Multi-level Disentangling Network for Cross-Subject Emotion Recognition Based on Multimodal Physiological Signals

Ziyu Jia, **Fengming Zhao**, **Yuzhe Guo**, **Hairong Chen** and **Tianzi Jiang***
Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{jia.ziyu, zhao.fengming, yuzhe.guo.cs, hairong.chen.david}@outlook.com,
tianzi.jiang.iacas@gmail.com

Abstract

Emotion recognition based on multimodal physiological signals is attracting more and more attention. However, how to deal with the consistency and heterogeneity of multimodal physiological signals, as well as individual differences across subjects, pose two significant challenges. In this paper, we propose a Multi-level Disentangling Network named MDNet for cross-subject emotion recognition based on multimodal physiological signals. Specifically, MDNet consists of a modality-level disentangling module and a subject-level disentangling module. The modality-level disentangling module projects multimodal physiological signals into modality-invariant subspace and modality-specific subspace, capturing modality-invariant features and modality-specific features. The subject-level disentangling module separates subject-shared features and subject-private features among different subjects from multimodal data, which facilitates cross-subject emotion recognition. Experiments on two multimodal emotion datasets demonstrate that MDNet outperforms other state-of-the-art baselines.

1 Introduction

Emotion recognition plays a pivotal role in affective computing [Cowie *et al.*, 2001]. In recent years, researchers have typically employed both non-physiological signals and physiological signals for emotion recognition [Shen *et al.*, 2019]. Non-physiological signals such as text, video, and audio are easily influenced by subjective factors and can be easily masked [Deng *et al.*, 2018]. Hence, it is not guaranteed to reflect human emotional states accurately. In contrast, some crucial physiological signals such as electroencephalogram (EEG), electromyogram (EMG), and electrooculogram (EOG) can objectively represent the true emotional state of the human body. As a result, emotion recognition based on physiological signals has gradually become a hot research topic [Ning *et al.*, 2023].

*Corresponding Author

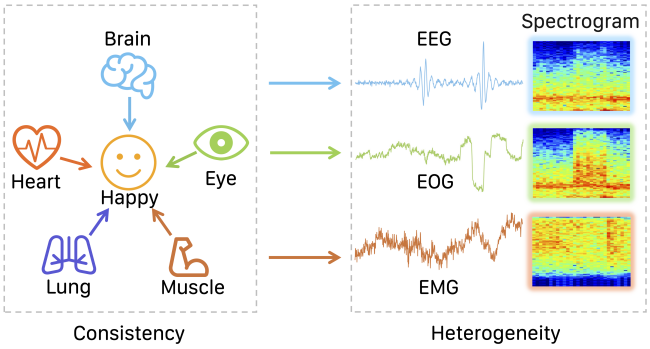


Figure 1: The consistency and heterogeneity in physiological signals. Consistency refers to the uniform patterns of physiological signals in representing the same emotional states across different modalities. Heterogeneity refers to the signals' distinct distributions in the temporal and spectral domains across different modalities.

Compared to unimodal physiological signals, multimodal physiological signals can represent emotional states comprehensively. Therefore, some emotion recognition methods based on multimodal physiological signals have achieved state-of-the-art performance [Jia *et al.*, 2021; Zhang *et al.*, 2020a; Abdullah *et al.*, 2021; Zhang *et al.*, 2022]. However, there are still two major challenges for emotion recognition based on multimodal physiological signals:

1) *How to model the consistency and heterogeneity of multimodal physiological signals simultaneously.* The consistency and heterogeneity are two important characteristics of multimodal physiological signals [Zhang *et al.*, 2020a; Khan *et al.*, 2023]. Specifically, the consistency of multimodal physiological signals refers to the interrelatedness of different physiological signals for the same physiological activity. In emotion recognition, different physiological signals can reflect the same emotional state, providing comprehensive information. As shown in Figure 1, both EEG and EOG reflect a happy emotional state. The heterogeneity of multimodal physiological signals refers to distinct features between different physiological signals [Jia *et al.*, 2023]. As shown in Figure 1, EEG and EOG exhibit notable differences in temporal and spectral domain distributions. However, most existing methods model either the consistency or heterogeneity of multimodal physiological signals individually. To

model consistency, joint models project different physiological signals into a common subspace for feature extraction [Lu *et al.*, 2015; Yin *et al.*, 2017; Xu *et al.*, 2021; Jia *et al.*, 2021], but ignore the heterogeneity between different physiological signals. To model heterogeneity, coordination models typically use different feature extractors for each physiological signal to capture their distinct features [Qiu *et al.*, 2018; Ma *et al.*, 2019b; Zhang *et al.*, 2020b; Jia *et al.*, 2022b; Lai, 2004], yet overlook consistency. Thus, how to model the consistency and heterogeneity of multimodal physiological signals simultaneously still remains a challenge.

2) *How to model individual differences across subjects in emotion recognition.* Due to structural and functional differences between subjects, physiological signals are different even though the subjects are in the same emotion state [Liu *et al.*, 2024; Zhang *et al.*, 2018]. This leads to poor performance in cross-subject emotion recognition [Fdez *et al.*, 2021]. Some researchers attempt to use domain adaptation to improve model performance in cross-subject emotion recognition [Zhao *et al.*, 2021; Li *et al.*, 2019a]. Although they can achieve excellent performance, Domain adaptation requires collecting extensive data from new subjects to customize the model in practical application, making the practical application of emotion recognition models inconvenient. Therefore, how to model individual differences across subjects in emotion recognition still remains a challenge.

To address these challenges, we propose a Multi-level Disentangling Network named MDNet for cross-subject emotion recognition based on multimodal physiological signals. MDNet consists of a modality-level disentangling module and a subject-level disentangling module. The modality-level disentangling module includes a modality-invariant encoder and several modality-specific encoders. The subject-level disentangling module comprises a subject-shared encoder and several subject-private encoders.

Overall, the main contributions of our work are summarized as follows:

- We propose a modality-level disentangling module, which captures modality-invariant features and modality-specific features. Both the consistency and heterogeneity of multimodal physiological signals are integrated into a unified framework.
- We develop a subject-level disentangling module, which separates both subject-shared features and subject-private features, in order to address the individual differences across subjects.
- Experiments on two multimodal emotion datasets show that MDNet achieves state-of-the-art performance.

2 Related Works

2.1 Multimodal Emotion Recognition

Multimodal emotion recognition models can mainly be classified into joint models and coordination models. Joint models extract features of multimodal physiological signals in the same subspace to model the consistency of multimodal signals [Xu *et al.*, 2021; Jia *et al.*, 2021; Lu *et al.*, 2015; Liu *et al.*, 2023; Yin *et al.*, 2017]. Xu *et al.* [2021] treat multimodal physiological signals as a multi-dimensional tensor. They integrate the multi-scale characteristics by fusing multi-core information. Jia *et al.* [2021] propose a two-stream heterogeneous graph recurrent neural network, which achieves multimodal feature fusion of temporal, spectral, and spatial features. Coordination models, on the other hand, extract features of multimodal physiological signals in separate subspaces and apply certain correlation constraints to these features to model the heterogeneity of multimodal signals [Qiu *et al.*, 2018; Jia *et al.*, 2022a; Zhang *et al.*, 2020b]. Qiu *et al.* [2018] utilize deep canonical correlation analysis (DCCA) for emotion recognition, which can learn separate representations for each modality in a non-linear way. Ma *et al.* [2019b] design a multimodal residual LSTM network for emotion recognition, projecting different modalities into separate LSTM branches to extract multimodal features.

Although joint models and coordination models can achieve high accuracy, they do not simultaneously model the consistency and heterogeneity of multimodal physiological signals. Our model combines the advantages of both joint models and coordination models, modeling the two characteristics within the same framework.

2.2 Cross-Subject Emotion Recognition

Individual differences exist in inter-subject physiological signals [Jia *et al.*, 2022b]. The variations of physiological signals from different subjects limits the model's performance for cross-subject emotion recognition. To address this challenge, researchers have proposed a series of methods to enhance emotion recognition models' generalizability [Ghifary *et al.*, 2016; Ma *et al.*, 2019a; Ganin *et al.*, 2016; Li *et al.*, 2018]. Li *et al.* [2019b] use source selection and style transformation mapping to reduce the domain differences between target and source, facilitating the acquisition of common features in EEG. Zhao *et al.* [2021] develop a plug-and-play domain adaptation method. This method customizes the model by inputting EEG from target subjects to enhance the model's specificity, which makes emotion recognition more generalizable and practicable as well. Li *et al.* [2019a] improve the model's generalizability by minimizing the classification error on the source while making the source and the target similar in latent representations. Li *et al.* [2021] propose a Transferable Attention Neural Network (TANN), which distinguishes the contribution of different samples for emotion recognition and uses local and global attention mechanisms to highlight the discriminative EEG data adaptively.

The methods above can effectively deal with individual differences across subjects and improve the accuracy of models in cross-subject emotion recognition. However, these methods only consider the individual differences in unimodal EEG and do not take a further concern of multimodal signals for cross-subject emotion recognition.

3 Multi-level Disentangled Network

Our goal is to achieve high performance in cross-subject emotion recognition by modeling key features from multimodal data. As shown in Figure 2, MDNet consists of a modality-

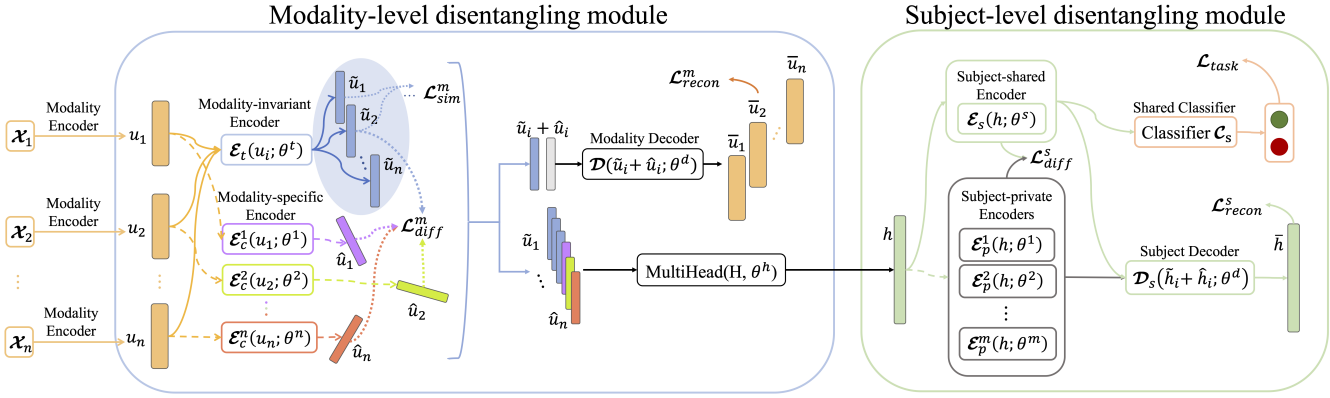


Figure 2: The overall structure of the proposed MDNet. MDNet takes the multimodal data $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_n$ as inputs. Modality encoders are used to initially extract the features of each modality. Modality-invariant encoder and modality-specific encoder project the initial features into the modality-invariant subspace and the modality-specific subspace, capturing modality-invariant features and modality-specific features. A multi-head self-attention mechanism is applied to obtain the subject-level representation. Then, subject-shared encoder and subject-private encoder separate subject-shared features and subject-private features. Finally, subject-shared features are used for cross-subject emotion recognition.

level disentangling module and a subject-level disentangling module. Specifically, the modality-level disentangling module captures modality-invariant features and modality-specific features from multimodal physiological signals. A self-attention mechanism is then used to fuse these features to obtain subject-level representation, reducing the information redundancy when fusing multimodal features. Afterwards, subject-level disentangling module separates subject-shared features and subject-private features from the subject-level representation. In order to deal with the individual differences across subjects, only subject-shared features are used for classification, enhancing the accuracy of emotion recognition.

3.1 Modality-level Disentangling Module

To model key features from multimodal data, modality-invariant features and modality-specific features are captured by projecting each modality into modality-invariant subspace and modality-specific subspace, respectively. In the modality-invariant subspace, modality-invariant features learn the multimodal consistency. In the modality-specific subspace, modality-specific features learn the multimodal heterogeneity. Afterwards, modality-invariant features and modality-specific features for each modality are obtained. A multi-head self-attention mechanism is applied to these features, which allows each feature to induce potential information from other features that are synergistic for emotion recognition.

Specifically, for each modality, a modality encoder is used to initially extract the modality features. The encoding process can be formally expressed as:

$$u_i = \text{ModalityEncoder}_i(\mathcal{X}_i) \quad (i = 1, 2, \cdots, n) \tag{1}$$

where $\mathcal{X}_i \in \mathbb{R}^{C_i \times L_i}$ represents the input of the i-th modality, C_i denotes the number of channels in the i-th modality, L_i denotes the number of sampling points in each channel of the i-th modality, and n represents the number of input modalities. Then, we obtain the initial features $u_i (i = 1, 2, \cdots, n) \in \mathbb{R}^d$ extracted from all modalities, where d represents the dimension of each feature.

The modality-level disentangling module consists of a modality-invariant encoder and n modality-specific encoders. The modality-invariant encoder re-encodes the initial features $u_i (i = 1, 2, \cdots, n)$ of the modalities, projecting them to the modality-invariant subspace. The encoding process of the modality-invariant encoder can be formally expressed as:

$$\tilde{u}_i = \mathcal{E}_t(u_i, \theta_t) \quad (i = 1, 2, \cdots, n) \tag{2}$$

where $\mathcal{E}_t$ represents the modality-invariant encoder composed of fully connected layers, θ_t denotes the parameters of the modality-invariant encoder, and $\tilde{u}_i$ represents the modality-invariant features of i-th modality. For all modalities, the modality-invariant encoder $\mathcal{E}_t$ shares the parameters θ_t, enabling the re-encoded features to learn the multimodal consistency. Hence, we obtain the modality-invariant features $\tilde{u}_i (i = 1, 2, \cdots, n) \in \mathbb{R}^d$ extracted from the initial features of all modalities.

The modality-invariant features of different modalities in the modality-invariant subspace have to be as similar as possible. We use Central Moment Discrepancy (CMD) to measure the similarity between modality-invariant features. The CMD decreases as the distribution of modality-invariant features become more similar. The modality similarity loss $\mathcal{L}^m_{sim}$ is defined as:

$$\mathcal{L}^m_{sim} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \text{CMD}(\tilde{u}_i, \tilde{u}_j) \tag{3}$$

For i-th modality, the i-th modality-specific encoder re-encodes the initial features u_i, projecting them to a modality-

specific subspace. The encoding process of the modality-specific encoder can be formally expressed as:

$$\hat{u}_i = \mathcal{E}_c^i(u_i, \theta^i) \quad (i = 1, 2, \cdots, n) \tag{4}$$

where $\mathcal{E}_c$ represents the modality-specific encoder composed of fully connected layers, θ^i denotes the parameters of the modality-specific encoder, and $\hat{u}_i$ represents the modality-specific features of i-th modality. For each modality, the modality-specific encoder $\mathcal{E}_c^i$ has different parameters θ^i, enabling the re-encoded features to learn the individual differences across subjects. Consequently, we obtain the modality-specific features $\hat{u}_i (i = 1, 2, \cdots, n) \in \mathbb{R}^d$ extracted from the initial features of all modalities.

Modality-invariant features and modality-specific features represent different aspects of a same modality. Orthogonality constraint is used to maximize the difference between modality-invariant features and modality-specific features. Similarly, the differences between modality-specific features of different modalities are maximized by the same orthogonality constraint. The modality difference loss $\mathcal{L}_{diff}^m$ is defined as:

$$\mathcal{L}_{diff}^m = \sum_{i=1}^{n} ||\tilde{u}_i^T \hat{u}_i||_F^2 + \sum_{i=1}^{n} \sum_{j=i+1}^{n} ||\hat{u}_i^T \hat{u}_j||_F^2 \tag{5}$$

where $|| \cdot ||_F^2$ represents the squared Frobenius norm.

During the re-encoding process, some information contained in the initial features is inevitably lost. Mean squared error is used to minimize the distortion of the re-encoded features. To measure the distortion of re-encoded features, we use a decoder composed of fully connected layers to reconstruct the initial features $\bar{u}_i = \mathcal{D}(\tilde{u}_i + \hat{u}_i, \theta^d)$. The modality reconstruction loss $\mathcal{L}_{recon}^m$ is defined as:

$$\mathcal{L}_{recon}^m = \frac{1}{n} \sum_{i=1}^{n} \frac{||u_i - \bar{u}_i||_2^2}{d} \tag{6}$$

where $|| \cdot ||_2^2$ represents the squared L_2-norm.

After obtaining modality-invariant features and modality-specific features of all modalities, we concatenate these features to obtain the modality-level representation $H = (\tilde{u}_1 || \hat{u}_1 || \tilde{u}_2 || \hat{u}_2 \cdots \tilde{u}_n || \hat{u}_n) \in \mathbb{R}^{2n \times d}$, where $||$ denotes concatenation. To induce interaction between features, we introduce the attention mechanism [Vaswani *et al.*, 2017], which can be formally expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{7}$$

where Q, K, V stand for query, key, and value, respectively, d_k is the dimension of the features. Multi-head attention consists of multiple parallel attention mechanisms. The head_i is defined as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{8}$$

where W_i^Q, W_i^K, W_i^V are the parameters of the i th attention mechanism.

In order to obtain the subject-level representation h, we apply the multi-head self-attention mechanism. Formally, the subject-level representation is computed as:

$$h = \text{MultiHead}(H, \theta^h) = (\text{head}_1 || \text{head}_2 || \cdots || \text{head}_n) W^O \tag{9}$$

where $\theta^h = \{W^Q, W^K, W^V, W^O\}$, the definition of head_i is as given above, and $||$ denotes concatenation.

3.2 Subject-level Disentangling Module

After obtaining the subject-level representation, we use the subject-level disentangling module to separate subject-shared features and subject-private features. Subject-shared features represent the similar expression of different subjects for the same emotion. On the other hand, subject-private features represent the personalized expression of different subjects for the same emotion.

The subject-level disentangling module consists of a subject-shared encoder and m subject-private encoders, where m is the subject's number in the training data. The subject-shared encoder re-encodes the subject-level representation to capture the subject-shared features, which can be formally expressed as:

$$\tilde{h} = \mathcal{E}_s(h, \theta_s) \tag{10}$$

where $\mathcal{E}_s$ represents the subject-shared encoder composed of fully connected layers, θ_s denotes the parameters of the subject-shared encoder, and $\tilde{h}$ represents the subject-shared features. For all subjects, the subject-shared encoder $\mathcal{E}_s$ shares parameters θ_s, enabling the re-encoded features to learn the similarities among different subjects. As a result, we obtain the subject-shared features $\tilde{h} \in \mathbb{R}^d$ extracted from the subject-level representation.

For each subject, a subject-private encoder re-encodes their subject-level representation to capture the subject-private features, which can be formally expressed as:

$$\hat{h} = \mathcal{E}_p^i(h, \theta_p^i) \tag{11}$$

where $\mathcal{E}_p^i$ represents the subject-private encoder composed of fully connected layers, θ_p^i denotes the parameters of the i-th subject-private encoder, and $\hat{h}$ represents the subject-private features. For each subject, the subject-private encoder $\mathcal{E}_p^i$ has different parameters θ_p^i, enabling the re-encoded features to learn the uniqueness of different subjects. As a result, we obtain the subject-private features $\hat{h} \in \mathbb{R}^d$ extracted from the subject-level representation.

Subject-shared features and subject-private features reflect different aspects of human emotional characteristics. In order to capture different aspects of the subject-level representation, an orthogonality constraint is applied to maximize the differences between subject-shared features and subject-private features. The subject difference loss $\mathcal{L}_{diff}^s$ is defined as:

$$\mathcal{L}_{diff}^s = ||\tilde{h}^T \hat{h}||_F^2 \tag{12}$$

where $|| \cdot ||_F^2$ represents the squared Frobenius norm.

We impose constraint using mean squared error which can minimize the distortion of the re-encoded features. To measure the distortion of features after the separation process, we use a decoder composed of fully connected layers to reconstruct the subject-level representation $\bar{h} = D(\tilde{h} + \hat{h}, \theta^d)$. The individual reconstruction loss $\mathcal{L}_{recon}^s$ is defined as:

$$\mathcal{L}_{recon}^{s} = \frac{||h - \bar{h}||_2^2}{d} \tag{13}$$

where $||\cdot||_2^2$ represents the squared L_2-norm, d is the dimension of h.

Finally, subject-shared features $\tilde{h}$ are fed into a classifier $\mathcal{C}_s$. The result can be formally computed as:

$$\hat{y} = \mathcal{C}_s(\tilde{h}) \tag{14}$$

where $\hat{y}$ is the emotion classification result.

The cross-entropy is used as the task loss $\mathcal{L}_{task}$, which is defined as:

$$\mathcal{L}_{task} = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i) \tag{15}$$

where y_i represents the label of the task, n is the number of samples in a batch.

3.3 Learning Loss

A joint loss function is employed to constrain the model, formally, the entire loss function can be represented as:

$$\mathcal{L} = \mathcal{L}_{task} + \alpha\mathcal{L}_{sim}^{m} + \beta\mathcal{L}_{diff}^{m} + \gamma\mathcal{L}_{recon}^{m} + \delta\mathcal{L}_{diff}^{s} + \epsilon\mathcal{L}_{recon}^{s} \tag{16}$$

where $\alpha, \beta, \gamma, \delta, \epsilon$ are the weights of the loss terms. The task loss $\mathcal{L}_{task}$ is used to ensure accurate classification. The Modality similarity loss $\mathcal{L}_{sim}^{m}$ ensures that the modality-invariant features remain similar. The modality difference loss $\mathcal{L}_{diff}^{m}$ guarantees the distinction between modality-specific features of different modalities and also between modality-specific features and modality-invariant features within the same modality. The modality reconstruction loss $\mathcal{L}_{recon}^{m}$ minimizes distortion of modality-specific features during the re-encoding process. The subject difference loss $\mathcal{L}_{diff}^{s}$ ensures a clear separation between subject-shared features and subject-private features. The subject reconstruction loss $\mathcal{L}_{recon}^{s}$ reduces feature distortion of subject-shared features and subject-private features during the separations.

4 Experiments

4.1 Datasets

We evaluate our MDNet on two public multimodal datasets: DEAP [Koelstra *et al.*, 2011] and MAHNOB-HCI [Soleymani *et al.*, 2011].

The DEAP dataset is a collection comprising 40 music videos, each with a duration of one minute. The dataset is created based on the responses of 32 subjects who conducted online self-assessments. During the experiment, the 32 subjects watch the selected videos while their Electroencephalogram (EEG) is recorded using a BioSemi EEG cap with 32 channels, conforming to the international 10-20 system standards. Additionally, 8 channels are dedicated to collecting peripheral physiological signals (PPS), including respiration rate, Electromyogram (EMG), and Electrooculogram (EOG). Each subject rates the videos based on arousal, valence, liking, dominance, and familiarity. In the preprocessing stage, all signals are downsampled to 128 Hz and EEG is passed through a bandpass filter from 4 Hz to 45 Hz. Electrooculogram (EOG) artifacts are removed. For the DEAP dataset, the arousal and valence ratings range from 1 to 9. Thus, ratings ≥ 5 are labeled as positive for high arousal or valence, and those < 5 as negative for low arousal or valence.

The MAHNOB-HCI dataset includes data from 30 subjects, each watching 20 video clips while recording physiological data for 20 trials. The length of these video clips ranges from 34.9 seconds to 117 seconds (average 81.4 seconds, standard deviation 22.5 seconds). Subjects provide subjective feedback using a score range from 1 to 9. This dataset includes EEG signals from 32 channels and 6 channels of peripheral physiological signals (PPS). Their EEG data are recorded using the Biosemi Active II system with 32 Ag/AgCl electrodes at a sampling rate of 512Hz. The peripheral physiological signals include Electrocardiogram (ECG), Galvanic Skin Response (GSR), respiration, and body temperature. We downsample all signals to 128Hz. In model evaluation, a fixed 5-point threshold is used to discretize the subjective feedback into binaries for two emotional dimensions, valence and arousal, where values ≥ 5 indicate positive, and < 5 as negative. However, data from subjects 12, 15, and 26 are missing, and thus excluded. Therefore, data from 27 out of the 30 subjects are used, consistent with existing models like TSecption [Ding *et al.*, 2022] and HetEmotionNet [Jia *et al.*, 2021].

4.2 Experiment Settings and Implementation

Multiple modalities are employed across two datasets: EEG, EOG, and EMG for the DEAP dataset, while EEG, ECG, and GSR for the MAHNOB-HCI dataset. To avoid data leakage during the evaluation process, we adopt a leave-one-subject-out cross-validation protocol on both datasets. Specifically, the training and test data are from different subjects, ensuring no data leakage. For example, in DEAP dataset, we use data from 31 subjects for training and the remaining subject's data for testing, repeating the validation process until each subject's data has been used as test data once. The entire experiment is conducted with 32 validations, and the average of all obtained test results is calculated as the final cross-validation performance.

To ensure a fair comparison, we apply the same treatment to our MDNet and all baseline models. Specifically, we set a 6s window with a 1s overlap. Considering the potential imbalance in label distribution in the original datasets, we resample and balance the training set labels. This helps avoid biases in the model during training and enhances its generalization ability, preventing imbalances in predicted labels.

We implement our MDNet based on the Pytorch framework. Our model is trained by Adam optimizer with a learning rate lr = 0.0003. The batch size is 64. The weights are α=0.002, β=0.1, γ=0.04, δ=0.5, and ϵ=0.05.

4.3 Baseline Methods

We compare our MDNet with eight state-of-the-art models, including SVM [Chatterjee and Bandyopadhyay, 2016], DGCNN [Song *et al.*, 2018], EEGNet [Lawhern *et al.*, 2018], ACRNN [Tao *et al.*, 2020], SST-EmotionNet [Jia *et al.*,

	DEAP		MAHNOB-HCI	
	Arousal	**Valence**	**Arousal**	**Valence**
SVM	0.518	0.516	0.472	0.525
DGCNN	0.598	0.594	0.622	0.578
EEGNet	0.617	0.590	0.645	0.605
ACRNN	0.635	0.595	0.586	0.623
SST-EmotionNet	0.597	0.565	0.606	0.590
HetEmotionNet	0.621	0.593	0.627	0.621
TSception	0.621	0.550	0.550	0.618
LGGNet	0.647	0.594	0.597	0.592
MDNet	**0.653**	**0.615**	**0.663**	**0.660**

Table 1: Comparative analysis of accuracy across different methods on DEAP and MAHNOB-HCI datasets. The accuracy is measured in terms of arousal and valence for both DEAP and MAHNOB-HCI datasets.

2020], HetEmotionNet [Jia *et al.*, 2021], TSception [Ding *et al.*, 2022], LGGNet [Ding *et al.*, 2023].

4.4 Experiment Analysis

To validate the effectiveness of our MDNet, we compare it with eight baseline methods on DEAP dataset and MAHNOB-HCI dataset, as shown in Table 1. Specifically, traditional models like EEGNet primarily contribute to extracting effective features from EEG signals to improve the performance of emotion recognition. However, these models neglect the significance of multimodal signals and do not utilize the fusion of multiple modalities to improve performance. In contrast, HetEmotionNet effectively extracts discriminative features from multimodal signals. It can model the heterogeneous information of different modalities, making it perform better than EEG-based models. TSception is able to extract discriminative features from multiple modalities, which in turn improves model performance. However, these models overlook the individual differences across subjects in multimodal signals, which often reduces the classification accuracy. In comparison, our MDNet employs disentangled representation learning to utilize the consistency and heterogeneity of multimodal physiological signals and reduce the impact of individual differences across subjects. As a result, our MDNet yields the highest accuracy in cross-subject emotion recognition tasks, and outperforms all the baseline models on both datasets.

4.5 Ablation Studies

To validate the effectiveness of our model, we conduct ablation experiments on DEAP dataset. These experiments are categorized into three types:

1) validating the effectiveness of modality-level disentangling module and subject-level disentangling module;

2) assessing the effectiveness of multimodal fusion;

3) examining the effectiveness of the loss terms.

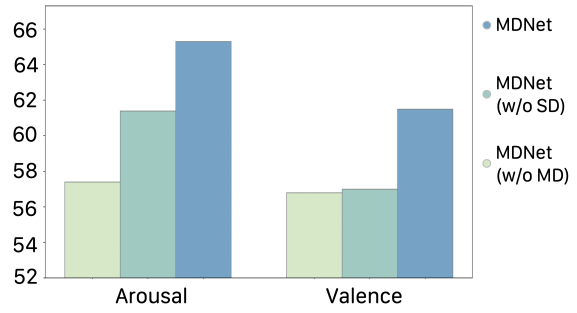


Figure 3: Ablation studies of disentangling modules on DEAP dataset. 'w/o SD' represents 'without the subject-level disentangling module', and 'w/o MD' represents 'without the modality-level disentangling module'.

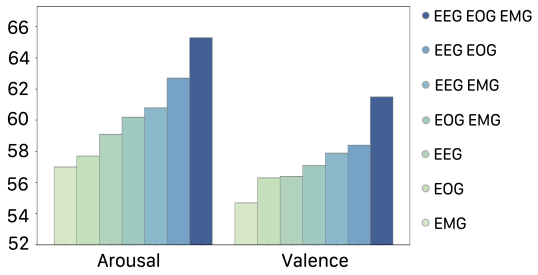


Figure 4: Ablation studies of multimodal fusion on DEAP dataset.

Ablation on disentangling modules: To validate the effectiveness of the disentangling modules, we conduct ablation experiments on modality-level disentangling module and subject-level disentangling module, as shown in Figure 3. The results indicate that the classification accuracy for valence and arousal significantly decreases in the absence of the modality-level disentangling module. Similarly, the accuracy drops without the subject-level disentangling module, particularly in classifying valence. This demonstrates that both disentangling modules improve the performance of our MDNet in cross-subject emotion recognition.

Ablation on multimodal fusion: To verify the superiority of multimodal fusion, we conduct ablation experiments by removing each modality (EEG, EMG, and EOG) individually from the original setting that uses three modalities. Figure 4 shows that removing any modality results in a decrease in classification accuracy. The removal of EEG signals has the most pronounced effect on accuracy reduction. This aligns with traditional cognition that EEG has the primary contribution to emotion recognition [Zhang *et al.*, 2020a].

It is also observed that using unimodal EEG yields better results than unimodal EOG, and using unimodal EOG performs better than unimodal EMG. This further validates the importance of EEG. The combination of multimodal data achieves better performance, which proves multimodal fusion has critical importance for emotion recognition.

Ablation on loss terms: To validate the effectiveness of the loss terms, we conduct ablation experiments, designing

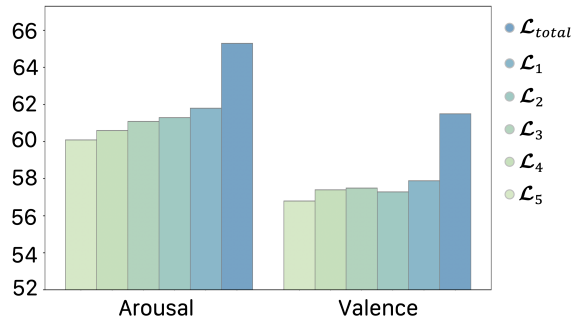
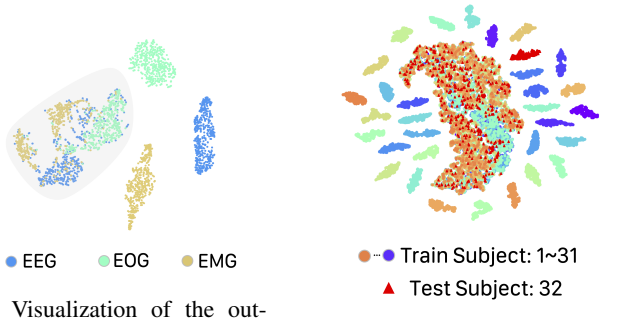


Figure 5: Ablation studies of loss terms on DEAP dataset.



(a) Visualization of the outputs of modality-invariant and modality-specific encoder for the 32nd subject (test subject).

(b) Visualization of the outputs of subject-shared encoder and subject-private encoders.

Figure 6: Visualizations of disentangling modules on DEAP dataset.

five variants:

- $\mathcal{L}_1 = \mathcal{L}_{total} - \mathcal{L}^s_{recon}$
- $\mathcal{L}_2 = \mathcal{L}_{total} - \mathcal{L}^s_{diff}$
- $\mathcal{L}_3 = \mathcal{L}_{total} - \mathcal{L}^m_{recon}$
- $\mathcal{L}_4 = \mathcal{L}_{total} - \mathcal{L}^m_{sim}$
- $\mathcal{L}_5 = \mathcal{L}_{total} - \mathcal{L}^m_{diff}$

To assess the importance of each loss term, we conduct ablation experiments by removing each loss term individually and retraining the model. Figure 5 indicates that the model performance is optimal when all loss terms are involved ($\mathcal{L}_{total}$). Removing any loss term leads to a decrease in accuracy. Observations on $\mathcal{L}_4$ and $\mathcal{L}_5$ reveal that the model is most sensitive to $\mathcal{L}^m_{diff}$ and $\mathcal{L}^m_{sim}$, which validate their importance in separating modality-specific features and modality-invariant features. Similarly, removing $\mathcal{L}^s_{diff}$ and $\mathcal{L}^s_{sim}$ also leads to a drop in classification accuracy, which verifies the vital role of $\mathcal{L}^s_{diff}$ and $\mathcal{L}^s_{sim}$ in separating subject-shared features and subject-private features. Overall, these results prove the importance of loss terms in maintaining and improving model performance and their necessity in the model optimization process.

4.6 Visualizations of Disentangling Modules

Validation of modality-level disentangling To explore the role of the modality-level disentangling module, we visualize the outputs of modality-invariant and modality-specific encoder. The t-SNE dimensionality reduction technique is utilized to transform high-dimensional data into lower-dimensional data. Additionally, a scatter plot visualization is applied to effectively present the outputs of the modality-invariant encoder and the modality-specific encoder which are shown in Figure 6a. In this figure, different colors represent different physiological signal modalities. The grey-shaded area covers the outputs of the modality-invariant encoder, while the outer area represents the outputs of the modality-specific encoder. This visualization reveals a distinction between the modality invariant representations and specific representations in the encoded data. This indicates that the modality-level disentangling module is effective in separating modality-invariant features from modality-specific features and in modeling the consistency and heterogeneity of multimodal data.

Validation of subject-level disentangling To validate the effectiveness of the subject-level disentangling module, we use the t-SNE dimensionality reduction and scatter plot visualization. Figure 6b shows the output feature distribution of the subject-shared encoder and subject-private encoders: the densely packed dots in the central area represent the shared emotional components, while the scattered dots around the periphery, separated from the central area, represent subject private components. Here, red triangles indicate data points from the test subject. This result demonstrates that the encoded outputs provide distinguishable shared and private features. The subject-shared encoder successfully extracts subject-shared features that are highly relevant to emotions, while the subject-private encoders effectively separate subject-private features.

5 Conclusion

In this paper, we propose a multi-level disentangling network for cross-subject emotion recognition based on multimodal physiological signals. To the best of our knowledge, it is the first to apply disentangled representation learning to simultaneously model consistency and heterogeneity in multimodal physiological signals, as well as the individual differences across subjects. The modality-level disentangling module captures modality-invariant features and modality-specific features to model the consistency and heterogeneity of multimodal physiological signals. The subject-level disentangling module separates subject-shared features and subject-private features, to model individual differences across subjects. Experimental results show that our model achieves state-of-the-art performance. In addition, ablation studies validate the effectiveness of the disentangling modules. To sum up, we propose a general framework, and we plan to further explore the application of MDNet in sleep stage classification and depression detection in the future.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant No. 62306317) and STI2030-Major Projects (Grant No. 2021ZD0200200) and Postdoctoral Fellowship Program of CPSF (Grant No. GZC20232992) and China Postdoctoral Science Foundation (Grant No. 2023M733738).

Contribution Statement

Fengming Zhao, Yuzhe Guo, and Hairong Chen have equal contributions to this paper.

References

[Abdullah *et al.*, 2021] Sharmeen M Saleem Abdullah Abdullah, Siddeeq Y Ameen Ameen, Mohammed AM Sadeeq, and Subhi Zeebaree. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(02):52–58, 2021.

[Chatterjee and Bandyopadhyay, 2016] Rajdeep Chatterjee and Tathagata Bandyopadhyay. Eeg based motor imagery classification using svm and mlp. In *2016 2nd International Conference on Computational Intelligence and Networks (CINE)*, pages 84–89. IEEE, 2016.

[Cowie *et al.*, 2001] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[Deng *et al.*, 2018] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller. Semisupervised autoencoders for speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1):31–43, 2018.

[Ding *et al.*, 2022] Yi Ding, Neethu Robinson, Su Zhang, Qiuhao Zeng, and Cuntai Guan. Tsception: Capturing temporal dynamics and spatial asymmetry from eeg for emotion recognition. *IEEE Transactions on Affective Computing*, 2022.

[Ding *et al.*, 2023] Yi Ding, Neethu Robinson, Chengxuan Tong, Qiuhao Zeng, and Cuntai Guan. Lggnet: Learning from local-global-graph representations for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Fdez *et al.*, 2021] Javier Fdez, Nicholas Guttenberg, Olaf Witkowski, and Antoine Pasquali. Cross-subject eeg-based emotion recognition through neural networks with stratified normalization. *Frontiers in neuroscience*, 15:626277, 2021.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

[Ghifary *et al.*, 2016] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1414–1430, 2016.

[Jia *et al.*, 2020] Ziyu Jia, Youfang Lin, Xiyang Cai, Haobin Chen, Haijun Gou, and Jing Wang. Sst-emotionnet: Spatial-spectral-temporal based attention 3d dense network for eeg emotion recognition. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2909–2917, 2020.

[Jia *et al.*, 2021] Ziyu Jia, Youfang Lin, Jing Wang, Zhiyang Feng, Xiangheng Xie, and Caijie Chen. Hetemotionnet: two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1047–1056, 2021.

[Jia *et al.*, 2022a] Ziyu Jia, Xiyang Cai, and Zehui Jiao. Multi-modal physiological signals based squeeze-and-excitation network with domain adversarial learning for sleep staging. *IEEE Sensors Journal*, 22(4):3464–3471, 2022.

[Jia *et al.*, 2022b] Ziyu Jia, Junyu Ji, Xinliang Zhou, and Yuhan Zhou. Hybrid spiking neural network for sleep electroencephalogram signals. *Science China Information Sciences*, 65(4):140403, 2022.

[Jia *et al.*, 2023] Ziyu Jia, Youfang Lin, Yuhan Zhou, Xiyang Cai, Peng Zheng, Qiang Li, and Jing Wang. Exploiting interactivity and heterogeneity for sleep stage classification via heterogeneous graph neural network. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Khan *et al.*, 2023] Pritam Khan, Priyesh Ranjan, and Sudhir Kumar. At2gru: A human emotion recognition model with mitigated device heterogeneity. *IEEE Transactions on Affective Computing*, 14(2):1520–1532, 2023.

[Koelstra *et al.*, 2011] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing*, 3(1):18–31, 2011.

[Lai, 2004] P Lai. Kernel and nonlinear canonical correlation analysis. *Int. J. Neural Systems*, 16(12):2639–2664, 2004.

[Lawhern *et al.*, 2018] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[Li *et al.*, 2018] Yang Li, Wenming Zheng, Yuan Zong, Zhen Cui, Tong Zhang, and Xiaoyan Zhou. A bi-hemisphere domain adversarial neural network model for eeg emotion recognition. *IEEE Transactions on Affective Computing*, 12(2):494–504, 2018.

[Li *et al.*, 2019a] Jinpeng Li, Shuang Qiu, Changde Du, Yixin Wang, and Huiguang He. Domain adaptation for eeg emotion recognition based on latent representation similarity. *IEEE Transactions on Cognitive and Developmental Systems*, 12(2):344–353, 2019.

[Li *et al.*, 2019b] Jinpeng Li, Shuang Qiu, Yuan-Yuan Shen, Cheng-Lin Liu, and Huiguang He. Multisource transfer learning for cross-subject eeg emotion recognition. *IEEE transactions on cybernetics*, 50(7):3281–3293, 2019.

[Li *et al.*, 2021] Yang Li, Boxun Fu, Fu Li, Guangming Shi, and Wenming Zheng. A novel transferability attention neural network model for eeg emotion recognition. *Neurocomputing*, 447:92–101, 2021.

[Liu *et al.*, 2023] Yucheng Liu, Ziyu Jia, and Haichao Wang. Emotionkd: A cross-modal knowledge distillation framework for emotion recognition based on physiological signals. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6122–6131, 2023.

[Liu *et al.*, 2024] Chenyu Liu, Xinliang Zhou, Zhengri Zhu, Liming Zhai, Ziyu Jia, and Yang Liu. Vbh-gnn: Variational bayesian heterogeneous graph neural networks for cross-subject emotion recognition. In *The Twelfth International Conference on Learning Representations*, 2024.

[Lu *et al.*, 2015] Yifei Lu, Wei-Long Zheng, Binbin Li, and Bao-Liang Lu. Combining eye movements and eeg to enhance emotion recognition. In *IJCAI*, volume 15, pages 1170–1176. Buenos Aires, 2015.

[Ma *et al.*, 2019a] Bo-Qun Ma, He Li, Wei-Long Zheng, and Bao-Liang Lu. Reducing the subject variability of eeg signals with adversarial domain generalization. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26*, pages 30–42. Springer, 2019.

[Ma *et al.*, 2019b] Jiaxin Ma, Hao Tang, Wei-Long Zheng, and Bao-Liang Lu. Emotion recognition using multimodal residual lstm network. In *Proceedings of the 27th ACM international conference on multimedia*, pages 176–183, 2019.

[Ning *et al.*, 2023] Xiaojun Ning, Jing Wang, Youfang Lin, Xiyang Cai, Haobin Chen, Haijun Gou, Xiaoli Li, and Ziyu Jia. Metaemotionnet: Spatial-spectral-temporal based attention 3d dense network with meta-learning for eeg emotion recognition. *IEEE Transactions on Instrumentation and Measurement*, 2023.

[Qiu *et al.*, 2018] Jie-Lin Qiu, Wei Liu, and Bao-Liang Lu. Multi-view emotion recognition using deep canonical correlation analysis. In *Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part V 25*, pages 221–231. Springer, 2018.

[Shen *et al.*, 2019] Jian Shen, Xiaowei Zhang, Gang Wang, Zhijie Ding, and Bin Hu. An improved empirical mode decomposition of electroencephalogram signals for depression detection. *IEEE transactions on affective computing*, 13(1):262–271, 2019.

[Soleymani *et al.*, 2011] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing*, 3(1):42–55, 2011.

[Song *et al.*, 2018] Tengfei Song, Wenming Zheng, Peng Song, and Zhen Cui. Eeg emotion recognition using dynamical graph convolutional neural networks. *IEEE Transactions on Affective Computing*, 11(3):532–541, 2018.

[Tao *et al.*, 2020] Wei Tao, Chang Li, Rencheng Song, Juan Cheng, Yu Liu, Feng Wan, and Xun Chen. Eeg-based emotion recognition via channel-wise attention and self attention. *IEEE Transactions on Affective Computing*, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Xu *et al.*, 2021] Hongyan Xu, Jiajia Tang, Jianhai Zhang, and Li Zhu. Emotion recognition using multi-core tensor learning and multimodal physiological signal. In *International Workshop on Human Brain and Artificial Intelligence*, pages 137–148. Springer, 2021.

[Yin *et al.*, 2017] Zhong Yin, Mengyuan Zhao, Yongxiong Wang, Jingdong Yang, and Jianhua Zhang. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer methods and programs in biomedicine*, 140:93–110, 2017.

[Zhang *et al.*, 2018] Lin Zhang, Harald Traue, and Dilana Hazer-Rau. Individual emotion recognition and subgroup analysis from psychophysiological signals. *Signal & Image Processing: An International Journal (SIPIJ) Vol*, 9, 2018.

[Zhang *et al.*, 2020a] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, 59:103–126, 2020.

[Zhang *et al.*, 2020b] Xiaowei Zhang, Jing Pan, Jian Shen, Zia Ud Din, Junlei Li, Dawei Lu, Manxi Wu, and Bin Hu. Fusing of electroencephalogram and eye movement with group sparse canonical correlation analysis for anxiety detection. *IEEE Transactions on Affective Computing*, 13(2):958–971, 2020.

[Zhang *et al.*, 2022] Yong Zhang, Cheng Cheng, and YiDie Zhang. Multimodal emotion recognition based on manifold learning and convolution neural network. *Multimedia Tools and Applications*, 81(23):33253–33268, 2022.

[Zhao *et al.*, 2021] Li-Ming Zhao, Xu Yan, and Bao-Liang Lu. Plug-and-play domain adaptation for cross-subject eeg-based emotion recognition. In *AAAI Conference on Artificial Intelligence*, 2021.