

Supplementary Materials: Correlation-Driven Multi-Modality Graph Decomposition for Cross-Subject Emotion Recognition

The subsequent sections consist of supplementary materials that offer in-depth insights into the extra analysis, extra implementations, and extra results, which could not be accommodated in the main text due to space limitations. The sections in supplementary materials are arranged in the order of when they are mentioned in the main manuscript. To distinguish them from the main manuscript, sections in the main text are denoted by "Section" followed by a number, whereas sections in this supplementary material are referenced as Appendix followed by an alphabet. For clarity, a concise summary of the supplementary materials is provided below:

(1) **Extra Analysis:** Initially, these supplementary materials introduce a more detailed analysis of the proposed CMMGD framework. Appendix A gives the distributions of multi-modality physiological signals across subjects, as the Principal Challenges paragraphs outlined in Section 1. Subsequently, Appendix B offers a comprehensive comparison between the proposed MGRT backbone and the most recent graph transformer architecture in emotion recognition. Additionally, Appendix D conducts a comparative analysis between the CMMGD framework and disentanglement-based methods to demonstrate the novelty and effectiveness of our approach.

(2) **Extra Implementations:** We further provide detailed implementations of the CMMGD framework. Appendix C details the computation process of the divergence loss (Equation (11)) as described in Section 4.2. Furthermore, Appendix E delineates the primary algorithm of CMMGD. Appendix F introduces the adopted two benchmark datasets, while Appendix G outlines the data pre-processing steps and hyperparameter configurations. Lastly, Appendix H introduces the descriptions of the baseline methods employed in the experiments.

(3) **Extra Results:** Finally, we present additional results. Appendix I offers visualizations of cross-modality correlation, whereas Appendix J presents brain topographic maps to further validate the efficacy and interpretability of our CMMGD framework.

A DISTRIBUTION ACROSS SUBJECTS

This section provides a visualization of the distribution of the multi-modality physiological signals across subjects. As shown in Figure 1, the distribution of the physiological signals varies across the first six subjects, highlighting the cross-subject divergence. Additionally, the distribution of different modalities differs significantly. This finding further emphasizes the need for a generalizable model that can simultaneously generalize across subjects and modalities.

B COMPARISON OF BACKBONE

This section delineates the architecture of EmoGTs [7] and MGRT, providing a graphical comparison as illustrated in Figure 2. EmoGTs is the latest graph transformer architectures designed for multi-modal emotion recognition tasks. It captures multimodal features by integrating the graph transformer with the graph convolutional network. The enhancements introduced in the MGRT backbone

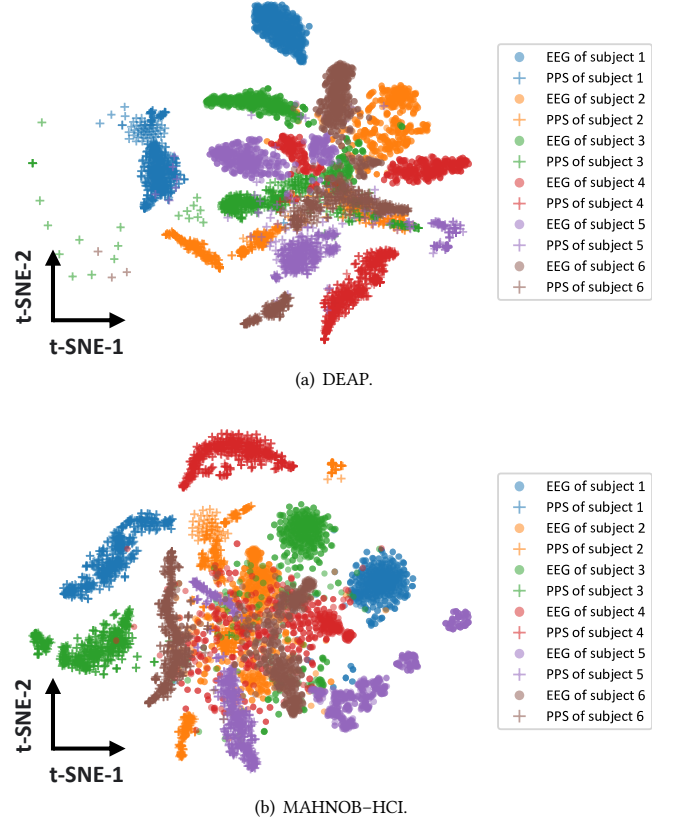


Figure 1: Distribution of the multi-modality physiological signals across subjects.

are two-fold: firstly, the parallel operation of the graph regularizer (GR) and the multi-modality self-attention (MMSA), and secondly, the removal of the linear transformer in the GR module.

As depicted in Figure 2, one EmoGTs layer comprises a GCN layer and a cross-modal attention-based transformer layer. The GCN layer is tasked with capturing the local features inherent to each modality, whereas the transformer layer is designed to encapsulate the global features. These two types of features are processed in a sequential manner, with the outputs of the GCN layer being inputs to the transformer layer, and they are iteratively updated. Nonetheless, given the relatively small scale of graphs derived from EEG and PPS modalities, there exists a potential for over-smoothing, which leads to the loss of local representation. This scenario posits the risk of either the GCN or transformer layer predominating the feature extraction process, thereby constituting a bottleneck. The parallel execution of GR and MMSA in MGRT serves to mitigate this risk, with the GR function acting as a regularizer for the global MMSA,

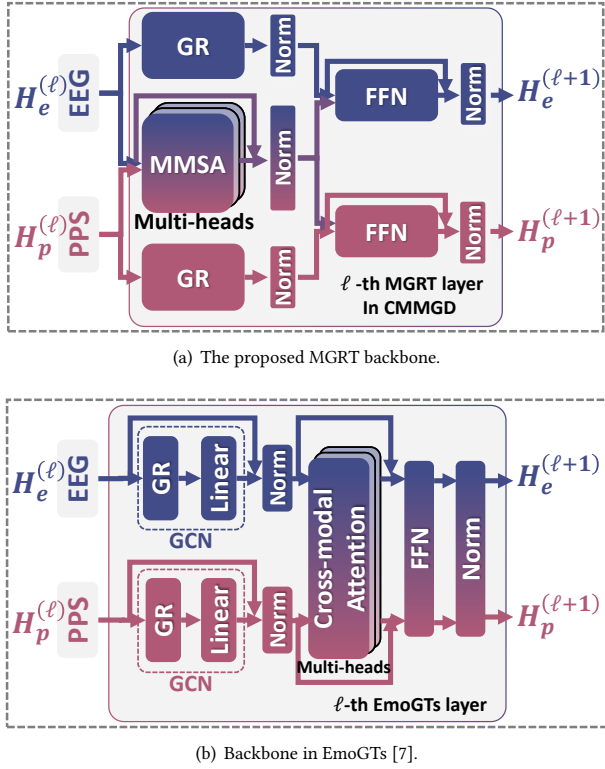


Figure 2: Comparison of the backbones of the latest EmoGTs and the proposed MGRT.

thereby ensuring an equitable amalgamation of local and global feature sets.

Previous studies have explored the parallel operation of graph convolutional networks and transformers, focusing on unimodal graphs [20]. In our work, we have extended an enhanced version of GraphGPS, denoted as GraphGPS*, for comparative analysis. The findings, as presented in Table 4, underscore the superior overall performance of the proposed MGRT backbone, which is attributed to the second-tier enhancement, in which we remove the linear transformer within the GR module. This refinement ensures the MGRT backbone is optimally configured for small-scale graphs, as typically observed in EEG and PPS modalities, by retaining solely the graph convolutional operation and eschewing the linear transformation.

C THE DETAILS OF THE DIVERGENCE LOSS

In Section 4.2, we introduce the divergence loss \mathcal{L}_{div} to encourage the model to capture the diverse features of the concordant and discrepant subgraphs. We realize this by leveraging the distance correlation (dCor) [24, 31] to quantify the dependencies between the concordant and discrepant subgraphs. This section provides a detailed derivation of the divergence loss \mathcal{L}_{div} .

The Distance Correlation. The dCor between random vectors X and Y , denoted by $\mathcal{R}(X, Y)$, satisfies $0 \leq \mathcal{R}(X, Y) \leq 1$, where $\mathcal{R}(X, Y) = 0$ indicates independence, and $\mathcal{R}(X, Y) = 1$ denotes

perfect dependence. Furthermore, in practice, we concentrate on the empirical distance correlation $\mathcal{R}_n(X, Y)$ of two random vectors X , and Y :

$$\mathcal{R}_n^2(X, Y) = \begin{cases} \frac{\mathcal{V}_n^2(X, Y)}{\sqrt{\mathcal{V}_n^2(X, X)\mathcal{V}_n^2(Y, Y)}}, & \text{if } \mathcal{V}_n(X, X)\mathcal{V}_n(Y, Y) > 0 \\ 0, & \text{if } \mathcal{V}_n(X, X)\mathcal{V}_n(Y, Y) = 0 \end{cases}, \quad (1)$$

where $\mathcal{V}_n^2(X, Y)$ is the empirical distance covariance, and $\mathcal{V}_n^2(X, X)$ and $\mathcal{V}_n^2(Y, Y)$ are the empirical variances of X and Y , respectively. These empirical statistics can be further defined by:

$$\begin{aligned} a_{k,l} &= \|X_k - X_l\|, \quad \bar{a}_{k,\cdot} = \frac{1}{n} \sum_{l=1}^n a_{k,l}, \\ \bar{a}_{\cdot,l} &= \frac{1}{n} \sum_{k=1}^n a_{k,l}, \quad \bar{a}_{\cdot,\cdot} = \frac{1}{n^2} \sum_{k,l=1}^n a_{k,l}, \\ A_{k,l} &= a_{k,l} - \bar{a}_{k,\cdot} - \bar{a}_{\cdot,l} + \bar{a}_{\cdot,\cdot}. \end{aligned} \quad (2)$$

In a similar manner, we can define $B_{k,l}$ for Y , where $B_{k,l} = b_{k,l} - \bar{b}_{k,\cdot} - \bar{b}_{\cdot,l} + \bar{b}_{\cdot,\cdot}$, and $b_{k,l} = \|Y_k - Y_l\|$. Then, the empirical distance covariance and empirical variances can be expressed as:

$$\mathcal{V}_n^2(X, Y) = \frac{1}{n^2} \sum_{k,l=1}^n A_{k,l}B_{k,l}. \quad (3)$$

$$\mathcal{V}_n^2(X, X) = \frac{1}{n^2} \sum_{k,l=1}^n A_{k,l}^2, \quad \mathcal{V}_n^2(Y, Y) = \frac{1}{n^2} \sum_{k,l=1}^n B_{k,l}^2. \quad (4)$$

The Divergence Loss. We derive the divergence loss \mathcal{L}_{div} based on the dCor. We randomly sample n_s vertices from the concordant and discrepant subgraphs, $n_s \leq \min(c_e, c_p)$. The features of these sampled vertices are denoted as $S_{\text{con}}, S_{\text{dis}} \in \mathbb{R}^{n_s \times d}$ respectively. The divergence loss \mathcal{L}_{div} based on dCor is formulated as:

$$\mathcal{L}_{\text{div}} = \frac{\mathcal{V}_{n_s}^2(S_{\text{con}}, S_{\text{dis}})}{\sqrt{\mathcal{V}_{n_s}^2(S_{\text{con}}, S_{\text{con}})\mathcal{V}_{n_s}^2(S_{\text{dis}}, S_{\text{dis}}) + \epsilon}}, \quad (5)$$

where ϵ is a small positive constant employed to prevent division by zero, and $\mathcal{V}_{n_s}^2(S_{\text{con}}, S_{\text{dis}})$ is the empirical distance covariance following (3), $\mathcal{V}_{n_s}^2(S_{\text{con}}, S_{\text{con}}), \mathcal{V}_{n_s}^2(S_{\text{dis}}, S_{\text{dis}})$ are the empirical variances following (4).

The divergence loss \mathcal{L}_{div} is designed to encourage the model to capture diverse representations of the concordant and discrepant subgraphs, thereby enhancing the generalizability of the model.

D COMPARISON OF CMMGD FRAMEWORK AND DISENTANGLEMENT-BASED METHODS

This section presents a comparison of the proposed CMMGD framework with disentanglement-based methods to demonstrate its novelty and effectiveness.

Commonalities Towards Generalizability. Previous studies have utilized disentanglement-based methods to tackle the issue of generalization [3, 4, 18]. These methods operate under the assumption that humans encode stimuli in a compositional manner, utilizing a small set of independent and primitive features [18]. This insight is particularly relevant in an emotion recognition task,

Algorithm 1: The CMMGD framework

Data: The EEG signals $X_e \in \mathbb{R}^{c_e \times T}$, the position of each EEG channel under the 10-20 system $\omega_e \in \mathbb{R}^{c_e \times 3}$, the weight of EEG spatial encoding λ_Ω ; the PPS signals $X_p \in \mathbb{R}^{c_p \times T}$; the number of concordant channels ρ ; the hidden dimension d ; the number of MGRT layers L_d ; the number of GRT layers L_f .

Result: The predicted emotional states $\hat{\mathcal{Y}}$.

```

1 begin
  // 1. Get MGRT backbone embedded features (Section 4.1).
2   $H_e^{(L_d)}, A_e, H_p^{(L_d)}, A_p \leftarrow \text{MGRT}(X_e, \omega_e, \lambda_\Omega, X_p);$  // Apply the MGRT backbone, following Algorithm 2.
  // 2. Decompose the multi-modality mixed graph to concordant and discrepant subgraphs (Section 4.2).
  // 2.1 Learn multi-modality mixed graph, following (7).
3  Calculate  $Z \leftarrow \Gamma_{:,0}$ , where every  $\Gamma_{ij} \leftarrow \text{Softmax}\left(\frac{p_{ij}+g}{\tau}\right)$ , and  $p_{ij} \leftarrow [H_{e,i}^{(L_d)} \| H_{p,j}^{(L_d)}] W_c + b_c$ ;
4  Get  $A_{\text{mixed}}$  by combining  $A_e, A_p, \Gamma_{ij}$ , and  $X_{\text{mixed}} \leftarrow [H_e^{(L_d)} \| H_p^{(L_d)}]$ ; // Combine the EEG and PPS graph.
  // 2.2 Graph decomposition.
5   $\text{idx}_{\text{con}} \leftarrow [\text{argsort}(\xi_e)_{1:\rho c_e} \| \text{argsort}(\xi_p)_{1:\rho c_p}]$ ;
6   $\text{idx}_{\text{dis}} \leftarrow [\text{argsort}(\xi_e)_{\rho c_e+1:c_e} \| \text{argsort}(\xi_p)_{\rho c_p+1:c_p}]$ ; // Ranking, following (9).
7   $X_{\text{con}} = X_u[\text{idx}_{\text{con}}, :], A_{\text{con}} = A_u[\text{idx}_{\text{con}}, \text{idx}_{\text{con}}]$ ;
8   $X_{\text{dis}} = X_u[\text{idx}_{\text{dis}}, :], A_{\text{dis}} = A_u[\text{idx}_{\text{dis}}, \text{idx}_{\text{dis}}]$ ; // Decompose the mixed graph, following (10).
  // 3. Cross-rebalance fusion (Section 4.3).
  // 3.1 Get GRT embedded features for each subgraph.
9   $H_{\text{con}}^{(L_f)} \leftarrow \text{GRT}(X_{\text{con}}, A_{\text{con}}, X_{\text{con}}, A_{\text{dis}})$ ;
10  $H_{\text{dis}}^{(L_f)} \leftarrow \text{GRT}(X_{\text{dis}}, A_{\text{dis}}, X_{\text{dis}}, A_{\text{dis}})$ ; // Apply the GRT backbone, following Algorithm 3.
  // 3.2 Cross-rebalance mechanism, following (15).
11 Calculate  $\alpha \leftarrow \frac{1}{2}(\psi_1 + \psi_2)$ , where  $\psi_1 \leftarrow \tanh(f_{\text{con}}^g(R_{\text{con}}))$ , and  $\psi_2 \leftarrow \tanh(f_{\text{dis}}^g(R_{\text{dis}}))$ ;
12  $R_{\text{con}} \leftarrow \omega(H_{\text{con}}^{(L_f)})$ , and  $R_{\text{dis}} \leftarrow \omega(H_{\text{dis}}^{(L_f)})$ ;
13  $\hat{\mathcal{Y}} \leftarrow \text{Softmax}(\alpha_1 f_{\text{con}}^c(\alpha_1 R_{\text{con}}) + \alpha_2 f_{\text{dis}}^c(\alpha_2 R_{\text{dis}}))$ ; // Cross-rebalance fusion, following (13).
14 return  $\hat{\mathcal{Y}}$ 

```

given that the human brain encompasses both primitive and complex features, which are abstract and high-level. Disentanglement methods aim to learn representations that capture various factors of variation in latent subspaces. The learned compositional structure improves the interpretability and generalizability of the model, supporting more difficult forms of generalization.

CMMGD vs. Disentanglement-Based Methods. The proposed CMMGD framework shares commonalities with disentanglement methods, as it decomposes multimodal signals into concordant and discrepant subgraphs, enhancing the generalizability of emotion recognition tasks. However, the CMMGD framework diverges from disentanglement-based methods in several key aspects.

Firstly, **CMMGD is specifically designed for multimodal scenarios.** Disentanglement methods often realize the disentanglement operation based on the unimodal tensor splitting. For example, let E denotes a unimodal embedding, Dong et al. [3] decompose E into E_s and E_c , where $E = [E_s | E_c]$. The E_s is usually called the modality-shared representation, while E_c is the modality-specific representation. They adopt an auxiliary loss for forcing E_s to be consistent across modalities. In contrast, the proposed CMMGD

adopts the correlation-driven decomposition strategy, which determines the decomposition based on the score measuring if one channel is close to the other. This strategy is naturally designed for multimodal scenarios involving two or more heterogeneous data types.

Secondly, **CMMGD is specifically designed for emotion recognition task.** As graphs have been widely adopted in analyzing physiological signals, the proposed framework primarily focuses on these graph structure data, decomposing the multi-modality mixed graph to concordant and discrepant subgraphs. It realizes the vertex-level graph decomposition. In contrast to the widely utilized graph disentanglement [16, 28], which considers more on the edge-level disentanglement, the attribute graph in emotion recognition task through physiological signals has a more significance on vertex attribute, which is the EEG or PPS features. In addition, since the human emotion process usually involves a small set of brain regions [1], the proposed CMMGD assigns each channel to the concordant or discrepant subgraph based on the correlation score, which is more interpretable and meaningful in the emotion recognition task.

Algorithm 2: The MGRT backbone

Data: EEG signals $X_e \in \mathbb{R}^{c_e \times T}$, position of each EEG channel $\omega_e \in \mathbb{R}^{c_e \times 3}$, the weight of EEG spatial encoding λ_Ω ; PPS signals $X_p \in \mathbb{R}^{c_p \times T}$; The number of MGRT layers L_d .

Result: EEG graph with vertex features $H_e^{(L_d)}$, and adjacent matrix A_e ; PPS graph with vertex features $H_p^{(L_d)}$, and adjacent matrix A_p .

```

1 begin
  // Construct the EEG and PPS prior edges.
2   Calculate  $\Omega_e$ , where every  $\Omega_{e,i,j} \leftarrow \|\omega_{e,i} - \omega_{e,j}\|_2$ ;
3   Calculate  $\Phi_e$ , where every  $\Phi_{e,i,j} \leftarrow \text{MI}(X_i, X_j)$ . And the adjacent matrix  $A_e \leftarrow \lambda_\Omega \Omega + (1 - \lambda_\Omega) \Phi_e$ ; // EEG edges.
4   Calculate  $\Phi_p$ , where every  $\Phi_{p,i,j} \leftarrow \text{MI}(X_i, X_j)$ . And the adjacent matrix  $A_p \leftarrow \Phi_p$ ; // PPS edges.
  // Temporal embedding.
5    $H_e^{(0)} = f_e^t(X_e)$ , and  $H_p^{(0)} = f_p^t(X_p)$ ; // Apply the one-dimensional convolutional layers.
  // MGRT layers.
6   for  $\ell \leftarrow 1$  to  $L_d$  do
    // Multi-modality self-attention module.
7      $H_{ep,\text{att}}^{(\ell)} \leftarrow [H_e^{(\ell-1)} \parallel H_p^{(\ell-1)}]$ ;
8      $H_{ep,\text{att}}^{(\ell)} \leftarrow \text{Norm}(H_{ep,\text{att}}^{(\ell)} + \text{Self-Attention}(H_{ep,\text{att}}^{(\ell)}))$ ; // Multi-heads self-attention, following (4).
9     Split  $H_{e,\text{att}}^{(\ell)} \leftarrow (H_{ep,\text{att}}^{(\ell)})_{1:c_e}$ , and  $H_{p,\text{att}}^{(\ell)} \leftarrow (H_{ep,\text{att}}^{(\ell)})_{c_e:c_e+c_p}$ ;
    // EEG and PPS graph regularization modules.
10     $H_{e,\text{gr}}^{(\ell)} \leftarrow \text{Norm}(\text{GR}(H_e^{(\ell-1)}, A_e))$ , and  $H_{p,\text{gr}}^{(\ell)} \leftarrow \text{Norm}(\text{GR}(H_p^{(\ell-1)}, A_p))$ ; // Graph regularization, following (5).
    // Integrating MMSA and GR.
11     $H_e^{(\ell)} \leftarrow \sigma(f_e^{\Theta_{\text{fin}}}(H_{e,\text{gr}}^{(\ell)} + H_{e,\text{att}}^{(\ell)}))$ , and  $H_p^{(\ell)} \leftarrow \sigma(f_p^{\Theta_{\text{fin}}}(H_{p,\text{gr}}^{(\ell)} + H_{p,\text{att}}^{(\ell)}))$ ; // Feed-forward, following (6).
12  return  $H_e^{(L_d)}, A_e, H_p^{(L_d)}, A_p$ 

```

E ALGORITHM OF CMMGD FRAMEWORK

This section outlines the algorithm of the proposed CMMGD framework, which is designed to enhance the generalizability of multi-modal emotion recognition across subjects. The algorithm is detailed in Algorithm 1.

We further provide the details of the MGRT backbone and GRT modules in Algorithm 2 and Algorithm 3, respectively.

Algorithm 3: The GRT modules

Data: The initial vertex features $H^{(0)}$, and the adjacent matrix A ; the number of GRT layers L_f .

Result: The embedded vertex features $H^{(L_f)}$.

```

1 begin
2   for  $\ell \leftarrow 1$  to  $L_f$  do
    // Self-attention module.
3      $H_{\text{att}}^{(\ell)} \leftarrow \text{Norm}(H_{\text{att}}^{(\ell)} + \text{Self-Attention}(H_{\text{att}}^{(\ell)}))$ ;
    // Graph regularization module.
4      $H_{\text{gr}}^{(\ell)} \leftarrow \text{Norm}(\text{GR}(H^{(\ell-1)}, A))$ ;
5      $H^{(\ell)} \leftarrow \sigma(f^{\Theta_{\text{fin}}}(H_{\text{gr}}^{(\ell)} + H_{\text{att}}^{(\ell)}))$ ;
6   return  $H^{(L_f)}$ 

```

F THE DETAILS OF DATASETS

This section introduces the two benchmark datasets in this study: the DEAP [9] and the MAHNOB-HCI [23]. Both of them contain physiological signals of two modalities. Table 1 briefly describes the two datasets.

DEAP. The DEAP dataset is a multimodal dataset for analyzing human affective states [9]. It comprises 32-channel EEG signals and 8-channel PPS, encompassing GSR, blood volume pressure, respiration pattern, skin temperature, EMG (two-channel), and EOG (two-channel). These signals were recorded from 32 participants while viewing 40 one-minute-long music video excerpts, each preceded by an additional 3-second pre-trial baseline signal. Following each trial, participants report their emotional state regarding arousal, valence, dominance, and preference, using nine discrete levels for each dimension.

We employ the official preprocessed data, which involves applying bandpass frequency filtering, eliminating EOG artifacts, and downsampling EEG signals and PPS to 128 Hz.

MAHNOB-HCI. The MAHNOB-HCI dataset is created for emotion recognition and implicit tagging research [23]. It contains 32-channel EEG signals and 6-channel PPS, including GSR, ECG (three-channel), respiration pattern, and skin temperature. The physiological signals are recorded when participants watch 20 emotional video excerpts between 34.9 and 117 seconds long (the mean is 81.4 seconds, and the standard deviation is 22.5 seconds).

The full MAHNOB-HCI dataset contains 30 participants. However, as instructed by [23], the 9th, 12th, and 15th persons are deleted since their original data are not intact. The data is collected at 256 Hz, and we downsample the signals to 128 Hz to keep consistency with the DEAP dataset. The first 30 seconds before the start of each trial are used as the baseline signal.

G THE DETAILS OF IMPLEMENTATION

Data Preprocessing. We downsample the EEG and PPS signals to 128 Hz to ensure consistency across datasets. For MAHNOB-HCI dataset, we apply a bandpass filter to the EEG signals within the range of 0.5-45 Hz, the same as the DEAP dataset. The baselines of the signals are removed by subtracting the mean value of baselines from the entire signal [6, 29]. The signals are then normalized to have zero mean and unit variance. To increase the number of data, we segment the signals into four-second non-overlapping windows to encourage the model to capture short-term dynamics. This is consistent with the approach by Liu et al. [14].

Experimental Setup. We employ the AdamW optimizer [15] with a learning rate of $1e-4$ and a weight decay term of $2e-1$. The training epochs are set to 180 for DEAP and 300 for MAHNOB-HCI. The batch size is set to 20 for each subject during training, which means the actual batch size is $20 \times (32 - 1) = 620$ samples for DEAP, and $20 \times (27 - 1) = 520$ samples for MAHNOB-HCI. We use the PyTorch library to implement the model, and the model can be run on a single NVIDIA RTX 4090 GPU with 24GB memory. Regarding the losses, we assign \mathcal{L}_{div} , \mathcal{L}_{aln} , and \mathcal{L}_{cross} values of $4e-1$, $2e-1$, and $1e-1$, respectively. We incorporate the Leave-One-Subject-Out (LOSO) cross-validation technique, and the evaluation metrics include accuracy and F1 score reported on the excluded subject.

The remaining hyperparameters are determined based on the sensitivity analysis in Section 5.4. We adopt the two types of basic data augmentation, including cropping the raw signals and multiplying the signals by a random factor sampled from a uniform distribution between 0 and 2. The augmentation ratio in Section 5.4 means the ratio of the augmented channels in each sample.

H THE DETAILS OF BASELINE METHODS

This section provides an introduction to each comparative method in Section 5.2. These methods include seven unimodal methods which solely adopt the EEG modality: LSVM-GSU [26], ACRNN [25], BiDANN [11], EEGFuseNet [12], Liu et al. [13], TMLP+SRDANN [10], TSception [2], and CAFNet [32]. We further make a comparison with seven multimodal methods that consider both EEG and PPS modalities. These methods contain RBM [22], MIL [21], MMResLSTM [17], RDFKM [30], CSDAMER [5], and EmotionKD [14].

Baseline Methods in Comparison of CMMGD. We present details of comparison methods in Table 2 and Table 3. The following are the details of methods solely adopting the EEG modality:

- **LSVM-GSU** [26]: The linear SVM-GSU (LSVM-GSU) proposes a maximum margin classifier that deals with uncertainty in data input. It conducts a leave-one-trial-out cross-validation to evaluate the model.

- **ACRNN** [25]: The ACRNN proposes the attention-based convolutional recurrent neural network to extract discriminative features from EEG signals. It adopts a channel-wise attention mechanism to adaptively assign different channels weights, and a convolutional neural network (CNN) to capture the spatial information. The temporal information is captured by the recurrent neural network (RNN).
- **BiDANN** [11]: The BiDANN introduces a global and two local domain discriminators that work adversarially to learn discriminative emotional features for each brain hemisphere. This model is inspired by the neuroscience findings that the left and right hemispheres of the human brain are asymmetric.
- **EEGFuseNet** [12]: The EEGFuseNet proposes a practical hybrid unsupervised deep convolutional recurrent generative adversarial network to learn generic and independent EEG features.
- **Liu et al.** [13]: Liu et al. [13] employ coordinate attention to endow the input signal with relative spatial information and then maps the EEG signal to higher dimensional space. A double-layer capsule network is constructed to utilize the relative location information of EEG.
- **TMPL+SRDANN** [10]: The TMPL+SRDANN designs the transposition multi-layer perceptron (TMPL) and sample-reweighted domain adaptation neural network (SRDANN) in one learning framework, attempting to learn more domain-invariant and class-discriminative EEG features.
- **TSception** [2]: The TSception model proposes to learn the temporal dynamics and spatial asymmetry of EEG signals by employing a multi-scale convolutional neural network.
- **CAFNet** [32]: The CAFNet develops a self-attention-based multi-channel long-short-term memory (LSTM) network and a confidence regression network to estimate true class probability.

The following are the details of methods considering both EEG and PPS modalities, leveraging multimodal fusion techniques:

- **RBM** [22]: This method proposes using a restricted Boltzmann machine to model the inherent dependencies among multimodal physiological signals. A support vector machine is adopted to recognize emotional states.
- **MIL** [21]: The MIL proposes a multiple instance learning-based framework to model time intervals by capturing the presence or absence of relevant states without the need to label the affective responses continuously, which is a crucial challenge in real-life applications.
- **MMResLSTM** [17]: The MMResLSTM introduces a multimodal residual LSTM, sharing the weights across the modalities in each LSTM layer to learn the correlation between the EEG and PPS.
- **RDFKM** [30]: The RDFKM constructs ensemble dense embeddings of multimodal features using kernel matrices and then utilizes a deep network architecture to learn task-specific representations of multi-modality signals.

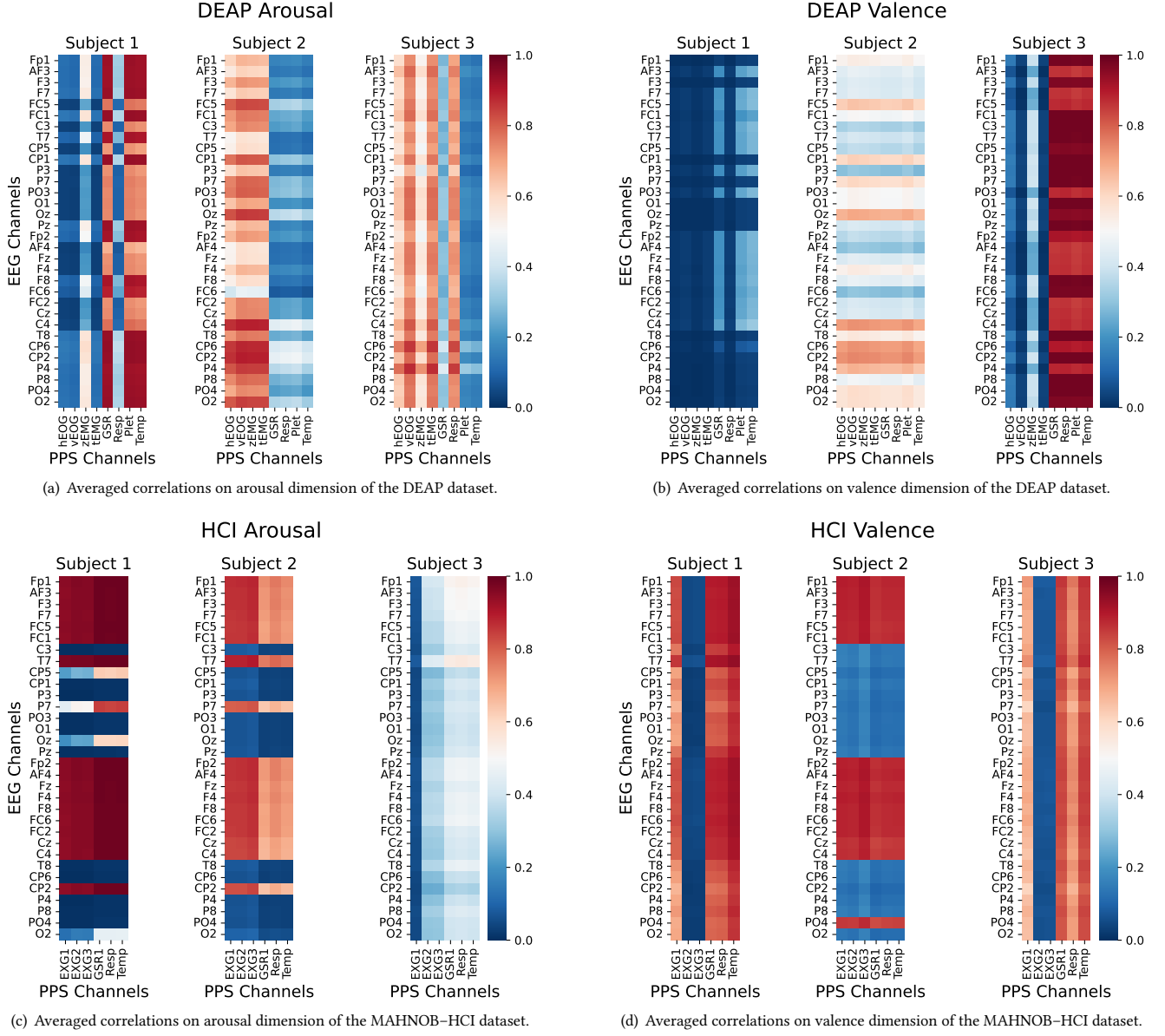


Figure 3: The visualization of averaged cross-modality correlations on the arousal and valence dimensions.

- **CSDAMER** [5]: The CSDAMER method realizes multimodal emotion recognition using CNN-SVM and data augmentation to realize multimodal emotion recognition. The performance of this model is borrowed from Liu et al. [14].
- **EmotionKD** [14]: The EmotionKD conducts cross-modal knowledge distillation that simultaneously models the heterogeneity and interactivity of EEG signals and PPS under a unified framework.

The comparison between the proposed CMMGD and these methods is conducted on the DEAP and MAHNOB-HCI datasets in

Section 5.2, and CMMGD has demonstrated superior performance over these methods.

Baseline Methods in Comparison of MGRT Backbone. We proceed to introduce the comparative methods in Table 4. In this experiment, we replace the MGRT backbone with the following methods and conduct the model training following the same settings and data preprocessing as the CMMGD to get a fair comparison. The following are the details of the two adopted graph convolutional network-based methods:

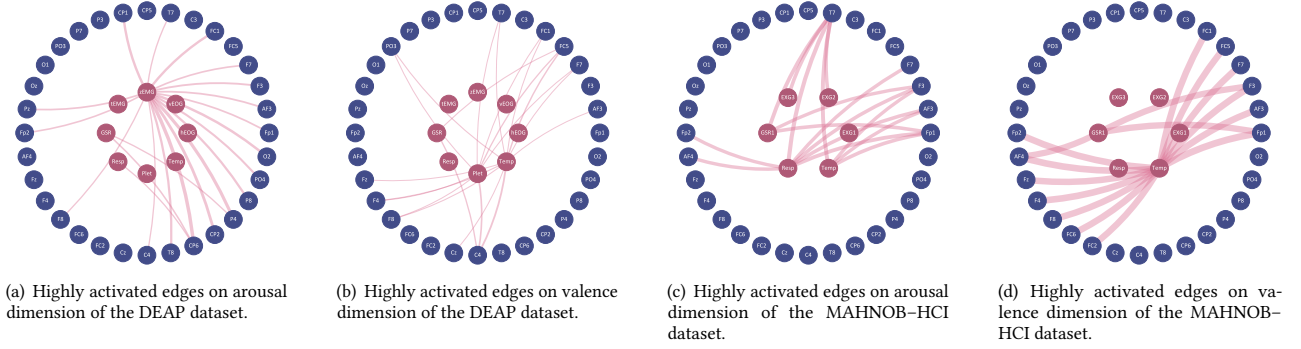


Figure 4: The visualization of highly activated edges on the arousal and valence dimensions. The wider line indicates the stronger correlation between the two channels.

- **GCN** [8]: The GCN layer is a benchmark graph neural network layer realizing the graph convolution operation. It is widely used in graph-based tasks.
- **GraphConv** [19]: The GraphConv layer is a powerful generalization of the graph neural network, taking higher-order graph structures at multiple scales into account. It has been demonstrated to be the benchmark in graph-level representation tasks.

We also replace the MGRT backbone with the Transformer [27], which is popular in a variety of sequence-based tasks. Subsequently, as the combination of the Transformer and graph-based methods, namely the graph transformer, we replace the MGRT backbone with the following graph transformer variants:

- **GraphGPS*** [20]: The GraphGPS aims to build a general, powerful, scalable graph transformer and provides a modular graph transformer framework. The original GraphGPS can handle the unimodal graph, and we extend it to the multimodal emotion recognition task by employing different graph convolutional networks and different feed-forward networks for EEG and PPS modalities, respectively. We name the extended model as GraphGPS*.
- **EmoGTs** [7]: The EmoGTs is the latest graph transformer for emotion recognition. It builds an elastic graph transformer network leveraging the Transformer for time series analysis and graph convolutional networks for topological analysis.

The comparison between the proposed MGRT backbone and these methods is conducted on the DEAP and MAHNOB-HCI datasets in Section 5.2, and the MGRT backbone has demonstrated superior performance over these methods.

I VISUALIZATION OF CROSS-MODALITY CORRELATION PATTERNS

Cross-Modality Correlation. This section visualizes the cross-modality correlation on the arousal and valence dimensions of the DEAP and MAHNOB-HCI dataset, shown in Figure 3, respectively. We select the first three subjects in both datasets and plot the cross-modality correlation scheme averaged of the correlation scores across all samples of the left-out subject. It is observed that the cross-modality correlation patterns are diverse. For instance, on the

valence dimension of the DEAP dataset, the GSR, Resp, Plet, and Temp channels of PPS modality from the 3rd subject have grand connections with EEG channels. However, their connections are weak considering the 1st and 2nd subjects. This indicates that the proposed CMMGD framework can adaptively capture the cross-modality correlation patterns of different subjects. Furthermore, the cross-modality correlation between the arousal and valence dimensions of the same subject is diverse. This indicates that the arousal and valence dimensions are not always positively correlated. The two dimensions are independent and should be considered separately in the emotion recognition task.

Highly Activated Correlations. We further visualize the highly activated cross-modality correlation on the arousal and valence dimensions of the DEAP and MAHNOB-HCI dataset, shown in Figure 4, respectively. The highly activated correlation contains the top 8% correlations across the first three subjects in both datasets. The line width of the edge represents the strength of the correlation, which is calculated by the correlation score. The wider line indicates the stronger correlation between the two channels. Visualizing the highly activated edges provides insights into the most crucial cross-modality correlation patterns in the emotion recognition task. These highly activated edges are inconsistent across different emotion dimensions, indicating the heterogeneous and complex nature of the emotion recognition task. The proposed CMMGD framework can effectively capture the essential cross-modality correlation patterns, and the subsequent graph decomposition and fusion mechanism based on the multi-modality correlation can further enhance the generalizability of emotion recognition across subjects.

J VISUALIZATION OF BRAIN ACTIVATION

This section provides the visualization of brain activation in the DEAP and MAHNOB-HCI datasets, shown in Figure 5. We visualize the activation of each channel in the EEG signals, and the activation is averaged on all samples of the left-out subject during the validation process. The activation value is the average of $H_e^{(L_d)}$ across all samples of the left-out subject. The brain activation topographic map is the interpolation of the activation values across the 32 channels. Each channel is placed at the corresponding position

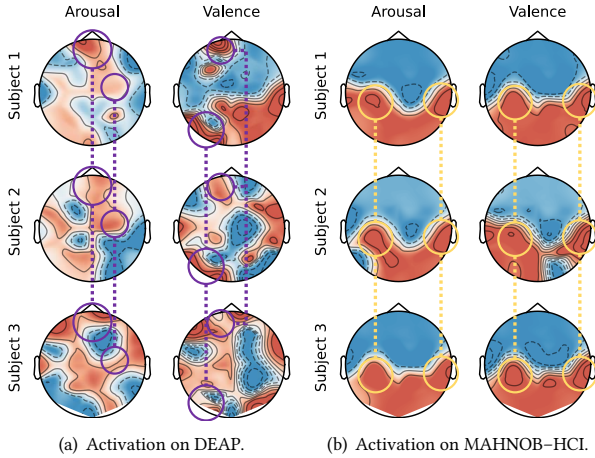


Figure 5: The visualization of brain activation in the DEAP and MAHNOB-HCI datasets.

on the 10-20 system. The high activation of the brain region is represented by the red color, while the low activation is represented by the blue color. The brain activation topographic maps are averaged on all samples on the left-out subject. The circles represent the partially found consistent brain regions across subjects.

In Figure 5, we mark partially consistent highly activated brain regions across subjects with circles, indicating that the proposed CMMGD framework can effectively capture the brain activation patterns. Specifically, considering the 1st and 2nd subjects, they have grand similar activation patterns on the arousal dimension. On the other hand, the 2nd and 3rd subjects have grand similar activation patterns on the valence dimension. Moreover, the activation patterns on the MAHNOB-HCI dataset are consistent across all three subjects, which indicates the robustness of the proposed CMMGD framework.

REFERENCES

- [1] Richard J Davidson and Nathan A Fox. 1982. Asymmetrical brain activity discriminates between positive and negative affective stimuli in human infants. *Science* 218, 4578 (1982), 1235–1237.
- [2] Yi Ding, Neethu Robinson, Su Zhang, Qiuhaio Zeng, and Cuntai Guan. 2023. TSception: Capturing Temporal Dynamics and Spatial Asymmetry From EEG for Emotion Recognition. *IEEE Transactions on Affective Computing* 14, 3 (July 2023), 2238–2250.
- [3] Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink. 2023. SimMMDG: A Simple and Effective Framework for Multi-Modal Domain Generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [4] Christina M. Funke, Paul Vicol, Kuan-Chieh Wang, Matthias Kuemmerer, Richard Zemel, and Matthias Bethge. 2022. Disentanglement and Generalization Under Correlation Shifts. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*.
- [5] Gengyuan Guo, Pengzhi Gao, Xiangwei Zheng, and Cun Ji. 2022. Multimodal Emotion Recognition Using CNN-SVM with Data Augmentation. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 3008–3014.
- [6] Ziyu Jia, Youfang Lin, Jing Wang, Zhiyang Feng, Xiangheng Xie, and Caijie Chen. 2021. HetEmotionNet: Two-Stream Heterogeneous Graph Recurrent Neural Network for Multi-modal Emotion Recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 1047–1056.
- [7] Wei-Bang Jiang, Xu Yan, Wei-Long Zheng, and Bao-Liang Lu. 2023. Elastic Graph Transformer Networks for EEG-Based Emotion Recognition. In *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5.
- [8] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.
- [9] Sander Koelstra, Christian Muehl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A Database for Emotion Analysis Using Physiological Signals. *IEEE Trans. Affect. Comput.* 3, 1 (2012), 18–31.
- [10] Wei Li, Bowen Hou, Xiaoyu Li, Ziming Qiu, Bo Peng, and Ye Tian. 2023. TMLP+SRDANN: A Domain Adaptation Method for EEG-based Emotion Recognition. *Measurement* 207 (Feb. 2023), 112379.
- [11] Yang Li, Wenming Zheng, Yuan Zong, Zhen Cui, Tong Zhang, and Xiaoyan Zhou. 2018. A bi-hemisphere domain adversarial neural network model for EEG emotion recognition. *IEEE Transactions on Affective Computing* 12, 2 (2018), 494–504.
- [12] Zhen Liang, Rushuang Zhou, Li Zhang, Linling Li, Gan Huang, Zhiguo Zhang, and Shin Ishii. 2021. EEGFuseNet: Hybrid Unsupervised Deep Feature Characterization and Fusion for High-Dimensional EEG With an Application to Emotion Recognition. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29 (2021), 1913–1925.
- [13] Shuaiqi Liu, Zeyao Wang, Yanling An, Jie Zhao, Yingying Zhao, and Yu-Dong Zhang. 2023. EEG Emotion Recognition Based on the Attention Mechanism and Pre-Trained Convolution Capsule Network. *Knowledge-Based Systems* 265 (April 2023), 110372.
- [14] Yucheng Liu, Ziyu Jia, and Haichao Wang. 2023. EmotionKD: A Cross-Modal Knowledge Distillation Framework for Emotion Recognition Based on Physiological Signals. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. Association for Computing Machinery, New York, NY, USA, 6122–6131.
- [15] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [16] Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. 2019. Disentangled Graph Convolutional Networks. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 4212–4221.
- [17] Jiaxin Ma, Hao Tang, Wei-Long Zheng, and Bao-Liang Lu. 2019. Emotion Recognition Using Multimodal Residual LSTM Network. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. Association for Computing Machinery, New York, NY, USA, 176–183.
- [18] Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. 2020. The Role of Disentanglement in Generalisation. In *International Conference on Learning Representations*.
- [19] Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. 2019. Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI'19)*. AAAI Press, Honolulu, Hawaii, USA, 4602–4609.
- [20] Ladislav Rampasek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. 2022. Recipe for a General, Powerful, Scalable Graph Transformer. In *Advances in Neural Information Processing Systems*.
- [21] Luca Romeo, Andrea Cavallo, Lucia Pepa, Nadia Bianchi-Berthouze, and Massimiliano Pontil. 2022. Multiple Instance Learning for Emotion Recognition Using Physiological Signals. *IEEE Transactions on Affective Computing* 13, 1 (Jan. 2022), 389–407.
- [22] Yangyang Shu and Shangfei Wang. 2017. Emotion Recognition through Integrating EEG and Peripheral Signals. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2871–2875.
- [23] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2012. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Trans. Affect. Comput.* 3, 1 (2012), 42–55.
- [24] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. 2007. Measuring and testing dependence by correlation of distances. *The Annals of Statistics* 35, 6 (2007), 2769–2794.
- [25] Wei Tao, Chang Li, Rencheng Song, Juan Cheng, Yu Liu, Feng Wan, and Xun Chen. 2023. EEG-Based Emotion Recognition via Channel-Wise Attention and Self Attention. *IEEE Transactions on Affective Computing* 14, 1 (Jan. 2023), 382–393.
- [26] Christos Tzelepis, Vasileios Mezaris, and Ioannis Patras. 2018. Linear Maximum Margin Classifier for Learning from Uncertain Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 12 (Dec. 2018), 2948–2962.
- [27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [28] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-Refined Convolutional Network for Multimedia Recommendation with Implicit Feedback. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*. Association for Computing Machinery, New York, NY, USA, 3541–3549.
- [29] Yilong Yang, Qingfeng Wu, Ming Qiu, Yingdong Wang, and Xiaowei Chen. 2018. Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network. In *2018 international joint conference on neural networks (IJCNN)*. IEEE, 1–7.

- [30] Xiaowei Zhang, Jinyong Liu, Jian Shen, Shaojie Li, Kechen Hou, Bin Hu, Jin Gao, Tong Zhang, and Bin Hu. 2021. Emotion Recognition From Multimodal Physiological Signals Using a Regularized Deep Fusion of Kernel Machine. *IEEE Transactions on Cybernetics* 51, 9 (Sept. 2021), 4386–4399.
- [31] Xingjian Zhen, Zihang Meng, Rudrasis Chakraborty, and Vikas Singh. 2022. On the versatile uses of partial distance correlation in deep learning. In *European Conference on Computer Vision*. Springer, 327–346.
- [32] Qi Zhu, Chuhang Zheng, Zheng Zhang, Wei Shao, and Daoqiang Zhang. 2023. Dynamic Confidence-Aware Multi-Modal Emotion Recognition. *IEEE Transactions on Affective Computing* (2023), 1–13.