- **KPCA [26]** uses a low transfer dimensional space using kernel methods.
- **TSVM [27]** uses the decision boundary in a semi-supervised manner and weights all training instances equally.
- **TCA [25]** aims to use some transfer components to embedding features into a high-dimensional space to preserve the shared attributes between two domains.
- **DDC [43]** is based on classic deep network architectures, and a linear-kernel MMD loss is added on the feature representation layer to maximize domain invariance.
- **DAN [44]** embeds all the task-specific layers' representations into a Reproducing Kernel Hilbert Space (RKHS), where the mean embeddings of two domain distributions can be matched.
- **DANN [45]** integrates a gradient reversal layer into the deep network, which can ensure that the features are domain-invariant and discriminative for the classification task.
- **JAN [46]** uses a transfer network by aligning the joint distribution of multiple domain-specific layers across multiple domains.
- **ADA [47]** produces statistically domain invariant embeddings, while minimizing the classification error on the labelled source domain by reinforcing associations between source and target data in the embedding space.
- **CDAN [48]** conditions the adversarial adaptation models using discriminative information to align different domains of multi-modal distributions.
- **CoGAN [49]** uses the joint distribution with just samples drawn from the marginal distributions by enforcing a weight-sharing constraint.
- **UNIT [34]** uses a VAE-GAN architecture to learn a joint distribution of data in different domains by using data from the marginal distributions in individual domains.
- **DAAN [50]** dynamically learns domain-invariant representations while quantitatively evaluating the relative importance of global and local domain distributions.

For fair comparison, we modified the above methods so that they were trained with the labelled source samples and the labelled target samples.

TABLE III
MEAN ACCURACY (%) FOR MODALITY FUSION EMOTION RECOGNITION ON THE SEED DATASET

Method	Input	1→2	1→3	2→1	2→3	3→1	3→2	Average
FLF-SO [11]	Source	78.50 ± 18.86	76.82 ± 22.44	73.06 ± 12.25	70.61 ± 23.19	71.98 ± 17.32	76.16 ± 14.88	74.71 ± 18.57***
BDAE-SO [40]		79.49 ± 9.84	77.00 ± 18.64	75.09 ± 11.08	77.82 ± 16.85	75.73 ± 14.56	79.56 ± 15.78	77.45 ± 14.46***
FLF-TO [11]	Target	53.38 ± 18.02	47.30 ± 10.73	63.49 ± 16.74	47.30 ± 10.73	63.49 ± 17.74	53.38 ± 18.03	54.72 ± 15.17***
BDAE-TO [40]		58.16 ± 13.76	49.73 ± 13.79	66.11 ± 13.92	49.73 ± 13.79	66.11 ± 13.92	58.16 ± 13.76	58.00 ± 13.82***
FLF [11]	Source + Target	80.32 ± 16.97	71.42 ± 25.72	69.42 ± 11.09	71.23 ± 22.87	72.59 ± 17.99	76.65 ± 16.74	73.60 ± 18.56***
DLF-SUM [11]		76.92 ± 14.56	78.12 ± 12.13	85.22 ± 15.29	82.79 ± 14.29	76.80 ± 9.25	76.71 ± 21.34	79.43 ± 14.48***
DLF-MAX [11]		76.39 ± 14.42	77.99 ± 12.11	82.32 ± 14.33	83.05 ± 15.13	77.54 ± 10.35	73.02 ± 19.62	78.39 ± 14.33***
DCA+LDA [42]		76.24 ± 13.47	67.50 ± 12.13	78.12 ± 17.28	71.32 ± 7.30	76.80 ± 15.16	76.24 ± 13.47	74.37 ± 8.16
BDAE [40]		83.78 ± 12.48	79.33 ± 17.96	76.96 ± 19.22	83.02 ± 17.10	##38; 77.42 ± 13.38	80.29 ± 13.31	80.13 ± 15.57***
MMDA-VAE		**93.27 ± 9.18**	**88.22 ± 13.27**	**89.47 ± 11.54**	**88.47 ± 15.52**	**89.60 ± 10.04**	**88.82 ± 11.17**	**89.64 ± 11.78**

* There are significant differences between our proposed method and other comparison methods (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).

TABLE IV
MEAN ACCURACY (%) FOR MODALITY FUSION EMOTION RECOGNITION ON THE SEED-IV DATASET

Method	Input	1→2	1→3	2→1	2→3	3→1	3→2	Average
FLF-SO [11]	Source	60.34 ± 17.63	60.55 ± 23.41	73.06 ± 12.25	67.12 ± 19.55	59.13 ± 11.95	64.87 ± 14.88	62.98 ± 16.58***
BDAE-SO [40]		60.38 ± 15.06	62.65 ± 15.22	63.99 ± 13.88	69.36 ± 14.88	56.52 ± 14.57	64.45 ± 13.74	62.89 ± 14.59***
FLF-TO [11]	Target	51.81 ± 16.92	14.33 ± 8.31	38.05 ± 12.23	14.33 ± 8.31	38.05 ± 12.23	51.81 ± 16.92	34.73 ± 12.12***
BDAE-TO [40]		44.89 ± 14.00	48.29 ± 14.00	35.58 ± 6.83	48.29 ± 14.00	35.58 ± 6.83	44.89 ± 14.00	37.77 ± 10.38***
FLF [11]	Source + Target	70.17 ± 14.18	51.95 ± 22.61	65.86 ± 12.03	57.26 ± 20.68	64.74 ± 12.24	71.65 ± 14.62	63.60 ± 16.06***
DLF-SUM [11]		65.78 ± 19.01	49.70 ± 16.25	62.16 ± 13.93	52.76 ± 17.53	62.66 ± 10.10	68.34 ± 13.79	60.23 ± 15.10***
DLF-MAX [11]		65.67 ± 17.06	41.66 ± 14.55	59.22 ± 10.29	49.53 ± 18.50	59.11 ± 11.84	66.39 ± 14.07	56.91 ± 14.39***
DCA+LDA [42]		66.50 ± 16.21	52.09 ± 15.41	61.95 ± 12.34	59.81 ± 14.89	62.52 ± 14.69	72.22 ± 10.12	63.39 ± 9.21
BDAE [40]		70.43 ± 13.14	54.86 ± 15.97	66.08 ± 8.86	60.39 ± 15.11	65.57 ± 8.93	71.84 ± 12.10	64.86 ± 12.35***
MMDA-VAE		**76.14 ± 13.61**	**70.60 ± 15.34**	**75.38 ± 11.41**	**75.10 ± 13.40**	**71.72 ± 6.25**	**73.97 ± 10.44**	**73.82 ± 11.74**

* There are significant differences between our proposed method and other comparison methods (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).

D. Statistical Analysis

In this paper, the *t*-test was conducted to analyse the difference of comparison results and other subsequent experiments. Post hoc analysis was conducted with a Benjamini & Hochberg correction. The significant level was set at 0.05. All results are presented as *mean* ± *std* deviation.

V. RESULTS

In this section, we designed a series of experiments to verify our method. Firstly, our methods are compared with the multi-modal fusion method and single-mode state domain adaptation method. Secondly, ablation studies were performed to show the effects of the two components. Thirdly, we designed the missing modality situation, the cross-subject situation, the different transfer strategy situation and the session-independent/-dependent situation to show comprehensive performance. In addition, visualization and sensitivity were conducted.

A. Comparison Results

Comparison With Modality Fusion Methods: The results on the SEED dataset and the SEED-IV dataset were summarized in Tables III and IV. Our proposed method significantly outperformed FLF-SO and BDAE-SO. Also, MMDA-VAE achieved a significant improvement compared with FLF-TO and BDAE-TO on both datasets. The results show that MMDA-VAE can obtain better performance than directly using the source data to classify the unlabelled target data or using a small quantity of labelled target samples to train models. On both two datasets, MMDA-VAE showed significant improvement over FLF, DLF-SUM, DLF-MAX, DCA+LDA and BDAE. Our method was 9.51% and 8.96% higher than the best performing modality fusion method, BDAE, on the two datasets. This demonstrates that our MMDA-VAE is designed to solve the problem of multi-modal domain adaptation and achieved significantly better results.

Comparison With Domain Adaptation Methods: The domain adaptation results on the SEED dataset and the SEED-IV dataset were shown in Tables V and VI, separately. The input of these comparison methods was the concatenation of EEG and eye movement for fair. Our method achieved high accuracies of 89.64% and 73.82% on the two datasets. On the SEED dataset and the SEED-IV dataset, the performance of MMDA-VAE was significantly better than conventional methods, KPCA, TSVM, and TCA. Compared with the classic deep domain adaptation methods, MMDA-VAE significantly

TABLE V
MEAN ACCURACY (%) FOR DOMAIN ADAPTATION EMOTION RECOGNITION ON THE SEED DATASET

Method	1→2	1→3	2→1	2→3	3→1	3→2	Average
KPCA [26]	79.74 ± 16.79	79.09 ± 18.06	75.04 ± 14.21	71.75 ± 19.86	72.77 ± 17.30	76.96 ± 18.22	75.89 ± 17.41***
TSVM [27]	79.24 ± 12.30	77.80 ± 14.21	74.48 ± 11.42	71.17 ± 16.85	73.02 ± 12.24	76.94 ± 14.98	75.44 ± 13.31***
TCA [25]	81.05 ± 13.07	78.93 ± 12.42	75.04 ± 14.21	72.74 ± 16.37	72.39 ± 9.24	77.48 ± 12.52	76.44 ± 12.81***
DDC [43]	89.02 ± 10.02	86.28 ± 16.57	83.46 ± 11.88	84.38 ± 19.05	80.34 ± 15.60	81.88 ± 17.17	84.23 ± 15.05*
DAN [44]	90.04 ± 10.54	80.85 ± 18.86	82.50 ± 14.94	85.68 ± 19.66	78.06 ± 13.30	86.62 ± 14.13	83.96 ± 15.24**
DANN [45]	89.09 ± 9.72	83.44 ± 18.86	83.87 ± 15.68	86.96 ± 18.76	78.24 ± 16.60	86.18 ± 12.23	84.63 ± 15.21*
JAN [46]	89.76 ± 8.04	84.80 ± 18.62	82.62 ± 13.74	87.62 ± 18.97	76.73 ± 13.92	86.80 ± 13.88	84.72 ± 14.19*
ADA [47]	86.07 ± 7.96	80.02 ± 24.17	77.18 ± 12.58	84.32 ± 14.12	77.41 ± 18.61	83.93 ± 15.07	81.49 ± 15.42*
CDAN [48]	82.88 ± 8.24	79.89 ± 19.08	79.98 ± 11.09	82.84 ± 17.92	78.80 ± 18.83	82.80 ± 15.72	81.20 ± 15.14***
CoGAN [49]	80.13 ± 6.75	78.53 ± 16.09	76.52 ± 10.39	78.76 ± 13.56	76.22 ± 13.64	80.96 ± 8.94	78.52 ± 11.56***
UNIT [34]	89.62 ± 8.76	82.35 ± 20.96	81.82 ± 10.32	84.47 ± 17.75	79.42 ± 13.88	84.19 ± 15.13	83.64 ± 14.47*
DAAN [50]	91.21 ± 8.52	86.38 ± 16.25	86.98 ± 12.77	87.84 ± 15.23	86.60 ± 10.74	88.40 ± 11.76	88.02 ± 9.34
MMDA-VAE	**93.27 ± 9.18**	**88.22 ± 13.27**	**89.47 ± 11.54**	**88.47 ± 15.52**	**89.60 ± 10.04**	**88.82 ± 11.17**	**89.64 ± 11.78**

* There are significant differences between our proposed method and other comparison methods (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).

TABLE VI
MEAN ACCURACY (%) FOR DOMAIN ADAPTATION EMOTION RECOGNITION ON THE SEED-IV DATASET

Method	1→2	1→3	2→1	2→3	3→1	3→2	Average
KPCA [26]	60.85 ± 17.52	60.98 ± 21.68	66.27 ± 12.32	69.48 ± 19.86	61.26 ± 14.63	67.32 ± 12.89	64.36 ± 16.48***
TSVM [27]	60.45 ± 12.24	61.11 ± 17.62	66.31 ± 11.42	68.91 ± 16.93	60.49 ± 11.93	65.78 ± 15.01	63.83 ± 14.19***
TCA [25]	63.91 ± 21.28	61.26 ± 20.68	67.02 ± 12.47	68.02 ± 16.43	59.31 ± 15.65	67.66 ± 17.11	64.53 ± 17.27***
DDC [43]	69.85 ± 12.57	69.09 ± 18.83	69.73 ± 16.33	74.96 ± 15.51	62.82 ± 15.28	67.09 ± 15.08	68.92 ± 15.60*
DAN [44]	69.60 ± 15.89	70.10 ± 18.42	69.15 ± 15.69	73.31 ± 15.91	61.22 ± 15.98	67.96 ± 15.90	68.56 ± 16.25**
DANN [45]	72.32 ± 12.39	68.52 ± 18.16	70.95 ± 16.31	74.74 ± 16.13	63.08 ± 15.25	67.96 ± 15.60	69.59 ± 15.63*
JAN [46]	70.52 ± 13.98	70.46 ± 18.64	71.39 ± 14.96	74.93 ± 14.98	63.53 ± 16.01	68.63 ± 15.50	69.91 ± 15.68*
ADA [47]	73.49 ± 13.25	67.88 ± 15.32	66.81 ± 15.58	73.00 ± 16.07	62.65 ± 15.98	68.50 ± 15.10	68.72 ± 15.21*
CDAN [48]	66.49 ± 16.30	63.72 ± 17.52	65.38 ± 16.62	72.82 ± 16.43	62.77 ± 9.58	68.57 ± 13.81	66.63 ± 17.95***
CoGAN [49]	63.84 ± 15.64	63.58 ± 17.45	64.38 ± 12.67	70.17 ± 12.21	61.59 ± 12.71	65.34 ± 10.49	64.82 ± 15.09***
UNIT [34]	72.46 ± 13.61	70.45 ± 16.02	70.59 ± 14.86	74.35 ± 15.97	66.14 ± 9.88	70.40 ± 15.96	70.73 ± 14.39*
DAAN [50]	74.83 ± 13.02	70.60 ± 15.34	74.51 ± 14.86	74.99 ± 17.57	68.29 ± 15.26	72.74 ± 15.00	72.41 ± 9.39
MMDA-VAE	**76.14 ± 13.61**	**70.60 ± 15.34**	**75.38 ± 11.41**	**75.10 ± 13.40**	**71.72 ± 6.25**	**73.97 ± 10.44**	**73.82 ± 11.74**

* There are significant differences between our proposed method and other comparison methods (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).

beat the DDC, DAN, DANN, JAN, DAAN, and ADA methods. In contrast to the deep generative models, CDAN, CoGAN, and UNIT, our method showed significantly better performance on both datasets. This demonstrates that our MMDA-VAE is more suitable for the domain adaptation problem of multi-modal data.

We selected several representative single-modal domain adaptation methods as comparison methods, and only a single modality was used for training and testing. The results were shown in Fig. 3. Compared to only using an EEG modality, our method significantly beats the conventional methods KPCA and TCA, on two datasets and beats the deep domain adaptation methods, DAN, DANN, and UNIT. As for only using eye movement, it was also significantly better than all the baselines on two datasets. This shows that using two modalities of emotion recognition performed better than the usage of a single modality. Moreover, in the comparison of EEG and eye movement modalities, the EEG modality performs better.

B. Ablation Study

We conducted ablation studies to evaluate the effects of two main components, which included cycle consistency loss and adversarial loss. Tables VII and VIII showed the results tested on the two datasets. On the SEED dataset, our method showed significant improvement over "No cycle consistency loss" with increases of 2.96% and increases of 1.86% "No adversarial loss". On the SEED-IV dataset, our MMDA-VAE also achieved 2.29% and 1.31% higher values than "No cycle consistency loss" and "No adversarial loss". Therefore, for the MMDA-VAE method, cycle consistency loss and adversarial loss have a great impact on the performance. The best

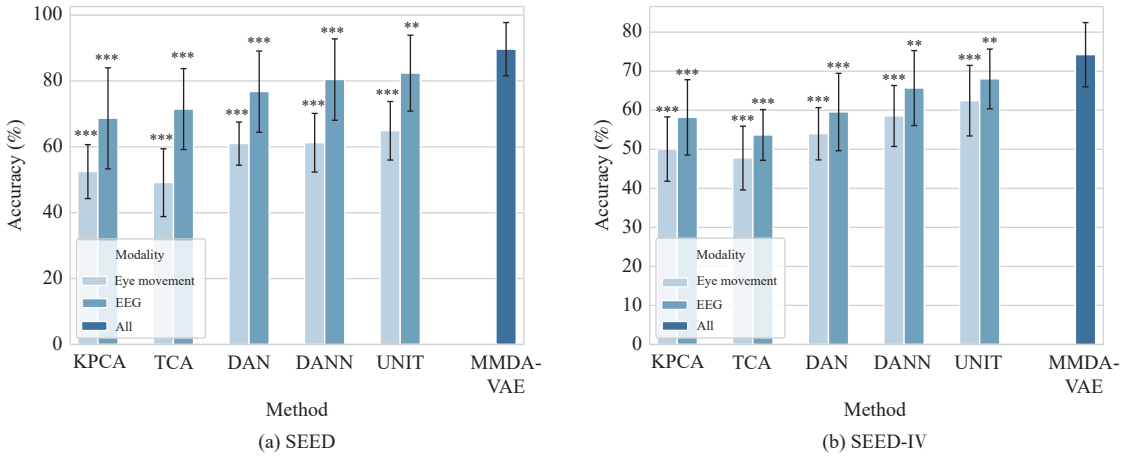


Fig. 3. Performance comparison for some domain adaptation methods using a single modality on (a) the SEED dataset and (b) the SEED-IV dataset. We carry out the significance test between our method and other single-modal domain adaptation methods (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).

TABLE VII
MEAN ACCURACY (%) FOR THE ABLATION STUDY ON THE SEED DATASET

Method	1→2	1→3	2→1	2→3	3→1	3→2	Average
No cycle consistency loss	90.34 ± 9.30	84.64 ± 14.62	84.57 ± 12.77	86.05 ± 15.71	86.32 ± 9.64	86.60 ± 11.39	86.68 ± 12.07**
No adversarial loss	92.18 ± 10.79	84.72 ± 13.56	86.18 ± 11.48	86.90 ± 16.70	87.56 ± 10.13	86.24 ± 13.36	87.78 ± 12.60*
No transfer loss	88.53 ± 12.23	82.88 ± 13.31	82.59 ± 12.40	85.00 ± 17.58	83.87 ± 13.27	84.15 ± 13.88	85.00 ± 13.82**
MMDA-VAE	**93.27 ± 9.18**	**88.22 ± 13.27**	**89.47 ± 11.54**	**88.47 ± 15.52**	**89.60 ± 10.04**	**88.82 ± 11.17**	**89.64 ± 11.78**

* There are significant differences between our proposed method and other comparison methods (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).

TABLE VIII
MEAN ACCURACY (%) FOR THE ABLATION STUDY ON THE SEED-IV DATASET

Method	1→2	1→3	2→1	2→3	3→1	3→2	Average
No cycle consistency loss	70.50 ± 14.11	67.49 ± 15.84	71.87 ± 11.71	73.10 ± 14.66	70.27 ± 7.91	72.18 ± 10.25	71.53 ± 12.20***
No adversarial loss	75.24 ± 13.45	68.29 ± 14.32	72.83 ± 11.02	74.26 ± 12.79	71.44 ± 7.11	72.99 ± 10.79	72.51 ± 11.58***
No transfer loss	71.99 ± 14.10	64.49 ± 17.94	70.21 ± 10.74	71.44 ± 13.23	68.72 ± 7.34	70.77 ± 11.57	69.59 ± 12.48***
MMDA-VAE	**76.14 ± 13.61**	**70.60 ± 15.34**	**75.38 ± 11.41**	**75.10 ± 13.40**	**71.72 ± 6.25**	**73.97 ± 10.44**	**73.82 ± 11.74**

* There are significant differences between our proposed method and other comparison methods (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).

accuracy is obtained when the two transfer losses are used together.

C. Missing Modality Results

In order to verify the feasibility of our method in the condition of a missing modality, we conducted a comparison strategy: "All" was equal to the original setting; "Only EEG" means that the training component and testing component of the target domain only contained EEG modality; "Only Eye" meant that there was only an eye movement modality in the target domain. The source domain contains the two modalities of EEG and eye movement. The results were shown in Tables IX and X. When we used all the modalities of the target domain, the results showed significant improvement when compared to only using the eye movement modality and only using EEG modality on both datasets. Besides, only using the EEG modality is better than only using the eye movement modality on the SEED-IV dataset. This result shows that using multi-modal data has better performance than using single-modal data. And among these single-modal datasets, EEG data provides more emotional information than eye movement data. It also demonstrates that our MMDA-VAE can handle the domain adaptation problem in the case of incomplete modalities between domains.

D. Cross-Subject Results

To verify the generalizability of our proposed method, we conducted cross-subject emotion recognition experiments. We selected the first session's data, and considered one subject in this session as the target subject in turn while considering the remaining subjects in the same session as source subjects. We trained multiple models on multiple source domains from other subjects, and reported the voting predictions of all the models on the target domain. We repeated two comparison experiments of modality fusion and domain adaptation in this way, which were same as the cross-session's baseline.

The results of the modality fusion methods on these two datasets were summarized in Table XI. Our method achieved the highest accuracies 85.07% and 75.52% on the two data-

TABLE IX
MEAN ACCURACY (%) FOR THE MISSING MODALITY EXPERIMENTS ON THE SEED DATASET

Method	1→2	1→3	2→1	2→3	3→1	3→2	Average
Only EEG	82.56 ± 8.04	81.44 ± 8.83	81.02 ± 9.72	81.87 ± 8.39	79.33 ± 9.84	79.47 ± 7.20	86.62 ± 12.86*
Only Eye	78.76 ± 12.35	76.24 ± 17.97	74.70 ± 10.30	74.85 ± 16.12	76.37 ± 12.26	77.46 ± 14.24	84.51 ± 13.70***
All	**93.27 ± 9.18**	**88.22 ± 13.27**	**89.47 ± 11.54**	**88.47 ± 15.52**	**89.60 ± 10.04**	**88.82 ± 11.17**	**89.64 ± 11.78**

* There are significant differences between our proposed method and other comparison methods (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).
† There is no significant difference between "Only EEG" and "Only Eye".

TABLE X
MEAN ACCURACY (%) FOR THE MISSING MODALITY EXPERIMENTS ON THE SEED-IV DATASET

Method	1→2	1→3	2→1	2→3	3→1	3→2	Average
Only EEG	71.67 ± 13.92	66.99 ± 13.62	67.48 ± 12.76	69.15 ± 16.30	67.40 ± 9.09	71.17 ± 15.11	68.80 ± 10.64**†
Only Eye	68.78 ± 11.28	55.12 ± 16.08	65.97 ± 9.59	63.55 ± 7.13	62.47 ± 6.77	67.91 ± 10.42	66.45 ± 10.50***
All	**76.14 ± 13.61**	**70.60 ± 15.34**	**75.38 ± 11.41**	**75.10 ± 13.40**	**71.72 ± 6.25**	**73.97 ± 10.44**	**73.82 ± 11.74**

* There are significant differences between our proposed method and other comparison methods (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).
† There are significant differences between "Only EEG" and "Only Eye" (†: $p < 0.05$).

TABLE XI
MEAN ACCURACY (%) FOR CROSS-SUBJECT MODALITY FUSION EMOTION RECOGNITION

Method	Input	Accuracy	
		SEED	SEED-IV
FLF-SO [11]	Source	63.53 ± 20.11**	63.87 ± 6.86***
BDAE-SO [40]		66.86 ± 8.83*	64.89 ± 7.17***
FLF-TO [11]	Target	63.49 ± 16.74*	38.05 ± 12.23***
BDAE-TO [40]		67.97 ± 14.68*	35.58 ± 6.83***
FLF [11]	Source + Target	69.14 ± 13.22**	67.40 ± 13.30*
DLF-SUM [11]		73.04 ± 10.60*	69.94 ± 9.32*
DLF-MAX [11]		72.42 ± 9.09*	67.99 ± 11.20*
DCA+LDA [42]		68.67 ± 15.28	66.79 ± 11.22**
BDAE [40]		75.22 ± 12.16**	70.21 ± 10.78*
MMDA-VAE		**85.07 ± 11.81**	**75.52 ± 10.21**

* There are significant differences between our proposed method and other comparison methods (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).

TABLE XII
MEAN ACCURACY (%) FOR CROSS-SUBJECT DOMAIN ADAPTATION EMOTION RECOGNITION

Method	Accuracy	
	SEED	SEED-IV
KPCA [26]	65.30 ± 10.16**	64.55 ± 9.82***
TSVM [27]	64.50 ± 7.20**	65.12 ± 13.83*
TCA [25]	76.63 ± 12.30*	66.36 ± 9.92**
DDC [43]	78.89 ± 12.84*	70.88 ± 8.66
DAN [44]	76.95 ± 11.52**	67.91 ± 8.55**
DANN [45]	79.84 ± 12.91	72.60 ± 11.78
JAN [46]	81.13 ± 13.27	73.76 ± 9.03
ADA [47]	78.59 ± 10.49	71.31± 9.18
CDAN [48]	73.64 ± 18.96*	72.85 ± 8.11
CoGAN [49]	73.79 ± 9.70***	69.87 ± 6.15*
UNIT [34]	76.64 ± 13.07*	73.33 ± 7.60
DAAN [50]	83.54 ± 9.81	75.36 ± 8.87
MMDA-VAE	**85.07 ± 11.81**	**75.52 ± 10.21**

* There are significant differences between our proposed method and other comparison methods (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).

sets. These results show that cross-subject experiments have achieved the same effect as cross-session experiments with modality fusion, which significantly exceeded the baseline. The result of domain adaptation methods on these two datasets were shown in Table XII. Our proposed method was significantly better than these conventional methods, KPCA, TSVM, and TCA. For the deep domain adaptation methods, our performance exceeded all the comparison methods of the two dataset, but only DDC, DAN, CDAN, CoGAN, and UNIT show significant differences. These results show that the performance of cross-subject experiments is consistent with that of cross-session experiments in the conventional domain adaptation, but the performance of cross-subject experiments is not as good as that of cross-session experiments when compared with the deep domain adaptation. On the whole, our MMDA-VAE can obtain optimal performance under two types of experimental settings, i.e., cross-session and cross-subject settings, indicating that it can effectively deal with various situations in the BCI's practical application.

E. Different Transfer Strategy Results

Our MMDA-VAE used a deep generation model to reconstruct multi-modal data, and operated on the reconstruction layer. We compared our transfer strategies with two transfer methods that operated on the feature-fusion layer. We added the MK-MMD constraint on the latent feature layer of MVAE, which is called MVAE-MMD [44] and added a discriminator to confuse the fused representation of MVAE from two domains, which is called MVAE-adversarial [45]. Tables XIII and XIV summarized the results on the SEED dataset and the SEED-IV dataset. On both two datasets, the averaged differences between the results of MVAE-MMD and MVAE-adversarial were 0.49% and 0.04%, respectively. The results showed that the MMD-based method and adversarial-based method had very similar performance on our basic framework MVAE. Our MMDA-VAE method achieved significantly

TABLE XIII
MEAN ACCURACY (%) FOR THE DIFFERENT TRANSFER STRATEGY EXPERIMENTS ON THE SEED DATASET

Method	1→2	1→3	2→1	2→3	3→1	3→2	Average
MVAE-MMD	88.16 ± 10.41	81.62 ± 16.05	81.93 ± 11.04	82.78 ± 17.89	82.31 ± 11.89	83.09 ± 12.90	83.31 ± 13.36***
MVAE-adversarial	88.00 ± 11.01	82.24 ± 15.00	80.53 ± 10.02	82.24 ± 18.36	81.58 ± 12.47	82.33 ± 14.04	82.82 ± 13.49***
MMDA-VAE	**93.27 ± 9.18**	**88.22 ± 13.27**	**89.47 ± 11.54**	**88.47 ± 15.52**	**89.60 ± 10.04**	**88.82 ± 11.17**	**89.64 ± 11.78**

* There are significant differences between our proposed method and other comparison methods (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).

TABLE XIV
MEAN ACCURACY (%) FOR THE DIFFERENT TRANSFER STRATEGY EXPERIMENTS ON THE SEED-IV DATASET

Method	1→2	1→3	2→1	2→3	3→1	3→2	Average
MVAE-MMD	70.05 ± 11.81	64.01 ± 15.22	70.08 ± 12.06	73.96 ± 11.65	68.31 ± 7.18	72.14 ± 11.87	69.76 ± 11.63***
MVAE-adversarial	69.43 ± 12.99	64.27 ± 14.43	69.94 ± 11.88	74.29 ± 11.55	67.90 ± 6.94	72.47 ± 11.51	69.72 ± 11.54***
MMDA-VAE	**76.14 ± 13.61**	**70.60 ± 15.34**	**75.38 ± 11.41**	**75.10 ± 13.40**	**71.72 ± 6.25**	**73.97 ± 10.44**	**73.82 ± 11.74**

* There are significant differences between our proposed method and other comparison methods (***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$).

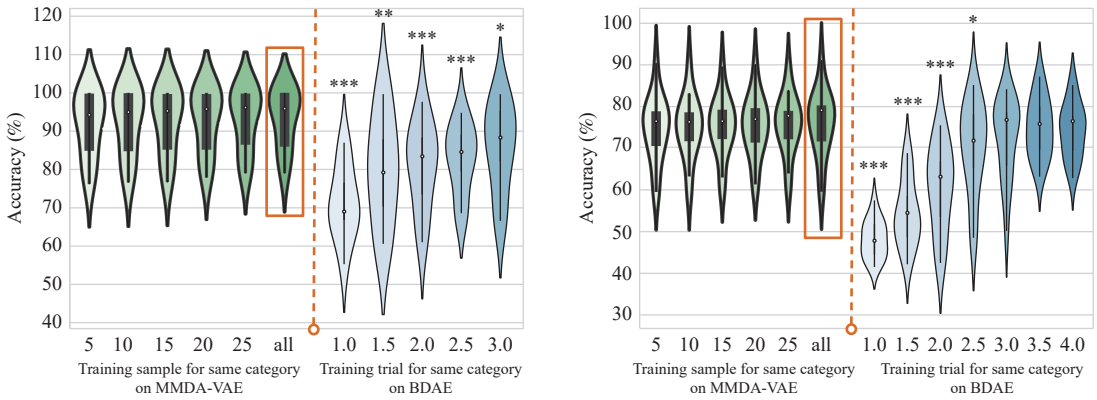


Fig. 4. Performance comparison for session-independent and session-dependent experiments on (a) the SEED dataset and (b) the SEED-IV dataset.

higher classification accuracy over the MVAE-MMD and MVAE-adversarial methods on both two datasets. The results show that our MMDA-VAE gets a better domain-invariant on the reconstruction layer, compared to directly constraining the latent space of MVAE.

F. Session-Independent and Session-Dependent Results

In order to verify the effectiveness of domain adaptation technology in the field of EEG emotion recognition, we integrated the session-independent and session-dependent experiments together for comparison. We keep the test set and valid set unchanged, and explored the performance of the model by changing the training set. The result curve is shown in Fig. 4. On the SEED dataset, we used the last 3 trials as the test set, with the 9th to 12th trials as the valid set. On the SEED-IV dataset, we used the last 4 trials as the test set, and the 16th to 20th trials as the valid set.

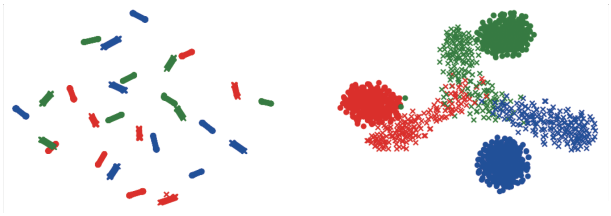
To investigate the influence of the number of calibration data samples (labelled target samples), we conducted a session-independent experiment on the MMDA-VAE model on the left side of the orange dotted line. We selected the samples from the first trial set for each category to be 5, 10, 15, 20, 25, and samples of the complete trial ("all"). We calculated the *t*-test between "all" and other cases, but there was no significant difference on both two datasets. It demonstrates that the number of calibration samples had little influence on the classification performance of MMDA-VAE.

To compare our method with the traditional session-dependent method, we trained a series of BDAE models which continuously increased the number of trials in each category. The compared MMDA-VAE benchmark is marked with the orange box, and 0.5 indicates that only the first half of the trial is used for training. On the SEED dataset, our method significantly outperformed the BDAE using 3.0 calibration trials in each category, achieving a 7.66% higher value. There was a similar phenomenon on the SEED-IV dataset. The accuracy was higher by 6.04%. This result shows that the MMDA-VAE can exceed the performance of traditional session-dependent multi-modal emotion recognition, effectively reducing the calibration time.

G. Space Visualization

T-distributed stochastic neighbour embedding (*t*-SNE) is a useful technique for dimensionality reduction. Here, we used *t*-SNE to visualize the distribution of the learnt latent representation. For simplicity, we randomly chose one target subject on task 1→2 of the SEED dataset. In Fig. 5(a), the source data and the target data were discrete, and the data

collected from the same video seemed like a strip. The emotional states were not easy to discriminate based on these original features. In Fig. 5(b), the discrepancy between two domains was obviously reduced in the latent representation space of our MMDA-VAE. Besides, the representations belonging to different categories became more distinguishable than before. As seen, our MMDA-VAE can effectively learn domain-invariant representations with discriminative category information.



(a) Before adaptation (b) After adaptation

Fig. 5. The distribution of source and target domain samples before and after the adaptation on task 1→2 of the SEED dataset.

H. Sensitivity Analysis

To clarify the effect of these hyper-parameters, we conduct experiments with different scaling constants of the λ_1–λ_5 and λ_{L2}. We chose the selected hyper-parameter from {1, 0.1, 0.01, 0.001, 0.0001} and kept other hyper-parameters optimal. The results on task 1→2 of the SEED dataset from one subject are shown in Fig. 6. For different λ_1, λ_3, λ_4, and λ_{L2}, the performance of our method fluctuates slightly. For different λ_2 and λ_5, our method is relatively variable. It shows that the weight of KL divergence loss had a greater impact on the result. Importantly, the performance is generally stable within the setting range (λ_1–λ_4 and λ_{L2} from {0.01, 0.001, 0.0001}; λ_5 from {0.001, 0.0001, 0.00001}).

VI. Discussion

In this paper, we combined EEG and eye movements to identify human emotional states. Our model made good use of the information of the two modalities. We can use a large amount of multi-modal data as the source domain to assist in providing a small amount of single-modal data in the target domain. Moreover, our method can naturally deal with the problem of missing modalities through the PoE structure. Although the performance of the modal missing data is lower than that of multi-modal data, the cost of multi-modal data collection is also a problem that needs to be balanced. It is meaningful for the application of emotional BCIs.

The traditional supervised learning method involves collecting a large amount of labelled data to train the classification model. In the case of a small amount of calibration data, we hope to use data collected by the same subject at different sessions to assist in model training, and reduce the calibration data collection time. In order to achieve this goal, we need to use domain adaptation technology to eliminate the difference in data distribution between each session. The results show that with a small amount of calibration data, our model has obvious advantages over the baseline without domain adaptation. We also chose the most widely used modal fusion method BDAE to train a series of session-dependent experiments. The results show that our domain adaptation model can be comparable to the performance of collecting sufficient data for traditional classification. This suggests that the domain adaptation method can use cross-session data to improve the classification performance and save calibration time, which is a promising approach in the actual use of emotional BCIs.

In addition to using data collected by the same subject in different sessions, our method can also train a cross-subject model using the data from the other subject. Since EEG has individual differences, there are also differences in distribution of the data between different subjects. Generally speaking, the difference in the data collected from different subjects is greater than the difference in the data collected from the same subject in different session [51]. We summarized cross-subject and cross-session experiments in Table XV. The cross-subject result on the SEED dataset was 4.57% lower than the cross-session result, while the cross-subject result on the SEED-IV dataset was 1.70% higher than the cross-session result. The subjects in each session of the SEED-IV dataset watched different evoked videos, but the subjects in each session of the SEED dataset watched the same evoked videos repeatedly. So the difference between each session may be greater on the SEED-IV dataset. Since emotion is not an instantaneous state, we additionally calculated the accuracy of trial-wise, that is, the label of the entire trial is voted by the samples removed from the same trial. We found that the classification accuracy of trial-wise and sample-wise methods have similar results when performing the tasks of within-subject and cross-subject domain adaptation.

We found that in Tables V and VI, the accuracy of the SEED-IV dataset was 15.82% lower than that of the SEED dataset. This may be due to the fact that the SEED dataset is a three-category emotion classification dataset, and the SEED-IV dataset is a four-category dataset. We therefore use the accuracy ratio from Tables V and VI to the chance level for the performance comparison, the SEED dataset: 89.64%/33.33% = 2.69 and the SEED-IV dataset: 73.82%/25% = 2.95. Looking at accuracy ratio, MMDA-VAE shows no performance degradation on the SEED-IV dataset and the results suggest that the number of categories caused the difference in the accuracy.

Although our method achieves better performance than compared methods, this method has several limitations. Firstly, our method is based on the structure of VAE, including decoder and encoder. This results in a larger number of parameters in this method compared with other methods. Secondly, this method assigns the same weight for two modalities. Future works should propose an algorithm to adaptively assign weights for different modalities based on the amount of the information they supply. Thirdly, other physiological signals have also been studied to estimate emotion states, including electromyogram signals (EMG) [42], electrocardiograms (ECG) [52], and galvanic skin responses (GSR) [53]. Thus, future studies should consider employing more multi-modal recordings, in combination with improvement of this method, to further improve the performance of emotion

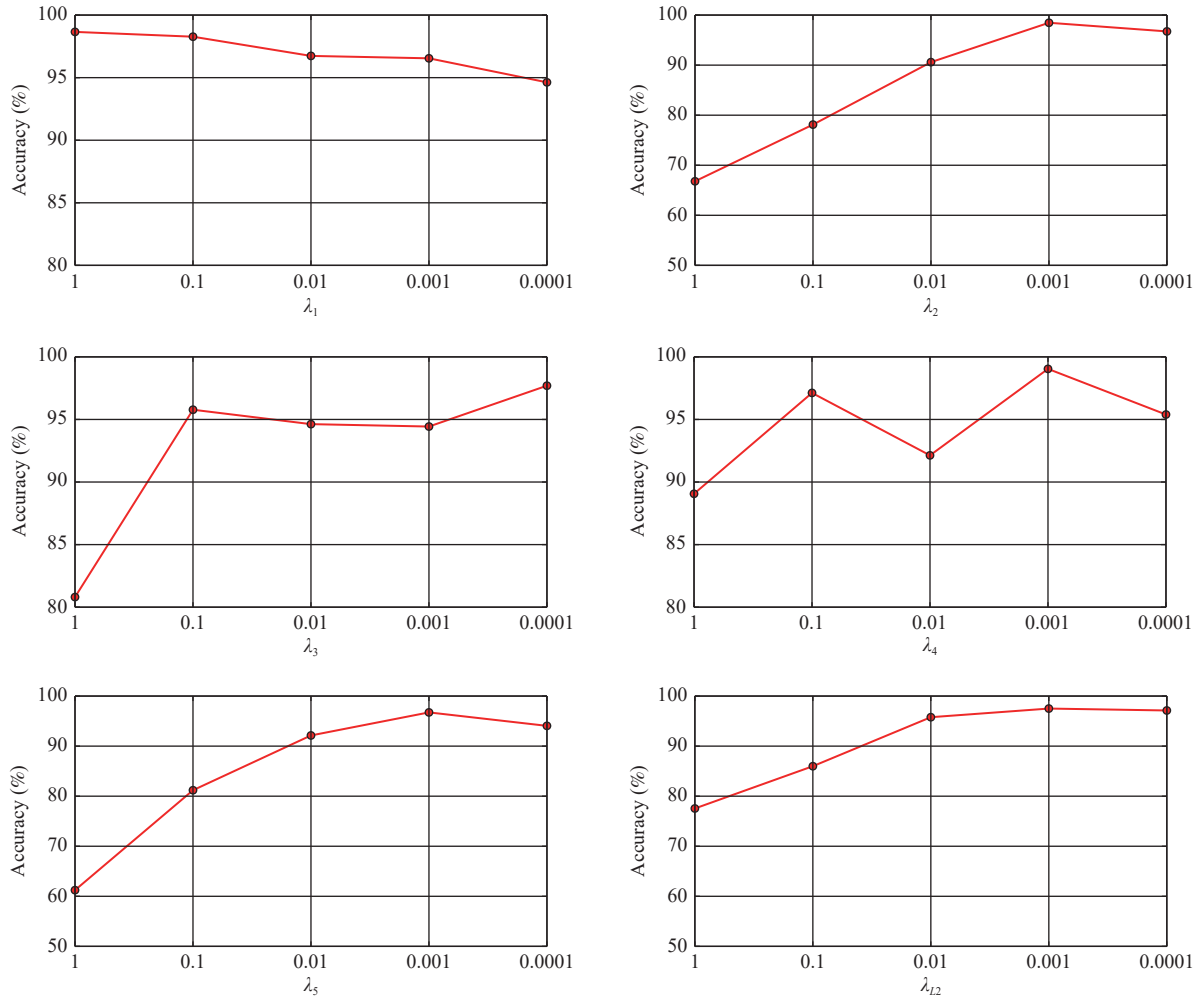


Fig. 6. The sensitivity analysis of hyper-parameters.

TABLE XV
MEAN ACCURACY (%) FOR DIFFERENT ACCURACY CALCULATION MODE

Validation method	Calculation mode	Accuracy	
		SEED	SEED-IV
Cross-session	Trial-wise	86.83 ± 13.04	72.74 ± 11.94
	Sample-wise	**89.64 ± 11.78**	**73.82 ± 11.74**
Cross-subject	Trial-wise	83.95 ± 15.82	73.75 ± 12.54
	Sample-wise	**85.07 ± 11.81**	**75.52 ± 10.21**

recognition. Finally, since this study is more concerned with improving the accuracy rate under the conditions of EEG practical usage, we do not further discuss the computational complexity.

VII. CONCLUSION

In this paper, we propose the MMDA-VAE method to achieve emotion recognition with small calibration samples. This method is based on the MVAE architecture, which ensures that the latent layer can extract the representation of modality fusion. We set up discriminators after the reconstruction layers of MVAE, and make the reconstruction data from the source and the target domains more confusing though adversarial learning. Also, we use the cycle-consistency constraint to convert the source domain and target domain data to each other during the reconstruction process. Our method can constrain the output space to narrow the modality gap and the domain gap, which helps the target subjects make use of the data of the other modalities and other sessions as much as possible. Indeed, the reconstructed structure leads to the increase of parameters. Our comprehensive experiments on two public datasets show the superior performance of the model. We also conducted some experiments with missing modalities. The results show that our MMDA-VAE can effectively deal with the difficulties in obtaining complete multi-modal data in practice. Our model provides a solution that overcomes the variability of data collected in different sessions and helps to reduce calibration time. This is a practical improvement in the field of emotion recognition based on EEG.

REFERENCES

[1] S. Spence, "Descartes' error: Emotion, reason and the human brain," BMJ, vol. 310, no. 6988, p. 1213, 1995. DOI: 10.1136/bmj.310.6988.1213.

[2] B. Shneiderman, C. Plaisant, M. Cohen, S. Jacobs, N. Elmqvist, and N. Diakopoulos, *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson, 2016.

[3] P. Ekman, "Expression and the nature of emotion," *Approaches to emotion*, vol. 3, pp. 19–344, 1984.

[4] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[5] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of Nonverbal Behavior*, vol. 28, no. 2, pp. 117–139, 2004.

[6] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. IEEE Face and Gesture*, pp. 827–834, 2011.

[7] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affective Computing*, vol. 1, no. 1, pp. 18–37, 2010.

[8] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 4, no. 2, p. R1, 2007. DOI: 10.1088/1741-2560/4/2/R01.

[9] J. Zhang, Z. Yin, P. Cheng, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, 2020.

[10] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affective Computing*, vol. 7, no. 1, pp. 17–28, 2015.

[11] Y. Lu, W.-L. Zheng, B. Li, and B.-L. Lu, "Combining eye movements and EEG to enhance emotion recognition," in *Proc. 24th Int. Joint Conf. Artificial Intelligence*, 2015.

[12] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *Proc. IEEE Winter Conf. Applications of Computer Vision*, pp. 1–9, 2016.

[13] C. Mühl, B. Allison, A. Nijholt, and G. Chanel, "A survey of affective brain computer interfaces: Principles, state-of-the-art, and challenges," *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 66–84, 2014.

[14] D. H. Lee and A. K. Anderson, "Reading what the mind thinks from how the eye sees," *Psychological Science*, vol. 28, no. 4, pp. 494–503, 2017.

[15] M. Soleymani, M. Pantic, and T. Pun, "Multi-modal emotion recognition in response to videos," *IEEE Trans. Affective Computing*, vol. 3, no. 2, pp. 211–223, 2011.

[16] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016. DOI: 10.1186/s40537-016-0043-6.

[17] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[18] E. Principi, D. Rossetti, S. Squartini, and F. Piazza, "Unsupervised electric motor fault detection by using deep autoencoders," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 2, pp. 441–451, 2019.

[19] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multi-modal learning for affective analysis and retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2008–2020, 2015.

[20] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multi-modal deep learning," in *Proc. Int. Conf. Neural Information Processing*, pp. 521–529, Springer, 2016.

[21] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multi-modal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[22] X. Kang, F. Ren, and Y. Wu, "Exploring latent semantic information for textual emotion recognition in blog articles," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 1, pp. 204–216, 2017.

[23] J. Chen, X. Luo, and M. Zhou, "Hierarchical particle swarm optimization-incorporated latent factor analysis for large-scale incomplete matrices," *IEEE Trans. Big Data*, 2021. DOI: 10.1109/TBDATA.2021.3090905

[24] W.-L. Zheng and B.-L. Lu, "Personalizing EEG-based affective models with transfer learning," in *Proc. 25th Int. Joint Conf. Artificial Intelligence*, pp. 2732–2738, AAAI Press, 2016.

[25] P. Sinno Jialin, I. W. Tsang, J. T. Kwok, and Y. Qiang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[26] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[27] R. Collobert, F. Sinz, J. Weston, L. Bottou, and T. Joachims, "Large scale transductive SVMs," *Journal of Machine Learning Research*, vol. 7, no. 1, pp. 1687–1712, 2006.

[28] E. Sangineto, G. Zen, E. Ricci, and N. Sebe, "We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer," in *Proc. ACM Int. Conf. Multimedia*, pp. 357–366, 2014.

[29] X. Chai, Q. Wang, Y. Zhao, Y. Li, D. Liu, X. Liu, and O. Bai, "A fast, efficient domain adaptation technique for cross-domain electroencephalography (EEG)-based emotion recognition," *Sensors*, vol. 17, no. 5, p. 1014, 2017. DOI: 10.3390/s17051014.

[30] Y. Li, W. Zheng, Y. Zong, Z. Cui, T. Zhang, and X. Zhou, "A bihemisphere domain adversarial neural network model for EEG emotion recognition," *IEEE Trans. Affective Computing*, vol. 12, no. 2, pp. 494–504, 2021.

[31] J. Li, S. Qiu, C. Du, Y. Wang, and H. He, "Domain adaptation for EEG emotion recognition based on latent representation similarity," *IEEE Trans. Cognitive and Developmental Systems*, vol. 12, no. 2, pp. 344–353, 2020.

[32] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv: 1312.6114, 2013.

[33] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola, "Style transfer from nonparallel text by cross-alignment," in *Advances in Neural Information Processing Systems*, pp. 6830–6841, 2017.

[34] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems*, pp. 700–708, 2017.

[35] M. Shang, X. Luo, Z. Liu, J. Chen, Y. Yuan, and M. Zhou, "Randomized latent factor model for high-dimensional and sparse matrices from industrial applications," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 1, pp. 131–141, 2018.

[36] X. Luo, Z. Liu, S. Li, M. Shang, and Z. Wang, "A fast non-negative latent factor model based on generalized momentum method," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 610–620, 2021.

[37] D. Wu, Y. He, X. Luo, and M. Zhou, "A latent factor analysis-based approach to online sparse streaming feature selection," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, 2021. DOI: 10.1109/TSMC.2021.3096065

[38] M. Wu and N. Goodman, "Multi-modal generative models for scalable weakly-supervised learning," in *Advances in Neural Information Processing Systems*, pp. 5575–5585, 2018.

[39] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[40] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, and A. Cichocki, "Emotionmeter: A multi-modal framework for recognizing human emotions," *IEEE Trans. Cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2018.

[41] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affective Computing*, vol. 10, no. 3, pp. 417–429, 2019.

[42] M. S. AL-Quraishi, I. Elamvazuthi, T. B. Tang, A.-Q. Muhammad, S. Parasuraman, and A. Borboni, "Multi-modal fusion approach based on

EEG and EMG signals for lower limb movement recognition," *IEEE Sensors Journal*, vol. 21, no. 24, pp. 27640–37650, 2021.

[43] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE Int. Conf. Computer Vision*, pp. 4068–4076, 2015.

[44] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Machine Learning*, pp. 97–105, 2015.

[45] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[46] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. 34th Int. Conf. Machine Learning-Volume 70*, pp. 2208–2217, JMLR. org, 2017.

[47] P. Haeusser, T. Frerix, A. Mordvintsev, and D. Cremers, "Associative domain adaptation," in *Proc. IEEE Int. Conf. Computer Vision*, pp. 2765–2773, 2017.

[48] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018.

[49] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in Neural Information Processing Systems*, pp. 469–477, 2016.

[50] C. Yu, J. Wang, Y. Chen, and M. Huang, "Transfer learning with dynamic adversarial adaptation network," in *Proc. IEEE Int. Conf. Data Mining*, pp. 778–786, 2019.

[51] R. Horlings, D. Datcu, and L. J. Rothkrantz, "Emotion recognition using brain activity," in *Proc. 9th Int. Conf. Computer Systems and Technologies and Workshop for Ph.D. Students in Computing*, 2008.

[52] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[53] M. Liu, D. Fan, X. Zhang, and X. Gong, "Human emotion recognition based on galvanic skin response signal feature selection and SVM," in *Proc. Int. Conf. Smart City and Systems Engineering*, pp. 157–160, 2016.

Yixin Wang received the Ph.D. degree with the Institute of Automation, Chinese Academy of Sciences (CASIA). She is currently an Engineer of Beijing Institute of Control and Electronic Technology. Her research interests include affective computing, machine learning, domain adaptation and brain-computer interface (BCI).

Shuang Qiu received the Ph.D. degree in biomedical engineering from Tianjin University. She is currently an Associate Professor with CASIA. Her research interests concern biosignal processing, machine learning and rehabilitation engineering, especially in brain-computer interface (BCI).

Dan Li received the Ph.D. degree in Institute of Automation, Chinese Academy of Sciences. She is currently a Lecturer of School of Mathematics and Information Sciences, Yantai University. Her current research interests include multi-modal deep learning, computational neuro-science, and applications in computer vision.

Changde Du received the Ph.D. degree in Institute of Automation, Chinese Academy of Sciences (CASIA). He is currently an Assistant Professor with CASIA. His current research interests include machine learning, Bayesian statistics and brain-inspired intelligence.

Bao-Liang Lu received the Dr. Eng. degree in electrical engineering from Kyoto University, Japan. He is currently a Full Professor at the Department of Computer Science and Engineering, Shanghai Jiao Tong University. His research interests include brain-like computing, neural networks, emotion AI, and affective brain-computer interface.

Huiguang He received the Ph.D. degree (with honor) in pattern recognition and intelligent systems from Institute of Automation, Chinese Academy of Sciences (CASIA). He is currently a Full Professor with CASIA. His research interests include pattern recognition, medical image processing, and brain-computer interface (BCI).