HeLo: Heterogeneous Multi-Modal Fusion with Label Correlation for Emotion Distribution Learning

Chuhang Zheng
College of Artificial Intelligence
Nanjing University of Aeronautics and Astronautics
Nanjing, China
h.zheng@nuaa.edu.cn

Chunwei Tian
School of Computer Science and Technology
Harbin Institute of Technology
Harbin, China
chunweitian@163.com

Jie Wen
School of Computer Science and Technology
Harbin Institute of Technology, Shenzhen
Shenzhen, China
jiewen_pr@126.com

Daoqiang Zhang
College of Artificial Intelligence
Nanjing University of Aeronautics and Astronautics
Nanjing, China
dqzhang@nuaa.edu.cn

Qi Zhu*
College of Artificial Intelligence
Nanjing University of Aeronautics and Astronautics
Nanjing, China
zhuqi@nuaa.edu.cn

Abstract

Multi-modal emotion recognition has garnered increasing attention as it plays a significant role in human-computer interaction (HCI) in recent years. Since different discrete emotions may exist at the same time, compared with single-class emotion recognition, emotion distribution learning (EDL) that identifies a mixture of basic emotions has gradually emerged as a trend. However, existing EDL methods face challenges in mining the heterogeneity among multiple modalities. Besides, rich semantic correlations across arbitrary basic emotions are not fully exploited. In this paper, we propose a multi-modal emotion distribution learning framework, named **HeLo**, aimed at fully exploring the heterogeneity and complementary information in multi-modal emotional data and label correlation within mixed basic emotions. Specifically, we first adopt cross-attention to effectively fuse the physiological data. Then, an optimal transport (OT)-based heterogeneity mining module is devised to mine the interaction and heterogeneity between the physiological and behavioral representations. To facilitate label correlation learning, we introduce a learnable label embedding optimized by correlation matrix alignment. Finally, the learnable label embeddings and label correlation matrices are integrated with the multi-modal representations through a novel label correlation-driven cross-attention mechanism for accurate emotion distribution learning. Experimental results on two publicly available datasets demonstrate the superiority of our proposed method in emotion distribution learning.

*Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754852

CCS Concepts

• **Human-centered computing** → **HCI design and evaluation methods**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Emotion Recognition, Emotion Distribution Learning, Label Correlation, Optimal Transport

ACM Reference Format:
Chuhang Zheng, Chunwei Tian, Jie Wen, Daoqiang Zhang, and Qi Zhu. 2025. HeLo: Heterogeneous Multi-Modal Fusion with Label Correlation for Emotion Distribution Learning. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746027.3754852

1 Introduction

Emotion recognition plays a pivotal role in enhancing human-computer interaction (HCI) [2, 19, 27], offering substantial benefits across various fields such as healthcare, education, and customer service [1]. Researchers have investigated various behavioral signals, including facial expressions [32] and voice signals [38], as well as physiological signals like electroencephalogram (EEG) [37], electrocardiogram (ECG) [12], and electromyogram (EMG) [5], to detect emotional states. While many studies have concentrated on identifying basic emotions using the valence-arousal (VA) [45] or valence-arousal-dominance (VAD) [18] model, there is an increasing awareness of the complexity and mixed nature of human emotions. Research indicates that individuals may simultaneously experience multiple emotions with varying intensities at the same time [35, 42]. However, most existing emotion recognition methods tend to focus on single-class models [14, 41, 44], which may not fully capture the varied, complex, and sometimes ambiguous nature of human emotions. Accurate identification of these complex emotional states is essential, as it reflects more realistic human experiences and interactions, underscoring the need for advanced emotion recognition systems that can interpret these nuances effectively.

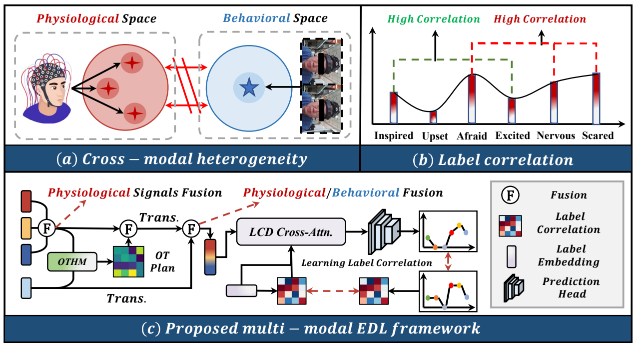


Figure 1: The challenges in emotion distribution learning and illustration of the proposed method. (a) cross-modal heterogeneity, (b) label correlation, (c) the schematic diagram of the proposed framework.

In contrast with multi-label emotion recognition [20, 39, 40], which typically classifies emotions into distinct and singular categories such as happiness, sadness, and anger, mixed emotion recognition aims to capture the complexity of coexisting emotions that do not easily fit into traditional labels. Furthermore, mixed emotion recognition evaluates the intensity of each emotion present, whereas multi-label approaches merely identify the presence or absence of each emotional state. Jia et al. [14] adopted an emotion distribution learning method that exploits label correlations locally for facial expression recognition. Zhang et al. [41] proposed a multi-task convolutional neural network for text emotion analysis. Existing single-modal methods struggle to fully utilize the complementary and correlative aspects between different modalities, overlooking the fact that emotions can be expressed through a combination of physiological responses and external manifestations.

Researchers have explored the efficacy of integrating multi-modal data to enhance emotion recognition performance [22]. Peng et al. [24] proposed CARAT to model fine-grained modality-to-label dependencies and exploit the modality complementarity by a shuffle-based aggregation strategy for multi-modal multi-label emotion recognition. Shu et al. [28] proposed an EDL convolutional neural network to predict the emotion distribution from the stacked physiological signals. Liu et al. [21] developed Emotion-Dict to learn a set of basic emotion elements in the latent feature space for emotion distribution learning. However, as the diagram shown in Figure 1 (a), these methods ignore the intrinsic correlation and heterogeneity among different modalities, resulting in a sub-optimal recognition performance. What's more, they cannot fully exploit the label correlation across different basic emotions. As can be seen from Figure 1 (b), "afraid" and "scared" may be highly correlated. Leveraging the label correlations as a guidance can enhance the performance of emotion distribution learning.

To address the above limitations, in this paper, we propose a multi-modal emotion distribution learning framework, named HeLo, as depicted in Figure 1 (c), which effectively mines the cross-modal heterogeneity within physiological and behavioral data, facilitating the learning of the label correlation to guide emotion distribution learning. In brief, the main contributions of our proposed model can be summarized as follows:

- Due to differences in heterogeneity across modalities, we first adopt a cross-attention mechanism to fuse the physiological data. Then, an optimal transport (OT)-based heterogeneity mining module is devised to effectively fuse the physiological and behavioral representations.
- For the learning of label correlation, we introduce a learnable label embedding, which is constrained by its learnable label correlation and ground-truth label correlation. Furthermore, the learnable label embeddings and label correlation are integrated through a novel label correlation-driven cross-attention mechanism.
- Extensive experiments on two multi-modal emotion datasets demonstrate that our proposed model outperforms state-of-the-art models in both subject-dependent and subject-independent scenarios.

2 Related Work

2.1 Label Distribution Learning

In order to address the problem that both the labels and their intensities are necessarily needed, Geng et al. systematically proposed label distribution learning [7]. According to the proposed LDL strategies, traditional LDL methods can be divided into three categories, i.e., Problem Transformation (PT) [8, 10], Algorithm Adaptation (AA) [9, 11], and Specialized Algorithms (SA) [7, 11]. Recently, Wen et al. proposed the learning objectives and evaluation metrics for label distribution learning, namely Cumulative Absolute Distance (CAD), Quadratic Form Distance (QFD), and Cumulative Jensen-Shannon divergence (CJS). LDL-LRR [13] adopted a novel ranking loss function for label distribution characteristics and combined it with KL-divergence as the loss term of the objective function. Gao et al. [6] proposed deep label distribution learning (DLDL), which utilizes label ambiguity in both feature learning and classifier learning to establish an end-to-end LDL framework.

To improve the performance of LDL, researchers have made attempts to exploit the label correlations. Jia et al. [14] employed a local low-rank structure to capture the label correlation implicitly to enhance facial expression recognition. Kou et al. [17] proposed TLRLDL, which leverages low-rank label correlations specifically within an auxiliary multi-label learning framework. Different from these methods, we develop an end-to-end EDL framework, which makes full use of the complementarity between physiology and behavior signals, and alleviates the negative impact of discrete emotional ambiguity on EDL through label correlation learning.

2.2 Muli-modal Emotion Recognition

As multi-modal data provide a more comprehensive understanding of human emotional states [23, 43, 46], multi-modal emotion recognition has emerged as an important topic in the field of affective computing. The dominant multi-modal emotion recognition methods can be divided into two subfields: single-label emotion recognition (SLER) and multi-label emotion recognition (MLER). SLER recognized emotions using a single emotion label for each discrete time. MAET [15] processed multi-modal inputs with specialized attention modules and alleviates subject discrepancy via adversarial training. While the goal of MLER is to assign an arbitrary number of emotion category labels to an input sample. Zhang et al. [39] integrated transformer and multi-head modality attention to effectively fuse text, visual, and audio data and predict emotions. Compared with SLER and MLER, emotion distribution learning (EDL) has gained increasing attention, as it evaluates the intensity of each basic emotion, which is crucial for a comprehensive representation of complex emotional space. EmotionDict [21] utilized behavior and physiological responses to offer complementary information for emotion recognition and disentangled emotion features of an emotional state into a weighted combination of a set of basic emotion elements. Shu et al.[28] proposed EDLConv the combine multi-modal peripheral physiological signals for emotion distribution learning. In this paper, we propose a novel EDL framework for integrating multi-modal emotional data, and the model's performance is improved with the guidance of learned label correlation.

3 Methodology

The overall framework of the proposed method is illustrated in Figure 2, which consists of physiological signals fusion, OT-based heterogeneity mining, and label correlation-driven cross-attention. We will briefly introduce these components in the remainder of this section.

3.1 Multi-Modal Physiological Signals Fusion

Taking the DMER dataset as an example, after data pre-processing, we get a sampled input x composed of four modalities, i.e., $x^e \in \mathbb{R}^{C_e \times d_e}$ for EEG, $x^g \in \mathbb{R}^{C_g \times d_g}$ for GSR, $x^p \in \mathbb{R}^{C_p \times d_p}$ for PPG, and $x^v \in \mathbb{R}^{C_v \times d_v}$ for video, where C_e, C_g, C_p, and C_v denote the number of signal channels, and d_e, d_g, d_p, and d_v are the extracted feature dimensions. As the heterogeneity within physiological signals (i.e., EEG, GSR, PPG, ECG, EDA, EMG) may be easier to handle than that of physiological and behavioral signals, we adopt a multi-head cross-attention mechanism to effectively fuse the physiological signals. Specifically, we first project EEG, GSR, and PPG signals into the same feature dimension. Then, we generate Queries on $x^e \in \mathbb{R}^{C \times d}$ and Keys, Values on $x^g \in \mathbb{R}^{C \times d}$ and $x^p \in \mathbb{R}^{C \times d}$, which can be formulated as follows:

$$Q^e = x^e \mathbf{W}_Q^e, K^m = x^m \mathbf{W}_K^m, V^m = x^m \mathbf{W}_V^m \tag{1}$$

where $\mathbf{W}_Q^e$, $\mathbf{W}_K^m$, $\mathbf{W}_V^m$ are learnable matrix, $m = \{g, p\}$ denotes the specific modality. Then, the cross-modal representation can be obtained by the self-attention mechanism:

$$\begin{aligned} x^{e \to g} &= \text{softmax}\left(\frac{Q^e\ (K^g)^T}{\sqrt{d}}\right) V^g \\ x^{e \to p} &= \text{softmax}\left(\frac{Q^e\ (K^p)^T}{\sqrt{d}}\right) V^p \end{aligned} \tag{2}$$

where d denotes a normalization parameter equal to the dimension of K^m. To learn multiple patterns of the association within physiological signals, multi-head attention mechanism and residual connection are employed:

$$x_{MHA}^{e \to m} = \left(x_{h_1}^{e \to m} \middle\| x_{h_2}^{e \to m} \cdots \middle\| x_{h_k}^{e \to m}\right) \mathbf{W}_{MHA} + LN\left(x^m\right) \tag{3}$$

where $x_{MHA}^{e \to m} \in \mathbb{R}^{C \times \frac{d}{h}}$ is the output of multi-head cross attention block, h is the number of heads, $\mathbf{W}_{MHA}$ denotes the transformation matrix, and || represents the concatenation operation. Then, the concatenation of two cross-modal representations is adopted as the final physiological representation:

$$x^{phy} = \left(x_{MHA}^{e \to g} \| x_{MHA}^{e \to p}\right) \tag{4}$$

Using EEG as the query guides the cross-attention mechanism to focus on the emotional information at the neural level, effectively capturing the relationship between EEG and other physiological signals.

3.2 Heterogeneity Mining with Optimal Transport

Physiological signals, i.e., EEG, GSR, ECG, and PPG, provide subconscious emotional responses, while behavioral data reflects discrete and context-specific insights that are influenced by conscious control and environmental interactions. It is worth noting that different data characteristics lead to significant cross-modal heterogeneity, which heightens the difficulty of multi-modal fusion. Optimal transport (OT) [3, 25, 29] learns the optimal matching flow between two probability distributions, which allows for the effective alignment of heterogeneous multi-modal data distributions, facilitating the extraction of coherent patterns from physiological and behavioral signals. Formally, for the previous extracted physiological representations and projected behavioral representations, $x^{Phy} \in \mathbb{R}^{2C \times d}$ and $x^v \in \mathbb{R}^{2C \times d}$, an optimal transport from the x^{Phy} to x^v can be defined by the discrete Kantorovich formulation [16] to search the overall optimal matching flow:

$$\mathcal{WD}\left(x^{Phy}, x^v\right) = \min_{f \in \Pi(u,v)} < T,\ \text{Cost}\ >_F \tag{5}$$

where $\mathcal{WD}(,)$ denotes the wasserstein distance, $\Pi(u, v) = \{T \in \mathbb{R}_+^{2C \times 2C} \mid T 1_{2C} = u, T^T 1_{2C} = v\}$, 1_{2C} represents a $2C$-dimensional all-one vector, $<,>_F$ refers to the Frobenius dot product, and Cost $\geq 0 \in \mathbb{R}^{2C \times 2C}$ is a cost matrix calculated by the cosine distance between x^{Phy} and x^v. The matrix $T \in \mathbb{R}^{2C \times 2C}$ is denoted as the optimal transport matching flow. Then, we apply the Pytorch framework-based batch-wise implementation of the Sinkhorn algorithm proposed in [4] to solve $\mathcal{WD}$. In this way, the obtained optimal transport matching flow T can be used as a "cross-modal correlation" to enhance the multi-modal fusion. As such, the final multi-modal representation is adopted as follows:

$$x^m = (\text{Transformer}_l(x^{Phy} \otimes T) || \text{Transformer}_l(x^v)) \tag{6}$$

where Transformer_l denotes a l-layer transformer encoder, $\otimes$ is the matrix multiplication, || represents the concatenation operation. The optimal transport (OT) module bridges modality heterogeneity by aligning physiological and behavioral feature spaces through optimal matching flows, thereby mitigating distribution discrepancies.

3.3 Label Correlation-driven Cross-Attention

In the emotion distribution learning scenario, correlation across different labels plays a vital role in the prediction of emotional

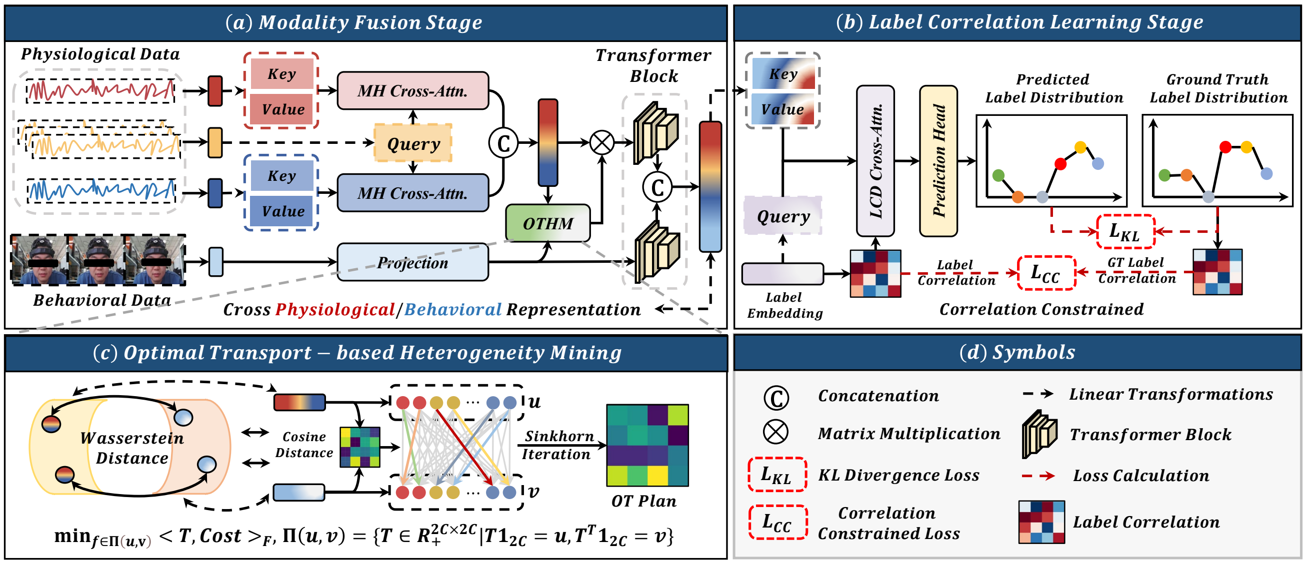


Figure 2: The pipeline of the proposed HeLo. (a) Modality fusion stage: We first adopt a cross-attention mechanism to fuse the physiological data. Then the OT-based heterogeneity mining module is introduced to integrate the physiological and behavioral representations. (b) Label correlation learning stage: A learnable label embedding is introduced to explore the label correlation that can be combined with the multi-modal features through a novel label correlation-driven cross-attention mechanism.

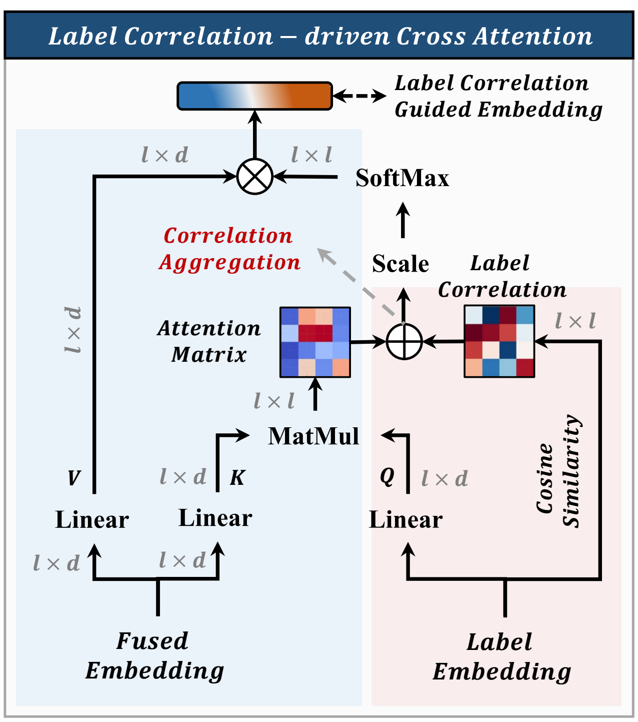


Figure 3: Illustration of the label correlation-driven cross-attention mechanism (LCDCA).

distributions. Therefore, we design a label correlation-driven cross-attention mechanism (LCDCA), shown in Figure 3, to effectively enhance the prediction performance of the model. Specifically, we introduce a learnable label embedding $x^L \in \mathbb{R}^{l\times d}$, where l is the number of emotion classes. To learn the label correlation, we first generate learnable label correlation $M^L \in \mathbb{R}^{l\times l}$ and ground-truth label correlation $M^{gt} \in \mathbb{R}^{l\times l}$ through the cosine similarity, whose (i, j)-th element is given as:

$$M^L_{i,j} = \frac{x^L_i \cdot x^L_j}{||x^L_i|| \cdot ||x^L_j||}, M^{gt}_{i,j} = \frac{L_i \cdot L_j}{||L_i|| \cdot ||L_j||} \tag{7}$$

where $L \in \mathbb{R}^{l\times 1}$ denotes the ground truth label distribution. Next, we adopt l-2 loss to constrain the learning of label correlation, and the correlation-constrained (CC) loss can be formulated as:

$$\mathcal{L}_{CC} = ||M^L - M^{gt}||^2_2 \tag{8}$$

Then, we generate Q^L on x^L, K^m and V^m on the projected multi-modal features x^m through linear transformations. The label correlation-guided representation can be obtained by:

$$x^o = softmax(\frac{Q^L(K^m)^T + M^L}{\sqrt{d}})V^m \tag{9}$$

With the LCDCA, the proposed model can effectively utilize the label correlation for emotion distribution learning. This design allows the model to dynamically aggregate global semantic relationships between features and labels, while simultaneously incorporating sample-specific correlations through attention matrix refinement. Finally, we adopt a prediction head consisting of a three-layer MLP and a softmax layer to generate the predicted emotion distribution $\hat{L} = (d_1, d_2, d_3, \cdots, d_l)$, where l is the number of emotion classes. The Kullback-Leibler divergence (KLD) is adopted as the loss function to measure the distance between the predicted emotion distribution and the ground truth emotion distribution. Therefore, the overall training of our method is:

$$\mathcal{L}_{Overall} = \mathrm{KLD}(\hat{L}, L) + \mathcal{L}_{CC} \tag{10}$$

where KLD(,) denotes the Kullback-Leibler divergence loss, L is the ground truth emotion label distribution, and $\mathcal{L}_{CC}$ is the correlation constrained loss.

4 Experiment

4.1 Multi-Modal Emotion Datasets

We evaluate our proposed method on two publicly available datasets: **DMER** [36] and **WESAD** [26] datasets. **DMER** dataset contains four modalities including EEG, galvanic skin response (GSR), photoplethysmogram (PPG), and facial videos collected from 80 participants. Each participant was asked to watch 32 video clips. In each trial, the subject first watched one video clip, and then completed the 10-item short positive and negative affect schedules (PANAS [33]). The score for each basic emotion ranged from 1 (none) to 5 (strong) and was then transformed into emotion distributions. The data of 73 subjects were used for further experiments in this paper. **WESAD** contains ECG, EMG, electrodermal activity (EDA), and 3-axis accelerometer (ACC) from 17 participants. The data were collected at a sampling rate of 700Hz, and the goal was to study four different affective states namely neutral, stressed, amused, and meditated. Upon completion of each trial, the ground truth labels for the affect states were collected using the PANAS [33]. 14 participants were utilized in our experiments. Notably, ECG, EMG, and EDA are the "Physiological data", and ACC denotes the "Behavioral Data" in our proposed method. Details of the data preprocessing can be referred to the ***supplementary materials***.

4.2 Evaluation Metrics

For the evaluation of our proposed model, following the previous emotion distribution learning studies [7, 21, 28], we adopt six distribution-based measurements, including four distance metrics (Chebyshev (↓), Clark (↓), Canberra (↓), and Kullback-Leibler (KL) (↓)) and two similarity measurements (Cosine (↑) and Intersection (↑)). ↓ indicates the metric is "the smaller the better", and ↑ denotes "the larger the better". Moreover, we reported the Average Rank denoting the mean rank of the six metrics following [21, 28].

4.3 Implementation Details

In our experiments, following previous works [21, 28], both subject-dependent and subject-independent settings are adopted to verify the method on the two datasets. In subject-dependent recognition, the data of each subject are split into the training set and testing set in a ratio of 8:2 randomly, and the average performance across all subjects is taken as the final result. For subject-independent recognition, we utilize the leave-one-subject-out cross-validation for the evaluation, and the average performance across all iterations is reported as the final result. Our proposed framework is trained on a 12GB NVIDIA GeForce RTX3080 GPU and developed under PyTorch. The training adopts Adam optimizer with a learning rate of 10^{-3}. The batch size is set to 128, the number of epochs for training is set to 300, and the number of attention heads is 4. The embedding size in the attention mechanism and feed-forward network is set to 128, and 64, respectively, and we set the depth of transformer blocks to 1. Our code is released at: https://github.com/kaio-99/HeLo.

4.4 Comparison with State-of-the-art Methods

To evaluate the effectiveness of our proposed method, we compare our model with several state-of-the-art models on both the DMER and AMIGOS datasets. These methods including label distribution learning methods (PT-SVM [7], AA-KNN [7], SA-CPNN [11]), EEG-based classification method (Conformer [30]), multi-modal emotion recognition method (MAET [15]), multi-modal multi-label emotion recognition method (CARAT [24]), recently proposed state-of-the-art single-modal label distribution learning methods (LDL-LRR [13], TLRLDL [17], CAD [34]), and multi-modal emotion distribution learning method (EmotionDict [21]).

The results of the subject-dependent recognition on the DMER and WESAD datasets are summarized in Table 1. As can be seen from the table, except for the KL metric on the DMER dataset, our proposed method outperforms all the other state-of-the-art methods over the six metrics on both datasets. Benefiting from deep learning's ability to extract high-dimensional features, the deep learning-based methods achieve relatively better performance than the traditional LDL methods (PT-SVM, AA-KNN, and SA-CPNN). Besides, visualization examples of the predicted emotion distributions using all the methods and the ground truths on the DMER dataset are shown in Figure 4. It can be seen that some methods, i.e., PT-SVM, SA-CPNN, Conformer, and MAET, are more likely to output similar distributions for different signal samples, leading to sub-optimal emotion distribution learning results. The qualitative results show that our method can predict emotion distributions most similar to the ground truths. The superiority of our method lies in that our model not only effectively utilizes the fusion of heterogeneous multi-modal data, but also leverages the learnable label correlation for the prediction of emotion distribution.

Table 2 illustrates the subject-independent evaluation of the comparison methods and our proposed model on the DMER datasets. Except for the KL metric on the WESAD dataset, our method achieves the best results across both datasets and shows the highest average rank. The subject-independent recognition is more challenging than the subject-dependent situation, showing the superiority of our proposed method in capturing the subject-invariant emotional features across different individuals.

4.5 Ablation Studies

Ablation on network components. To verify the rationality of the modules of the proposed framework, we conduct ablation experiments on the DMER and WESAD datasets by removing each module of our method, i.e., cross-attention-based physiological signals fusion (CAPF), optimal transport-based heterogeneity mining (OTHM), and label-correlation-driven cross-attention (LCDCA), and the results are illustrated in Figure 5. The experimental results show that each module in the model has a positive contribution to the emotion distribution learning performance of the model, and the proposed method achieves the best performance through the effective integration of each module. CAPF effectively fuses the

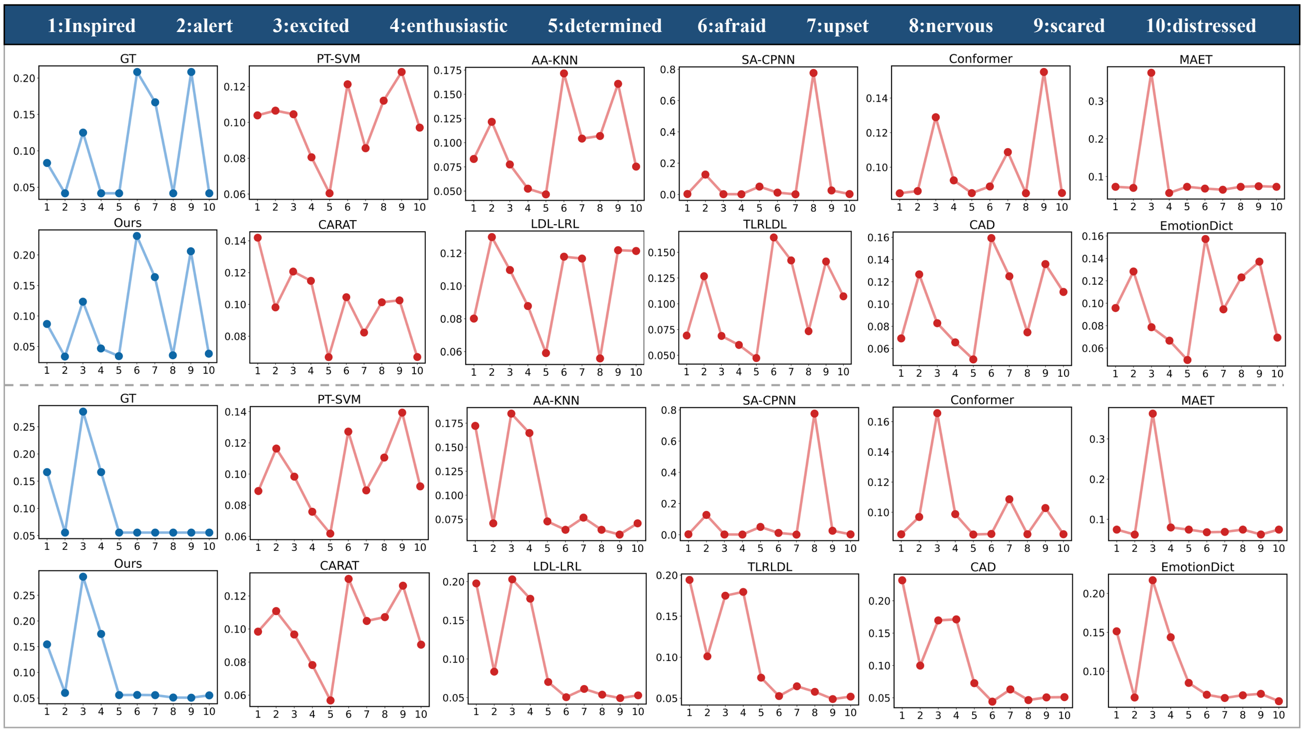


Figure 4: Predicted emotion distributions of our model and comparison methods. We show two panels of two test samples of Subject No.45. GT indicates the ground truth distribution. The numbers 1 to 10 correspond to emotions inspired, alert, excited, enthusiastic, determined, afraid, upset, nervous, scared, and distressed.

Table 1: Subject-dependent comparison of experimental results of our method and 10 comparison methods on six measures on the DMER and WESAD datasets. ↓ indicates "the smaller the better", and ↑ indicates "the larger the better". The parentheses show the corresponding ranks on each evaluation metric and the *Average Ranks*. The boldfaced scores denote the best performances and the underlined scores are the second runners. The results of the proposed HeLo are presented in a blue background.

Dataset	Measure	TKDE16 [7]	TKDE16 [7]	TPAMI13 [11]	TNSRE22 [30]	ACMMM23 [15]	AAAI24 [24]	TKDE23 [13]	IJCAI24 [17]	ICCV23 [34]	TAFFC23 [21]	Ours
		PT-SVM	AA-KNN	SA-CPNN	Conformer	MAET	CARAT	LDL-LRR	TLRLDL	CAD	EmotionDict	
DMER	Chebyshev (↓)	0.0851 (7)	0.0949 (10)	0.0917 (8)	0.0661 (3)	0.0522 (2)	0.0752 (4)	0.0936 (9)	0.0847 (6)	0.1017 (11)	0.0753 (5)	**0.0446 (1)**
	Clark (↓)	0.5787 (6)	0.6649 (10)	0.6333 (8)	0.4743 (3)	0.3762 (2)	0.5951 (7)	0.6529 (9)	0.5380 (4)	0.7517 (11)	0.5463 (5)	**0.3256 (1)**
	Canberra (↓)	1.5294 (7)	1.2357 (3)	1.8092 (9)	1.2581 (5)	0.9875 (2)	1.5965 (8)	1.8224 (10)	1.3271 (5)	1.9415 (11)	1.4771 (6)	**0.8664 (1)**
	KL (↓)	0.0837 (6)	0.1409 (10)	0.1048 (8)	0.0593 (3)	**0.0288 (1)**	0.1137 (9)	0.0644 (4)	0.0925 (7)	0.1460 (11)	0.0766 (5)	0.0323 (2)
	Cosine (↑)	0.9237 (8)	0.9315 (6)	0.9058 (9)	0.9464 (3)	0.9688 (2)	0.9332 (5)	0.9016 (10)	0.9339 (4)	0.8923 (11)	0.9313 (7)	**0.9714 (1)**
	Intersection (↑)	0.8417 (8)	0.8425 (7)	0.8150 (9)	0.8716 (3)	0.9113 (2)	0.8474 (6)	0.8136 (10)	0.8537 (5)	0.8077 (11)	0.8492 (5)	**0.9128 (1)**
	Average Rank (↓)	7 (7)	7.66 (8)	8.5 (9)	3.33 (3)	1.83 (2)	6.16 (6)	9.5 (10)	4.66 (4)	11 (11)	5.5 (5)	**1.16 (1)**
WESAD	Chebyshev (↓)	0.0515 (8)	0.0346 (4)	0.0998 (11)	0.0209 (3)	0.0347 (5)	0.0354 (6)	0.0357 (7)	0.0622 (10)	0.0569 (9)	0.0090 (2)	**0.0073 (1)**
	Clark (↓)	0.3940 (8)	0.2779 (4)	0.7285 (11)	0.1838 (3)	0.3503 (7)	0.2812 (5)	0.2822 (6)	0.5381 (10)	0.4618 (9)	0.0733 (2)	**0.0653 (1)**
	Canberra (↓)	1.0249 (8)	0.7381 (4)	2.0746 (11)	0.4810 (3)	0.9100 (7)	0.7430 (5)	0.7466 (6)	1.3586 (10)	1.2199 (9)	0.1876 (2)	**0.1614 (1)**
	KL (↓)	0.0320 (8)	0.0243 (5)	0.1226 (10)	0.0110 (3)	0.0251 (6)	0.0210 (4)	0.0273 (7)	0.2051 (11)	0.0440 (9)	0.0017 (2)	**0.0010 (1)**
	Cosine (↑)	0.9728 (8)	0.9793 (7)	0.8918 (11)	0.9909 (3)	0.9867 (4)	0.9824 (5)	0.9821 (6)	0.9541 (10)	0.9624 (9)	0.9985 (2)	**0.9992 (1)**
	Intersection (↑)	0.9024 (8)	0.9297 (4)	0.7896 (11)	0.9550 (3)	0.9213 (7)	0.9296 (5)	0.9292 (6)	0.8746 (10)	0.8812 (9)	0.9820 (2)	**0.9905 (1)**
	Average Rank (↓)	8 (8)	5.33 (5)	10.83 (11)	3 (3)	6 (6)	5 (4)	6.33 (7)	10.16 (10)	9 (9)	2.5 (2)	**1 (1)**

physiological signals to enhance the model's performance. OTHM achieves better performance than concatenation, indicating its proficiency in capturing heterogeneous information across physiological and behavioral features. Furthermore, LCDCA greatly improves the model's performance as it integrates the label correlations for the learning of emotion distribution.

Ablation on the modalities. We also assessed the impact of each modality by separately removing each modality from our full model, the results are shown in Table 3 and 4. These experiments were

Table 2: Subject-independent comparison of experimental results of our method and 10 comparison methods on six measures on the DMER and WESAD datasets. ↓ indicates "the smaller the better", and ↑ indicates "the larger the better". The parentheses show the corresponding ranks on each evaluation metric and the *Average Ranks*. The boldfaced scores denote the best performances and the underlined scores are the second runners. The results of the proposed HeLo are presented in a blue background.

Dataset	Measure	TKDE16 [7] PT-SVM	TKDE16 [7] AA-KNN	TPAMI13 [11] SA-CPNN	TNSRE22 [30] Conformer	ACMMM23 [15] MAET	AAAI24 [24] CARAT	TKDE23 [13] LDL-LRR	IJCAI24 [17] TLRLDL	ICCV23 [34] CAD	TAFFC23 [21] EmotionDict	Ours
DMER	Chebyshev (↓)	0.1462 (11)	0.1070 (10)	0.0928 (3)	0.0933 (4)	0.0939 (5)	0.1023 (9)	0.1012 (8)	0.0964 (6)	0.0925 (2)	0.0945 (7)	**0.0882 (1)**
	Clark (↓)	1.2877 (11)	0.6860 (9)	0.6440 (3)	0.6515 (4)	0.6605 (5)	0.6909 (10)	0.6856 (8)	0.6702 (6)	0.6379 (2)	0.6676 (7)	**0.6289 (1)**
	Canberra (↓)	2.0459 (11)	1.8308 (7)	1.8542 (8)	1.7689 (2)	1.7863 (3)	1.9161 (10)	1.9071 (9)	1.8244 (6)	1.8104 (5)	1.8092 (4)	**1.7603 (1)**
	KL (↓)	0.5528 (11)	0.1227 (9)	0.1089 (5)	0.1061 (2)	0.1088 (4)	0.1245 (10)	0.1222 (8)	0.1142 (7)	0.1076 (3)	0.1104 (6)	**0.1027 (1)**
	Cosine (↑)	0.8295 (11)	0.8843 (10)	0.9022 (4)	0.9038 (2)	0.9016 (5)	0.8899 (9)	0.8917 (8)	0.8964 (7)	0.9036 (3)	0.9001 (6)	**0.9148 (1)**
	Intersection (↑)	0.7762 (11)	0.8050 (8)	0.8102 (6)	0.8161 (2)	0.8147 (3)	0.8026 (10)	0.8037 (9)	0.8087 (7)	0.8122 (4)	0.8119 (5)	**0.8194 (1)**
	Average Rank (↓)	11 (11)	8.83 (9)	4.83 (5)	2.66 (2)	4.16 (4)	9.66 (10)	8.33 (8)	6.5 (7)	3.16 (3)	5.83 (6)	**1 (1)**
WESAD	Chebyshev (↓)	0.0471 (7)	0.0500 (9)	0.0894 (11)	0.0436 (3)	0.0516 (10)	0.0454 (5)	0.0466 (6)	0.0441 (4)	0.0481 (8)	0.0421 (2)	**0.0403 (1)**
	Clark (↓)	0.4506 (8)	0.4516 (10)	0.6394 (11)	0.4048 (3)	0.4514 (9)	0.4270 (5)	0.4471 (6)	0.4145 (4)	0.4488 (7)	0.3631 (2)	**0.3455 (1)**
	Canberra (↓)	1.1928 (10)	1.1762 (7)	1.7384 (11)	1.0806 (4)	1.1666 (6)	1.1874 (9)	1.1817 (8)	1.0746 (3)	1.1647 (5)	0.9530 (2)	**0.9329 (1)**
	KL (↓)	0.0347 (5)	0.0423 (9)	0.0944 (11)	0.0369 (7)	0.0445 (10)	0.0345 (4)	0.0343 (3)	0.0372 (8)	0.0366 (6)	**0.0282 (1)**	0.0283 (2)
	Cosine (↑)	0.9727 (5)	0.9663 (10)	0.9154 (11)	0.9699 (7)	0.9678 (9)	0.9731 (3)	0.9731 (3)	0.9695 (8)	0.9711 (6)	0.9773 (2)	**0.9790 (1)**
	Intersection (↑)	0.8935 (7)	0.8933 (9)	0.8254 (11)	0.9015 (3)	0.8967 (4)	0.8944 (6)	0.8945 (5)	0.8917 (10)	0.8935 (7)	0.9126 (2)	**0.9154 (1)**
	Average Rank (↓)	7 (8)	9 (10)	11 (11)	4.5 (3)	8 (9)	5.33 (5)	5.16 (4)	6.16 (6)	6.5 (7)	1.83 (2)	**1.16 (1)**

Table 3: Ablation study in modalities on the DMER.

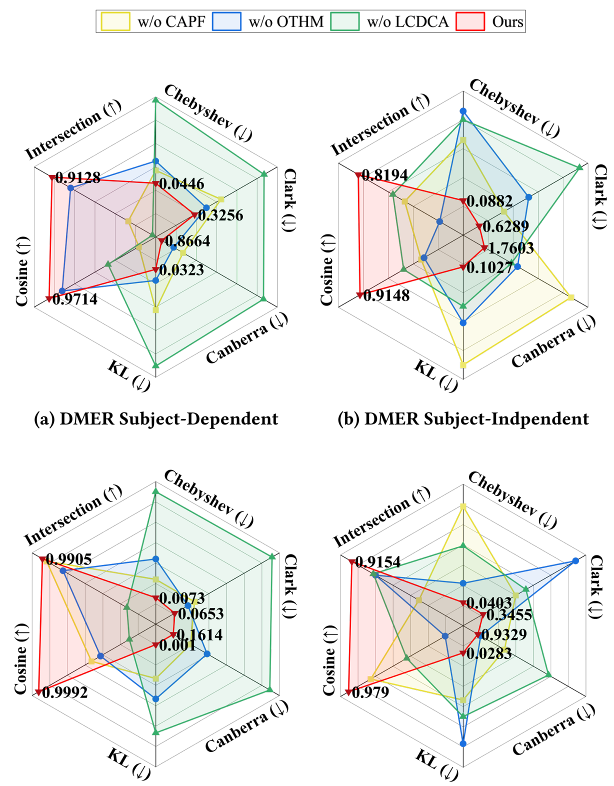
Settings	Measure	w/o EEG	w/o GSR	w/o PPG	w/o Video	**Ours**
Subject-Dependent	Chebyshev (↓)	0.0529	0.0574	0.0487	0.0552	**0.0446**
	Clark (↓)	0.3372	0.3550	0.3671	0.3380	**0.3256**
	Canberra (↓)	0.9028	0.8987	0.8820	0.9065	**0.8664**
	KL (↓)	0.0381	0.0429	0.0357	0.0377	**0.0323**
	Cosine (↑)	0.9677	0.9591	0.9692	0.9628	**0.9714**
	Intersection (↑)	0.9102	0.9007	0.9086	0.9033	**0.9128**
Subject-Independent	Chebyshev (↓)	0.0977	0.0953	0.0964	0.0926	**0.0882**
	Clark (↓)	0.6906	0.6742	0.6470	0.6635	**0.6289**
	Canberra (↓)	1.9261	1.9077	1.8074	1.8250	**1.7603**
	KL (↓)	0.1175	0.1244	0.1085	0.1134	**0.1027**
	Cosine (↑)	0.9080	0.9075	0.9122	0.9107	**0.9148**
	Intersection (↑)	0.8197	0.8056	0.8138	0.8121	**0.8194**

Table 4: Ablation study in modalities on the WESAD.

Settings	Measure	w/o ECG	w/o EMG	w/o EDA	w/o ACC	**Ours**
Subject-Dependent	Chebyshev (↓)	0.0088	0.0074	0.0079	0.0082	**0.0073**
	Clark (↓)	0.0744	0.0722	0.0762	0.0695	**0.0653**
	Canberra (↓)	0.1730	0.1630	0.1658	0.1725	**0.1614**
	KL (↓)	0.0022	0.0016	0.0014	0.0016	**0.0010**
	Cosine (↑)	0.9915	0.9960	0.9955	0.9971	**0.9992**
	Intersection (↑)	0.9873	0.9899	0.9892	0.9885	**0.9905**
Subject-Independent	Chebyshev (↓)	0.0502	0.0419	0.0429	0.0474	**0.0403**
	Clark (↓)	0.3877	0.3597	0.3721	0.3842	**0.3455**
	Canberra (↓)	1.0748	0.9516	0.9579	1.0029	**0.9329**
	KL (↓)	0.0322	0.0308	0.0344	0.0315	**0.0283**
	Cosine (↑)	0.9688	0.9722	0.9641	0.9658	**0.9790**
	Intersection (↑)	0.8870	0.9095	0.9017	0.8991	**0.9154**

conducted in both subject-dependent and subject-independent settings across both datasets. The results indicate that each modality contributes significantly to the overall performance of multi-modal emotion distribution learning, effectively integrating multiple modalities can enhance the emotion distribution performance.

Ablation on the number of attention heads. The number of attention heads is usually an essential parameter for models based on attention mechanisms. The result in Figure 6 indicates that, the model's performance in both subject-dependent and subject-independent settings initially improves as the number of heads increases, and subsequently declines. The model reaches its optimal performance across all six evaluation metrics when the number of attention heads was 4. However, the overall impact of the number of attention heads on the model's performance is relatively limited.



(a) DMER Subject-Dependent (b) DMER Subject-Indpendent

(c) WESAD Subject-Dependent (d) WESAD Subject-Independent

Figure 5: Ablation of network components in HeLo.

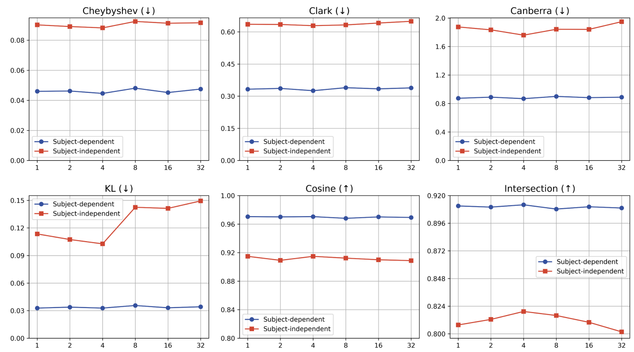


Figure 6: Effect of attention heads on the DMER

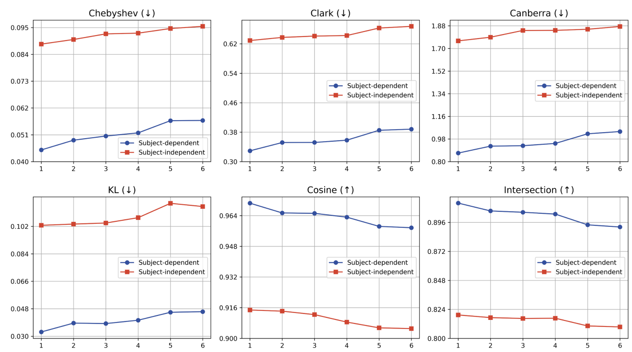
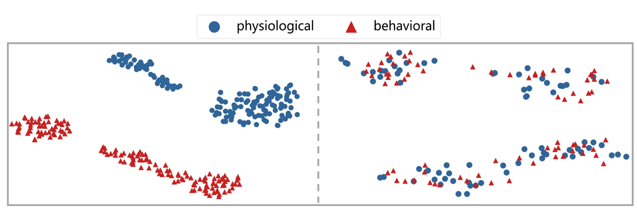


Figure 7: Effect of transformer block depth on the DMER.

Ablation on the depth of transformer block. The depth of end-to-end models is considered a crucial factor in determining their fitting ability. Figure 7 illustrates that, the model achieved optimal classification performance when the transformer block depth was 1. As the depth increased, the classification accuracy showed a slight tendency to decrease and then leveled off. This is because a greater transformer block depth might cause overfitting.



(a) Before OTHM (b) After OTHM

Figure 8: Visualization of the feature embedding.

4.6 Visualization

Visualization of the cross-modal representations. We visualize the multi-modal feature space in the testing set by utilizing t-SNE [31]. In Figure 8 (a), before the OTHM module, there is a noticeable separation between physiological (blue) and behavioral (red) representations. This pronounced heterogeneity illustrates the challenges faced in multi-modal fusion due to distinct feature distributions, which can impede effective integration and analysis. With the help of OTHM, shown in Figure 8 (b), the cross-modal representations exhibit a closer arrangement and overlap more significantly than before. This reduction in heterogeneity suggests that OTHM has effectively minimized the disparities between the physiological and behavioral features, facilitating a better-aligned feature space and a more coherent multi-modal fusion.

Visualization of the learned label correlations. To verify the effectiveness of the proposed LCDCA module, we visualize the learned label correlation M^L in Figure 7 to illustrate the interpretability, and the results on the testing set of DMER dataset are shown in Figure 9. As can be seen from Figure 9, the label correlations differ from different trials and different subjects, which jointly attend to rich semantic information from various perspectives. From the horizontal view, "afraid" is highly correlated with "nervous", "scared". Besides, "enthusiastic", and "determined" are usually correlated with "excited", "alert". All of these conclusions are consistent with human intuition, so as to demonstrate the better emotion distribution learning performance of our proposed model.

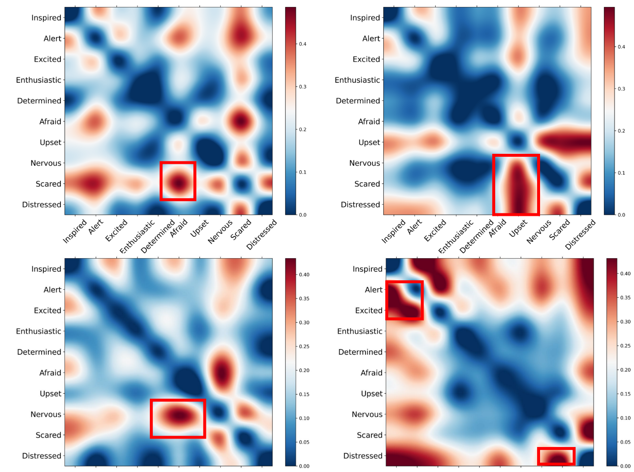


Figure 9: Label correlations visualization. A higher red intensity value indicates a stronger correlation.

5 Conclusion

In this paper, we introduce a novel multi-modal emotion distribution learning framework, HeLo, which effectively addresses the challenges of capturing complex, mixed emotional states from heterogeneous physiological and behavioral data. The CAPF module ensures the robust fusion of multi-modal physiological signals, and the OTHM module efficiently mines the heterogeneity between physiological and behavioral representations, bridging the gap between these diverse data sources and enhancing feature alignment. Moreover, the LCDCA fully exploits label correlations, guiding the learning of emotion distributions in a semantically meaningful way. Experimental results on two multi-modal emotion datasets demonstrate the efficacy of our proposed method. In the future, we plan to utilize LLMs to extract the label embeddings, which provide strong semantic priors to our model instead of random initialization.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62076129, 62371234, and 62136004), Natural Science Foundation of Jiangsu Province (No. BK20231438) and Key Research and Development Plan of Jiangsu Province (No. BE2022842).

References

[1] Soraia M Alarcao and Manuel J Fonseca. 2017. Emotions recognition using EEG signals: A survey. *IEEE Transactions on Affective Computing* 10, 3 (2017), 374–393.

[2] Yekta Said Can, Bhargavi Mahesh, and Elisabeth André. 2023. Approaches, applications, and challenges in physiological emotion recognition—a tutorial overview. *Proc. IEEE* (2023).

[3] Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. 2022. Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in Neural Information Processing Systems* 35 (2022), 39090–39102.

[4] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. 2020. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*. PMLR, 1542–1553.

[5] Bo Cheng and Guangyuan Liu. 2008. Emotion recognition from surface EMG signal using wavelet transform and neural network. In *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*. IEEE, 1363–1366.

[6] Bin-Bin Gao, Chao Xing, Chen-Wei Xie, Jianxin Wu, and Xin Geng. 2017. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing* 26, 6 (2017), 2825–2838.

[7] Xin Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering* 28, 7 (2016), 1734–1748.

[8] Xin Geng and Rongzi Ji. 2013. Label Distribution Learning. In *2013 IEEE 13th International Conference on Data Mining Workshops*. IEEE, 377–383.

[9] Xin Geng, Kate Smith-Miles, and Zhi-Hua Zhou. 2010. Facial Age Estimation by Learning from Label Distributions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 24. 451–456.

[10] Xin Geng, Qin Wang, and Yu Xia. 2014. Facial age estimation by adaptive label distribution learning. In *2014 22nd International Conference on Pattern Recognition*. IEEE, 4465–4470.

[11] Xin Geng, Chao Yin, and Zhi-Hua Zhou. 2013. Facial age estimation by learning from label distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 10 (2013), 2401–2412.

[12] Yu-Liang Hsu, Jeen-Shing Wang, Wei-Chun Chiang, and Chien-Han Hung. 2017. Automatic ECG-based emotion recognition in music listening. *IEEE Transactions on Affective Computing* 11, 1 (2017), 85–99.

[13] Xiuyi Jia, Xiaoxia Shen, Weiwei Li, Yunan Lu, and Jihua Zhu. 2023. Label Distribution Learning by Maintaining Label Ranking Relation. *IEEE Transactions on Knowledge & Data Engineering* 35, 02 (2023), 1695–1707.

[14] Xiuyi Jia, Xiang Zheng, Weiwei Li, Changqing Zhang, and Zechao Li. 2019. Facial emotion distribution learning by exploiting low-rank label correlations locally. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9841–9850.

[15] Wei-Bang Jiang, Xuan-Hao Liu, Wei-Long Zheng, and Bao-Liang Lu. 2023. Multimodal adaptive emotion transformer with flexible modality inputs on a novel dataset with continuous labels. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5975–5984.

[16] Leonid V Kantorovich. 2006. On the Translocation of Masses. *Journal of Mathematical Sciences* 133, 4 (2006).

[17] Zhiqiang Kou, Jing Wang, Jiawei Tang, Yuheng Jia, Boyu Shi, and Xin Geng. 2024. Exploiting Multi-Label Correlation in Label Distribution Learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, Kate Larson (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4326–4334. doi:10.24963/ijcai.2024/478 Main Track.

[18] Peter J Lang. 1995. The emotion probe: Studies of motivation and attention. *American psychologist* 50, 5 (1995), 372.

[19] Cunbo Li, Peiyang Li, Yangsong Zhang, Ning Li, Yajing Si, Fali Li, Zehong Cao, Huafu Chen, Badong Chen, Dezhong Yao, et al. 2023. Effective emotion recognition by learning discriminative graph topologies in EEG brain networks. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[20] Shan Li and Weihong Deng. 2019. Blended emotion in-the-wild: Multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. *International Journal of Computer Vision* 127, 6 (2019), 884–906.

[21] Fang Liu, Pei Yang, Yezhi Shu, Fei Yan, Guanhua Zhang, and Yong-Jin Liu. 2023. Emotion dictionary learning with modality attentions for mixed emotion exploration. *IEEE Transactions on Affective Computing* (2023).

[22] Minxu Liu, Donghai Guan, Chuhang Zheng, and Qi Zhu. 2025. Multi-Modal Discriminative Network for Emotion Recognition across Individuals. *IEEE Transactions on Cognitive and Developmental Systems* (2025), 1–13. doi:10.1109/TCDS.2025.3552124

[23] Wang Kay Ngai, Haoran Xie, Di Zou, and Kee-Lee Chou. 2022. Emotion recognition based on convolutional neural networks and heterogeneous bio-signal data sources. *Information Fusion* 77 (2022), 107–117.

[24] Cheng Peng, Ke Chen, Lidan Shou, and Gang Chen. 2024. CARAT: Contrastive Feature Reconstruction and Aggregation for Multi-Modal Multi-Label Emotion Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 14581–14589.

[25] Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. 2016. Mapping estimation for discrete optimal transport. *Advances in Neural Information Processing Systems* 29 (2016).

[26] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. 400–408.

[27] Ben Shneiderman and Catherine Plaisant. 2010. *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India.

[28] Yezhi Shu, Pei Yang, Niqi Liu, Shu Zhang, Guozhen Zhao, and Yong-Jin Liu. 2022. Emotion distribution learning based on peripheral physiological signals. *IEEE Transactions on Affective Computing* 14, 3 (2022), 2470–2483.

[29] Andrew H Song, Richard J Chen, Guillaume Jaume, Anurag J Vaidya, Alexander S Baras, and Faisal Mahmood. 2024. Multimodal Prototyping for cancer survival prediction. *arXiv preprint arXiv:2407.00224* (2024).

[30] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. 2022. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2022), 710–719.

[31] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008).

[32] Monu Verma, M Satish Kumar Reddy, Yashwanth Reddy Meedimale, Murari Mandal, and Santosh Kumar Vipparthi. 2021. Automer: Spatiotemporal neural architecture search for microexpression recognition. *IEEE Transactions on Neural Networks and Learning Systems* 33, 11 (2021), 6116–6128.

[33] D Watson. 1994. The PANAS-X: Manual for the positive and negative affect schedule-expanded form. *The University of Iowa* (1994).

[34] Changsong Wen, Xin Zhang, Xingxu Yao, and Jufeng Yang. 2023. Ordinal label distribution learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23481–23491.

[35] Patti Williams and Jennifer L Aaker. 2002. Can mixed emotions peacefully coexist? *Journal of consumer research* 28, 4 (2002), 636–649.

[36] Pei Yang, Niqi Liu, Xinge Liu, Yezhi Shu, Wenqi Ji, Ziqi Ren, Jenny Sheng, Minjing Yu, Ran Yi, Dan Zhang, et al. 2024. A Multimodal Dataset for Mixed Emotion Recognition. *Scientific Data* 11, 1 (2024), 847.

[37] Mengqing Ye, CL Philip Chen, and Tong Zhang. 2022. Hierarchical dynamic graph convolutional network with interpretability for EEG-based emotion recognition. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[38] Lu Yi and Man-Wai Mak. 2020. Improving speech emotion recognition with adversarial data augmentation network. *IEEE Transactions on Neural Networks and Learning Systems* 33, 1 (2020), 172–184.

[39] Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multi-modal multi-label emotion detection with modality and label dependence. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language processing (EMNLP)*. 3584–3593.

[40] Dong Zhang, Xincheng Ju, Wei Zhang, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 14338–14346.

[41] Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. 2018. Text Emotion Distribution Learning via Multi-Task Convolutional Neural Network.. In *IJCAI*. 4595–4601.

[42] Guozhen Zhao, Yulin Zhang, Guanhua Zhang, Dan Zhang, and Yong-Jin Liu. 2020. Multi-target positive emotion recognition from EEG signals. *IEEE Transactions on Affective Computing* 14, 1 (2020), 370–381.

[43] Chuhang Zheng, Wei Shao, Daoqiang Zhang, and Qi Zhu. 2023. Prior-driven dynamic brain networks for multi-modal emotion recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 389–398.

[44] Wei-Long Zheng and Bao-Liang Lu. 2016. Personalizing EEG-based affective models with transfer learning. In *Proceedings of the twenty-fifth International Joint Conference on Artificial Intelligence*. 2732–2738.

[45] Qi Zhu, Chuhang Zheng, Zheng Zhang, Wei Shao, and Daoqiang Zhang. 2024. Dynamic Confidence-Aware Multi-Modal Emotion Recognition. *IEEE Transactions on Affective Computing* 15, 3 (2024), 1358–1370. doi:10.1109/TAFFC.2023.3340924

[46] Qi Zhu, Ting Zhu, Lunke Fei, Chuhang Zheng, Wei Shao, David Zhang, and Daoqiang Zhang. 2025. Multi-Modal Cross-Subject Emotion Feature Alignment and Recognition with EEG and Eye Movements. *IEEE Transactions on Affective Computing* (2025), 1–15. doi:10.1109/TAFFC.2025.3554399