A Cross-Modal Densely Guided Knowledge Distillation Based on Modality Rebalancing Strategy for Enhanced Unimodal Emotion Recognition

Shuang Wu[1], Heng Liang[2], Yong Zhang[3*], Yanlin Chen[4], Ziyu Jia[5*]
[1]South China University of Technology
[2]The University of Hong Kong
[3]Huzhou University
[4]New York University
[5]Institute of Automation, Chinese Academy of Sciences
frostfree.ws@gmail.com, hengliang@connect.hku.hk, zhyong@zjhu.edu.cn, yc3156@nyu.edu, jia.ziyu@outlook.com

Abstract

Multimodal emotion recognition has garnered significant attention for its ability to integrate data from multiple modalities to enhance performance. However, physiological signals like electroencephalogram are more challenging to acquire than visual data due to higher collection costs and complexity. This limits the practical application of multimodal networks. To address this issue, this paper proposes a cross-modal knowledge distillation framework for emotion recognition. The framework aims to leverage the strengths of a multimodal teacher network to enhance the performance of a unimodal student network using only the visual modality as input. Specifically, we design a prototype-based modality rebalancing strategy, which dynamically adjusts the convergence rates of different modalities to mitigate modality imbalance issue. It enables the teacher network to better integrate multimodal information. Building upon this, we develop a Cross-Modal Densely Guided Knowledge Distillation (CDGKD) method, which effectively transfers knowledge extracted by the multimodal teacher network to the unimodal student network. Our CDGKD uses multi-level teacher assistant networks to bridge the teacher-student gap and employs dense guidance to reduce error accumulation during knowledge transfer. Experimental results demonstrate that the proposed framework outperforms existing methods on two public emotion datasets, providing an effective solution for emotion recognition in modality-constrained scenarios.

1 Introduction

Emotion is a complex psychological and physiological process reflecting an individual's subjective perception of the external environment. It also plays a crucial role in interpersonal communication and human-computer interaction (HCI). With the rapid advancement of HCI technologies, emotion recognition has become a major research focus in affective computing [Zhao *et al.*, 2018], with applications spanning intelligent assistants, educational technologies, and medical diagnostics [Shen *et al.*, 2017]. However, achieving high accuracy in emotion recognition remains challenging due to the inherent diversity and complexity of emotional expressions [Liu *et al.*, 2024]. To address this, researchers have explored various multimodal data sources in recent years, including text, speech, facial expressions, and physiological signals, to improve recognition accuracy [Lian *et al.*, 2023; Udahemuka *et al.*, 2024; Wang *et al.*, 2025]. Among these, the visual modality is widely used due to its ease of acquisition and rich temporal dynamics [Canal *et al.*, 2022]. Complementing the visual modality, electroencephalogram (EEG) is a representative physiological signal that reflects the neural activities underlying emotions, offering more objective neural evidence and being less susceptible to intentional manipulation [Jia *et al.*, 2024]. The two modalities exhibit strong complementarity in emotion recognition: the visual modality captures external features of emotional expressions [Liu *et al.*, 2023a], while EEG provides neural evidence of intrinsic emotional states [Ding *et al.*, 2022; Cheng *et al.*, 2024b]. By integrating these two modalities, researchers can leverage synergistic effects, further enhancing recognition performance and advancing HCI development.

However, visual modality data is often easier to acquire in practical applications, whereas physiological signals such as EEG are relatively difficult to collect due to equipment limitations and high collection costs. This limitation hinders emotion recognition systems from fully leveraging the advantages of multimodal fusion in resource-constrained environments. Cross-modal knowledge distillation offers a potential solution to address this issue by transferring knowledge from a multimodal teacher network to a unimodal student network [Gupta *et al.*, 2016]. It compensates for the limited representation capacity of student networks, improving their classification performance with unimodal inputs. However, existing methods are still inadequate for effectively distilling information from multimodal data (e.g., video and physiological signals) into unimodal networks. The key challenge is efficiently ex-

*Corresponding Author

tracting knowledge from multimodal data and transferring it effectively to unimodal networks.

Modality imbalance hinders the effective extraction of knowledge from multimodal data. Theoretically, multimodal data provides multiple perspectives and should outperform unimodal networks [Peng *et al.*, 2022]. However, in some cases, the best-performing unimodal networks even outperform multimodal networks, as shown in Table 1. This phenomenon reflects modality imbalance in multimodal fusion [Wang *et al.*, 2020]. Significant differences in feature representation capabilities and convergence rates between modalities hinder weaker modalities from learning effectively, limiting overall fusion performance [Fan *et al.*, 2023]. To address the modality imbalance problem, existing methods can be broadly categorized into two types. The first category enhances weaker modalities by introducing unimodal supervision or auxiliary modules [Wang *et al.*, 2020; Du *et al.*, 2021; Zhang *et al.*, 2024]. However, these methods heavily depend on unimodal network quality and increase computational overhead. The second category dynamically adjusts the convergence rates of modalities, suppressing stronger modalities to balance the learning contributions between them [Wang *et al.*, 2020; Xiao *et al.*, 2020; Peng *et al.*, 2022]. Nevertheless, this strategy may weaken dominant modalities, which could hinder the overall network performance. Thus, whether by improving weaker modalities or balancing convergence rates, current methods face limitations and struggle to balance enhancing weaker modalities and optimizing overall fusion performance.

Modal	Arousal	Valence
EEG	60.61	61.27
Visual	**61.47**	63.02
Fusion (Visual+EEG)	60.93	**64.50**

Table 1: Unimodal networks can outperform multimodal networks in some cases. Experiments are conducted on the DEAP dataset.

Another challenge is bridging the gap between different networks to facilitate effective cross-modal knowledge transfer, enhancing the performance of unimodal networks. Multimodal teacher networks typically have more complex deep network structures, with multiple input channels and specialized branches for processing features from different modalities, whereas student networks are lightweight and rely on a single modality. Due to the significant differences in network complexity and feature representation capabilities, direct knowledge distillation often results in incomplete knowledge transfer, limiting the performance of the student network [Cho and Hariharan, 2019]. To address this issue, existing methods have attempted to mitigate these structural differences by introducing intermediate auxiliary networks, such as teacher assistant (TA) networks [Mirzadeh *et al.*, 2020]. Assistant-based distillation methods can be broadly categorized into two types. The first is single-level TA distillation, where a single-level TA network is constructed to transfer knowledge from the teacher network to the TA network, which then transfers it to the student network, as shown in Figure 1a. However, a single-level TA network often struggles to bridge the complex gap between teacher and student networks, limiting knowledge transfer effectiveness [Liang *et al.*, 2023]. The second approach is multi-level TA knowledge distillation (MTAKD), which employs multiple TA networks to transfer knowledge step by step from the teacher network to the student network, as shown in Figure 1b. Although this method can alleviate the gap between teacher and student networks at a finer granularity, it may suffer from error accumulation during the progressive training [Son *et al.*, 2021]. Specifically, errors introduced during the knowledge transfer of one level may be further amplified by subsequent networks, thereby affecting the performance of the student network. Designing an effective distillation mechanism that enables the student network to efficiently comprehend the teacher network's knowledge while reducing error propagation during the transfer process remains a crucial challenge.

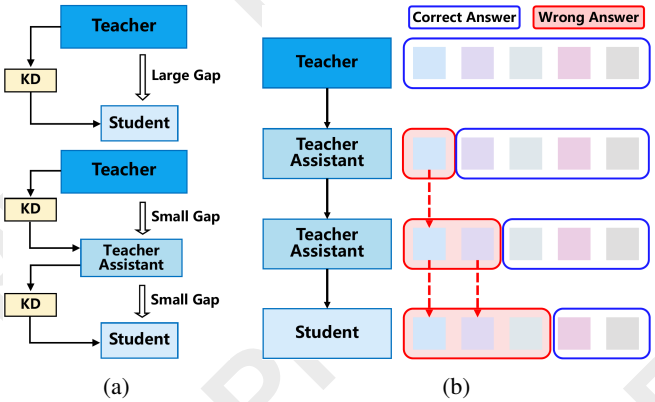

Figure 1: (a) The gap between the teacher and student networks may limit the effectiveness of knowledge distillation. TA network bridge this gap by reducing the differences, enabling more efficient knowledge transfer. (b) The MTAKD method suffers from the problem of "error accumulation": as higher-level TAs transfer knowledge to lower-level TAs, errors accumulate across levels, eventually affecting the learning performance of the student network.

To address the above challenges, this paper proposes an innovative cross-modal knowledge distillation framework for emotion recognition. The framework aims to effectively extract and integrate emotional features through a multimodal fusion strategy and transfer them to a unimodal student network via knowledge distillation, enhancing its performance in emotion recognition task. The main contributions of this paper are summarized as follows:

- We propose a prototype loss-based modality rebalancing strategy to balance the convergence rates between modalities, enabling the teacher network to effectively fuse multimodal features.
- We develop a Cross-Modal Densely Guided Knowledge Distillation (CDGKD) method that bridges the gap between the teacher and student networks while mitigating error accumulation.
- The proposed framework effectively transfers multimodal knowledge to enhance unimodal network performance, achieving state-of-the-art results on two emotion

recognition datasets.

2 Related Works

Multimodal fusion networks have shown considerable advantages over unimodal networks in emotion recognition tasks [Ngiam *et al.*, 2011; Zadeh *et al.*, 2017; Cheng *et al.*, 2025]. For example, Liu et al. [2023b] propose EmotionKD, using a multimodal teacher network to extract heterogeneous and interactive features between EEG and Galvanic Skin Response (GSR) signals. Li et al. [2023] design a Transformer-based audio-visual framework to capture cross-modal correlations, improving emotion intensity estimation and classification. Wu et al. [2023] develop a bionic dual-system architecture that integrates facial expression features with remote physiological signals and leveraging reinforcement learning to achieve efficient recognition of complex emotions. These works show the promise of multimodal fusion for improving adaptability and robustness in varied scenarios.

In the domain of multimodal emotion recognition, the integration of facial video and EEG signals has demonstrated considerable potential due to their complementary characteristics. Facial video, characterized by its intuitiveness and ease of acquisition, effectively captures explicit emotional states through observable facial expressions [Zhang *et al.*, 2022]. Liu et al. [2023a] propose a method based on the Transformer architecture, which makes full use of the rich temporal dynamics in the visual modality to accurately capture the subtle changes in facial expressions. However, the robustness of the visual modality is often compromised by external factors such as lighting conditions, occlusion, and deliberate emotional masking. In contrast, EEG signals objectively reflect implicit emotional states by recording brain activity and are particularly adept at identifying concealed emotional responses. The TSception network employs multi-scale convolution to capture both the temporal dynamics and spatial asymmetry of EEG signals, demonstrating the unique advantages of EEG in implicit emotion recognition [Ding *et al.*, 2022]. Therefore, the combination of visual and EEG signals can leverage their complementary strengths, enabling more comprehensive feature representations and improved robustness in emotion recognition tasks.

Some preliminary studies have explored the integration of visual and EEG modalities [Tan *et al.*, 2021]. For example, Huang et al. [2019] employ a decision-level fusion approach to combine visual and EEG data, effectively enhancing the accuracy of emotion recognition. Saffaryazdi et al. [2022] further introduce the facial micro-expression modality and combine it with EEG and other physiological signals to more comprehensively capture emotional features. Cheng et al. [2024a] propose the Dense Graph Convolutional with Joint Cross-Attention network, leveraging the spatial topology and consistency information between visual and EEG to further improve emotion recognition performance. Despite these advancements, existing approaches often overlook the challenge of modality imbalance.

In addition to the challenges within multimodal fusion, practical constraints such as computational cost and deployment complexity have limited the real-world application of multimodal systems. To address these issues, cross-modal knowledge distillation has emerged as a promising solution [Gupta *et al.*, 2016]. By transferring knowledge from a high-performance multimodal teacher network to a lightweight unimodal student network, this approach can achieve comparable performance with significantly lower resource requirements. For instance, Liu et al. [2023b] demonstrate how interactive knowledge from EEG and GSR signals can be effectively distilled into a unimodal GSR network. Similarly, Aslam et al. [2024] propose aligning structural relationships from multiple teacher networks to improve the adaptability of student networks. However, the effectiveness of this distillation process is constrained by the gap between the teacher and student networks, with the limited capacity of the student network often hindering effective knowledge transfer [Cho and Hariharan, 2019].

3 Methodology

This study proposes a cross-modal knowledge distillation framework designed to enhance visual unimodal emotion recognition task. The framework comprises two key components: a Prototype-Based Modality Rebalancing Strategy for effective emotional feature extraction and multimodal fusion, and a Cross-Modal Densely Guided Knowledge Distillation method to transfer these features to unimodal student networks for improved performance. As shown in Figure 2, the framework first receives paired facial video and EEG data, extracting visual and EEG features through the visual encoder and EEG encoder, respectively. These features are used to calculate the prototype loss and then passed to the classifier to compute the cross-entropy loss after concatenation. Both losses guide the backpropagation process of the multimodal teacher network. Subsequently, the knowledge from the trained teacher network will guide the learning of the TA network and the student network.

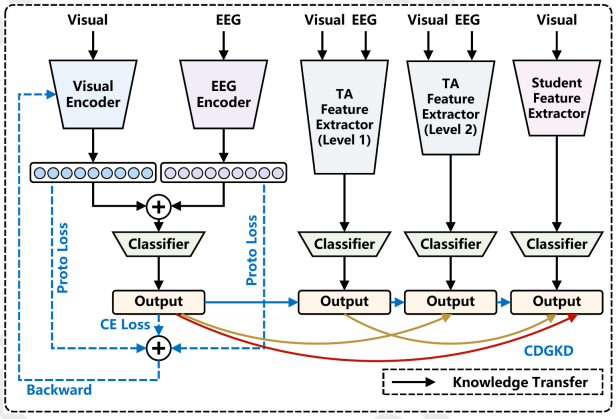

Figure 2: The proposed cross-modal knowledge distillation framework. The teacher network balances the convergence rates of the visual and EEG modalities using prototype loss, generating high-quality multimodal knowledge. This knowledge is gradually transferred to the student network through multi-level TA networks, thereby enhancing the student network's emotion recognition performance using only the visual modality.

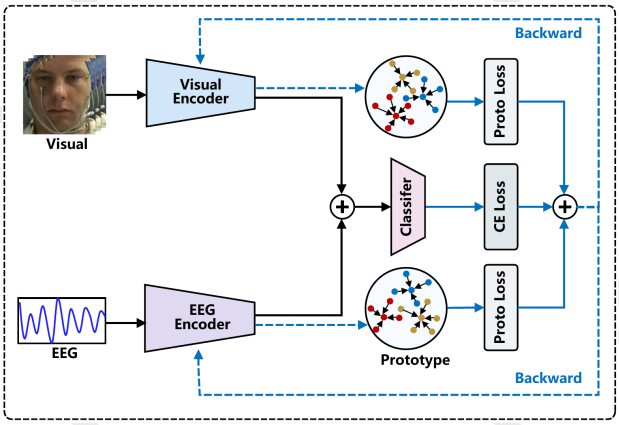

Figure 3: Overview of the multimodal teacher network training process. Visual and EEG inputs are encoded separately, then concatenated and passed to a shared classifier for prediction. Modality-specific features are compared with class prototypes to compute the prototype loss, which encourages intra-class feature compactness and enables the estimation of each modality's convergence rate. The final loss combines cross-entropy loss and prototype losses, guiding the teacher network to learn balanced multimodal representations for later distillation.

3.1 Prototype-Based Modality Rebalancing Strategy

In emotion recognition tasks involving the fusion of visual and EEG modalities, there is a disparity in the feature utilization efficiency between different modalities. Features from the weaker modality are prone to being suppressed, which limits overall performance. To address this issue, a prototype loss-based modality rebalancing strategy is designed, as shown in Figure 3. By dynamically adjusting the modality weights, this approach promotes the convergence of the weaker modality and alleviates the imbalance problem.

A prototype is a feature vector that represents the center of a data class [Snell *et al.*, 2017]. In the embedding space, each class's data points are assumed to cluster around a central point (prototype). For emotion class k, the prototype in the visual modality m_v and EEG modality m_e is defined as the mean of the feature representations of all samples in that class. The definition of the class-wise prototype c_k^m in each modality is given by:

$$c_k^m = \frac{1}{N_k} \sum_{i=1}^{N_k} z_i^m, \quad m \in \{m_v, m_e\} \tag{1}$$

where N_k is the number of samples in class k, and z_i^m is the feature representation of sample x_i in modality m. The class prototype aggregates the sample distribution information, guiding the modality features to converge towards the class center, thereby enhancing the discriminability of emotion classification.

For sample x_i in modality m, its class distribution is determined by the distance between the feature representation z_i^m and the class prototype c_k^m. Specifically, the probability distribution of each class is calculated using the Softmax function:

$$p_i^m(y=k|x_i^m) = \frac{\exp(-d(z_i^m, c_k^m))}{\sum_{k'} \exp(-d(z_i^m, c_{k'}^m))} \tag{2}$$

where $d(\cdot,\cdot)$ is a distance function defined as the Euclidean distance: $d(z_i^m, c_k^m) = \|z_i^m - c_k^m\|^2$. The training objective is to make the sample's predicted result more likely to match the correct class. To achieve this, given N samples, the loss function is defined as the average negative log probability of each sample belonging to its true class k:

$$\mathcal{L}_{proto}^m = \frac{1}{N} \sum_{i=1}^{N} [-\log(p_i^m(y=k|x_i^m))] \tag{3}$$

The class distribution p_i^m not only provides the classification basis for the sample, but the sum of the probabilities for the correct class also measures the sample's aggregation degree. The higher the intra-class compactness of the sample features, the better the discriminability of the features, which in turn reflects the convergence rate within the modality.

Specifically, during training, the convergence rate of modality m is represented by the sum of the classification probabilities of the samples in the current batch:

$$r_t^m = \sum_{i \in B_t^m} p_i^m \tag{4}$$

where B_t^m is the batch of data at training step t, and p_i^m is the probability of sample x_i belonging to its true class in modality m, as calculated by equation (2). This metric is computed using only feature representations and prototypes, which are computationally independent of the fusion method and classifier structure. The final loss function $\mathcal{L}_{teacher}$ for the multimodal network is defined as the weighted sum of cross-entropy loss and prototype loss:

$$\mathcal{L}_{teacher} = (1-\alpha)\mathcal{L}_{CE} + \alpha\left(\lambda_{m_v}\mathcal{L}_{proto}^{m_v} + \lambda_{m_e}\mathcal{L}_{proto}^{m_e}\right) \tag{5}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss, and α controls the contribution of the classification and prototype losses to the overall objective. The values of λ_{m_v} and λ_{m_e} are determined by calculating the ratio of modality convergence rates r_t^m, and they dynamically adjust the convergence rates of the modalities during training:

$$\begin{cases} \lambda_{m_v} = \text{clip}\left(0, \frac{r_t^{me}}{r_t^{mv}} - 1, 1\right), \lambda_{m_e} = 0 & \frac{r_t^{me}}{r_t^{mv}} > 1 \\ \lambda_{m_v} = 0, \lambda_{m_e} = \text{clip}\left(0, \frac{r_t^{mv}}{r_t^{me}} - 1, 1\right) & \frac{r_t^{me}}{r_t^{mv}} \leq 1 \end{cases} \tag{6}$$

where $\text{clip}(a,b,c)$ is a clipping function that restricts b to lie within the range $[a,c]$. This method encourages the slower-learning modality to utilize its prototype features, while the faster modality maintains its original learning strategy, alleviating the modality imbalance issue.

3.2 Cross-Modal Densely Guided Knowledge Distillation

To address the issues of information loss and error accumulation in multi-level knowledge transfer, we propose a Cross-Modal Densely Guided Knowledge Distillation (CDGKD)

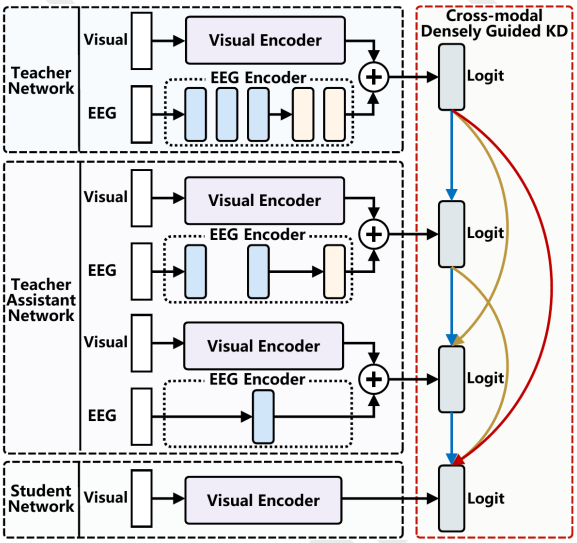

Figure 4: Overview of the CDGKD framework. Each TA or the student network is supervised by a randomly sampled subset from the combined set of higher-level models, including the teacher and previously trained TAs.

method for emotion recognition. As shown in Figure 4, our CDGKD improves unimodal emotion recognition by guiding the student network using knowledge from the teacher and intermediate TA networks during cross-modal transfer from EEG and visual modalities to the visual modality. Specifically, the student is influenced not only by the immediately preceding TA, but also by combined supervision from all previous TAs and the teacher. This strategy mitigates error accumulation by leveraging multiple knowledge sources, improving the student's performance.

The TA networks are derived from the teacher network by simplifying its EEG encoder, while preserving core feature extraction capabilities. The visual encoder remains unchanged, enabling visual information to combine with representations from progressively simplified EEG encoders, providing diverse cross-modal guidance. Since the student only uses the visual modality, TA networks gradually reduce reliance on EEG and strengthen visual guidance. This ensures smooth knowledge transition and enhances student performance on visual tasks.

The entire CDGKD training adopts a level-by-level strategy to ensure clarity and effectiveness in knowledge transfer. Each TA network is trained independently and receives guidance from more complex previous networks (i.e., the teacher or a higher-level TA) via distillation loss. Let T denotes the teacher network, $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ represents the set of TA networks, where A_i is the i-th TA, and S denotes the student network. We define the hierarchical guidance relationships as follows:

$$A_i \leftarrow \{A_j \in \mathcal{A} \mid j < i\} \cup \{T\}, \quad \text{for } i = 1, 2, \ldots, n \tag{7}$$

$$S \leftarrow \{T\} \cup \mathcal{A} \tag{8}$$

For the training process of the i-th TA network A_i, the loss function is defined as:

$$\mathcal{L}_{CDGKD} = (1-\beta)\mathcal{L}_{CE} + \beta \frac{1}{|\mathcal{S}_i|} \sum_j \mathrm{KL}(l_{A_i} \parallel l_{A_{i-j}}) \tag{9}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss used to measure the classification performance of the current network A_i. $\mathcal{S}_i$ represents the training target set $\mathcal{S}_i \subseteq \{l_T, l_{A_1}, \ldots, l_{A_{i-1}}\}$, where l_T and l_{A_k} denote the logits of the teacher network and the k-th level TA network, respectively. $|\mathcal{S}_i|$ represents the size of the logits set. $\mathrm{KL}(l_{A_i} \parallel l_{A_{i-j}})$ represents the Kullback-Leibler divergence between the logits of the current TA network A_i and a selected higher-level network A_{i-j}, $j \in \{1, \ldots, i-1\}$, measuring the distribution difference between them. The coefficient β controls the contribution of the classification loss and distillation loss to the total loss.

The training of the final student network is similar to that of the TA networks, both involving the use of $\mathcal{L}_{CDGKD}$. Through level-by-level training and the stochastic learning strategy, each network in the hierarchy can acquire effective information from the higher-level knowledge while maintaining flexibility, improving the performance of the final unimodal visual student network.

To further enhance the flexibility and robustness of the proposed CDGKD, we introduce a stochastic learning strategy into the knowledge distillation process. Specifically, during the training of each TA or the final student network, knowledge sources are randomly selected from the teacher network T and the set of previous TA networks $\mathcal{A}$, dynamically influencing the training process by varying the knowledge paths. This strategy effectively reduces the network's reliance on specific knowledge paths by dynamically adjusting the knowledge connections, lowering the risk of overfitting.

The stochastic learning strategy irregularly drops certain knowledge connections, reducing the interference from complex teacher or TA output distributions on the student network. This allows the student network to focus on the most relevant and reliable modality knowledge for the emotion recognition task, alleviating overfitting issues due to the simpler structure of the student network. Additionally, this strategy significantly reduces the accumulation of erroneous knowledge during stepwise transfer, enabling the student network to concentrate on learning core and reliable knowledge.

4 Experiments

4.1 Datasets

This study uses two public datasets, DEAP [Koelstra *et al.*, 2011] and MAHNOB-HCI [Soleymani *et al.*, 2011], to evaluate the proposed cross-modal distillation framework.

DEAP dataset is a multimodal dataset that includes EEG signals and facial video data. In the experiment, 32 participants watched 40 one-minute music clips, followed by self-assessments of arousal, valence, liking, and dominance. The EEG signals were recorded at 512 Hz with a 32-channel device, and synchronized facial videos were recorded for 22 participants. Only the data from these 22 participants, who had both EEG and facial video recordings are used.

MAHNOB-HCI dataset is a similar multimodal emotion dataset, containing EEG and front-facing color video data.

The EEG signals were recorded using the Biosemi Active II system with 32 electrodes at 256 Hz. In the experiment, 30 participants watched 20 video clips, followed by self-assessments of arousal, valence, and other emotional dimensions. Data from 24 participants who completed the experiment with both EEG and facial video recordings are used.

4.2 Baselines

To validate the effectiveness of the knowledge distillation framework, we reproduce some novel cross-modal knowledge distillation methods and representative distillation methods in a cross-modal experimental setup as baselines, including KD [Hinton, 2015], Fitnets [Romero *et al.*, 2014], NST [Huang and Wang, 2017], TAKD [Mirzadeh *et al.*, 2020], EmotionKD [Liu *et al.*, 2023b], and AMBOKD [Li *et al.*, 2024]. Additionally, to validate the performance of the multimodal teacher network, we reproduce some novel and well-performing multimodal emotion recognition network architectures as baselines, including CNN+SVM [Huang *et al.*, 2019], CNN+LSTM [Saffaryazdi *et al.*, 2022], EmotionKD [Liu *et al.*, 2023b], and DGC+JCA [Cheng *et al.*, 2024a].

4.3 Experiment Settings

In this experiment, we perform binary classification tasks for two emotion dimensions: valence and arousal. Specifically, we use a rating of 5 (the median on a 9-point scale) as the classification threshold, dividing valence into low valence and high valence, and arousal into low arousal and high arousal. For visual modality data, we first extract one image every 40 frames, representing the visual modality as a sequence of frames. Then, we perform face detection on each frame, crop the face region, and resize the cropped image to 224×224 to prepare it for input into the network. To align with practical application scenarios and ensure cross-experiment independence between different datasets (training, validation, and test sets), we perform the data split at the trial level, ensuring that slices from the same trial do not appear in different datasets. Specifically, the entire dataset is divided into training, validation, and test sets in a 8:1:1 ratio. For each sample in the original data (continuous data of several seconds), we split it into samples every 4 seconds along the time axis, generating short-term sample segments for network training.

In this experiment, the visual modality features are extracted using the Pyramid Vision Transformer (PVT) network [Wang *et al.*, 2021], which captures emotion-related information at different levels by employing multi-scale feature extraction and global context networking. The EEG modality features are extracted using the TSception encoder [Ding *et al.*, 2022], a network specifically designed for EEG signals that can extract emotion-related features from both the temporal and spatial dimensions. With these two specially designed encoders, this method effectively leverages the valuable information from both visual and EEG signals to support emotion recognition tasks.

We use the Adam optimizer with a fixed learning rate of $1e$-4, and employ early stopping during training, which halts the training process if the validation set metric does not improve for 15 consecutive epochs. The evaluation metric is classification accuracy. Training is conducted with a batch size of 128 for up to 100 epochs. The hyperparameter α in the multimodal teacher network's loss function ($\mathcal{L}_{teacher}$) and the hyperparameter β in the distillation loss function ($\mathcal{L}_{CDGKD}$) are both set to 0.5. All experiments are conducted using TensorFlow on NVIDIA RTX 3090 GPUs.

4.4 Results

To validate the effectiveness of the proposed teacher network training strategy and knowledge distillation method, comprehensive experimental evaluations are conducted on both the multimodal teacher network and the unimodal student network guided by the multimodal teacher network. These networks are compared with existing knowledge distillation and emotion recognition baselines.

Performance Analysis of the Proposed CDGKD. The results in Table 2 show that the proposed CDGKD achieves state-of-the-art performance on both datasets. Traditional KD methods focus solely on the logits distribution, using the teacher network's output as additional labels to improve the student network's performance. However, these methods provide limited knowledge to the student network. In contrast, Fitnets and NST leverage intermediate layer feature maps, offering more information than traditional KD method. However, they struggle in cross-modal knowledge distillation, where structural differences between teacher and student networks make bridging the gap challenging. AMBOKD introduces an adaptive modality balancing module to address modality imbalance but still doesn't tackle the structural gap between teacher and student networks. TAKD narrows the gap between teacher and student networks by using TA networks for gradual knowledge transfer. Compared to TAKD, our CDGKD integrates knowledge from all higher-level networks to guide the student, providing a more diverse set of knowledge sources while avoiding error accumulation, achieving superior classification performance.

Method	DEAP		MAHNOB-HCI	
	Arousal	Valence	Arousal	Valence
KD	62.11	64.34	58.42	61.58
Fitnets	63.33	62.33	56.69	62.38
NST	57.36	64.65	57.75	62.38
TAKD	64.42	64.50	59.48	61.18
EmotionKD	62.71	63.36	60.53	62.58
AMBOKD	63.38	65.57	61.56	61.98
Our	**65.97**	**65.74**	**63.02**	**63.97**

Table 2: Comparison of student network performance across knowledge distillation baselines on DEAP and MAHNOB-HCI datasets.

Performance Analysis of the Proposed Modality Rebalancing Strategy. The results in Table 3 show that the multimodal teacher network, enhanced with the prototype loss-based modality rebalancing strategy, achieves state-of-the-art performance on both datasets and particularly improves results in the arousal and valence tasks. Specifically, our method achieves an average performance of 69.22% on the DEAP dataset, about 3% higher than the best-performing baseline. On the MAHNOB-HCI dataset, it achieves 69.75%

on the valence task, outperforming the baselines. EmotionKD and DGC+JCA use the same feature extractor for both modalities, simplifying the architecture, but they fail to capture the unique characteristics of each modality. CNN+SVM and CNN+LSTM use dedicated extractors for each modality, which capture modality-specific characteristics effectively. However, the differences in extractor complexity and convergence rates exacerbate the modality imbalance, with the weaker modality suppressed by the stronger one, limiting overall performance. Our method explicitly balances the convergence rates of both modalities with prototype loss, achieving superior performance.

Method	DEAP		MAHNOB-HCI	
	Arousal	Valence	Arousal	Valence
CNN+SVM	64.87	64.32	62.71	60.31
CNN+LSTM	65.10	62.40	-	-
EmotionKD	62.88	66.61	60.66	64.72
DGC+JCA	64.38	57.75	60.78	63.80
Our	**68.68**	**69.75**	**65.27**	**69.75**

Table 3: Comparison of the teacher network based on Prototype-Based Modality Rebalancing Strategy with baselines on DEAP and MAHNOB-HCI datasets.

4.5 Ablation Experiments

To further evaluate the effectiveness of the proposed prototype loss-based modality rebalancing strategy and CDGKD method, we conduct ablation experiments on the DEAP dataset. The specific setups are as follows:

Ablation of the Modality Rebalancing Strategy. To verify whether the prototype loss alleviates modality imbalance and improves performance, we design two experimental settings:

- Variant 1: No prototype loss term ($\alpha = 0$), trained using only cross-entropy loss.
- Variant 2: With prototype loss term ($\alpha = 0.5$), combining cross-entropy loss with prototype loss.

As shown in Table 4, adding the prototype loss term significantly improves performance, with an increase of approximately 8% in the arousal task and 5% in the valence task. This improvement confirms its effectiveness in addressing modality imbalance by balancing the convergence rates of both modalities, promoting the learning of the weaker modality and enhancing overall network performance.

Variant	Arousal	Valence
Variant 1	60.93	64.50
Variant 2	68.68	69.75

Table 4: Performance comparison of teacher network with and without the prototype-based modality rebalancing strategy.

Variant	Arousal	Valence
Variant 3	61.47	63.02
Variant 4	64.42	64.50
Variant 5	65.97	65.74

Table 5: Performance comparison of student network under different experimental setups of the proposed CDGKD and stochastic learning strategy.

Ablation of Our CDGKD and Stochastic Learning Strategy. To evaluate the synergy between our CDGKD method and the stochastic learning strategy, we design the following experimental setups:

- Variant 3: No knowledge distillation, the student network is trained independently.
- Variant 4: Using the CDGKD without the stochastic learning strategy.
- Variant 5: Using the CDGKD with the stochastic learning strategy (full scheme).

As shown in Table 5, the student network's performance is lowest without knowledge distillation. Introducing the proposed CDGKD significantly improves performance, highlighting the importance of dense guidance and multi-level TA networks in bridging the gap between the teacher and student networks and preventing error accumulation. The full scheme combines our CDGKD with the stochastic learning strategy and achieves the best performance, improving 4.5% on the arousal task and 2.72% on the valence task compared to Variant 3. The stochastic learning strategy introduces diversity in knowledge transfer, mitigating overfitting and enhancing the student network's generalization and robustness. In conclusion, the combination of CDGKD and the stochastic learning strategy significantly improves the unimodal student network's performance.

5 Conclusion

We propose an innovative cross-modal knowledge distillation framework that leverages a multimodal teacher network to fuse visual and EEG features and efficiently transfer them to a unimodal student network. To address the issue of modality imbalance, a prototype-based modality rebalancing strategy is designed to enhance multimodal feature fusion, obtaining high-quality multimodal knowledge representations. Additionally, the proposed CDGKD effectively bridges the structural differences between the teacher and student networks, enabling efficient transfer of the multimodal knowledge extracted by the teacher network and improving the performance of the unimodal student network. Experimental results demonstrate that the proposed framework improves the emotion recognition performance of the unimodal visual network on the DEAP and MAHNOB-HCI datasets. This study is the first to apply cross-modal knowledge distillation from a multimodal teacher network to guide a unimodal visual network for emotion recognition task, providing new insights into emotion recognition under modality-constrained conditions.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62306317).

References

[Aslam *et al.*, 2024] Muhammad Haseeb Aslam, Marco Pedersoli, Alessandro Lameiras Koerich, and Eric Granger. Multi teacher privileged knowledge distillation for multimodal expression recognition. *arXiv preprint arXiv:2408.09035*, 2024.

[Canal *et al.*, 2022] Felipe Zago Canal, Tobias Rossi Müller, Jhennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sobieranski. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582:593–617, 2022.

[Cheng *et al.*, 2024a] Cheng Cheng, Wenzhe Liu, Lin Feng, and Ziyu Jia. Dense graph convolutional with joint cross-attention network for multimodal emotion recognition. *IEEE Transactions on Computational Social Systems*, 2024.

[Cheng *et al.*, 2024b] Cheng Cheng, Wenzhe Liu, Lin Feng, and Ziyu Jia. Emotion recognition using hierarchical spatial–temporal learning transformer from regional to global brain. *Neural Networks*, 179:106624, 2024.

[Cheng *et al.*, 2025] Cheng Cheng, Wenzhe Liu, Xinying Wang, Lin Feng, and Ziyu Jia. Disd-net: A dynamic interactive network with self-distillation for cross-subject multi-modal emotion recognition. *IEEE Transactions on Multimedia*, 2025.

[Cho and Hariharan, 2019] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019.

[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*, 14(3):2238–2250, 2022.

[Du *et al.*, 2021] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multi-modal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*, 2021.

[Fan *et al.*, 2023] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 20029–20038. IEEE, 2023.

[Gupta *et al.*, 2016] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

[Hinton, 2015] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Huang and Wang, 2017] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[Huang *et al.*, 2019] Yongrui Huang, Jianhao Yang, Siyu Liu, and Jiahui Pan. Combining facial expressions and electroencephalography to enhance emotion recognition. *Future Internet*, 11(5):105, 2019.

[Jia *et al.*, 2024] Ziyu Jia, Fengming Zhao, Yuzhe Guo, Hairong Chen, Tianzi Jiang, and Brainnetome Center. Multi-level disentangling network for cross-subject emotion recognition based on multimodal physiological signals. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 3069–3077, 2024.

[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.

[Li *et al.*, 2023] Jia Li, Yin Chen, Xuesong Zhang, Jiantao Nie, Ziqiang Li, Yangchen Yu, Yan Zhang, Richang Hong, and Meng Wang. Multimodal feature extraction and fusion for emotional reaction intensity estimation and expression classification in videos with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 5838–5844. IEEE, 2023.

[Li *et al.*, 2024] Zixing Li, Chao Yan, Zhen Lan, Xiaojia Xiang, Han Zhou, Jun Lai, and Dengqing Tang. Adaptive modality balanced online knowledge distillation for brain-eye-computer based dim object detection. *arXiv preprint arXiv:2407.01894*, 2024.

[Lian *et al.*, 2023] Hailun Lian, Cheng Lu, Sunan Li, Yan Zhao, Chuangao Tang, and Yuan Zong. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, 25(10):1440, 2023.

[Liang *et al.*, 2023] Heng Liang, Yucheng Liu, Haichao Wang, Ziyu Jia, and Brainnetome Center. Teacher assistant-based knowledge distillation extracting multi-level features on single channel sleep eeg. In *IJCAI*, pages 3948–3956, 2023.

[Liu *et al.*, 2023a] Yuanyuan Liu, Wenbin Wang, Chuanxu Feng, Haoyu Zhang, Zhe Chen, and Yibing Zhan. Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition*, 138:109368, 2023.

[Liu *et al.*, 2023b] 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.

[Mirzadeh *et al.*, 2020] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020.

[Ngiam *et al.*, 2011] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[Peng *et al.*, 2022] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022.

[Romero *et al.*, 2014] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[Saffaryazdi *et al.*, 2022] Nastaran Saffaryazdi, Syed Talal Wasim, Kuldeep Dileep, Alireza Farrokhi Nia, Suranga Nanayakkara, Elizabeth Broadbent, and Mark Billinghurst. Using facial micro-expressions in combination with eeg and physiological signals for emotion recognition. *Frontiers in Psychology*, 13:864047, 2022.

[Shen *et al.*, 2017] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, Wenwu Zhu, et al. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844, 2017.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[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.

[Son *et al.*, 2021] Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. Densely guided knowledge distillation using multiple teacher assistants. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9375–9384. IEEE, 2021.

[Tan *et al.*, 2021] Ying Tan, Zhe Sun, Feng Duan, Jordi Solé-Casals, and Cesar F Caiafa. A multimodal emotion recognition method based on facial expressions and electroencephalography. *Biomedical Signal Processing and Control*, 70:103029, 2021.

[Udahemuka *et al.*, 2024] Gustave Udahemuka, Karim Djouani, and Anish M Kurien. Multimodal emotion recognition using visual, vocal and physiological signals: a review. *Applied Sciences*, 14(17):8071, 2024.

[Wang *et al.*, 2020] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020.

[Wang *et al.*, 2021] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021.

[Wang *et al.*, 2025] Jing Wang, Zhiyang Feng, Xiaojun Ning, Youfang Lin, Badong Chen, and Ziyu Jia. Two-stream dynamic heterogeneous graph recurrent neural network for multi-label multi-modal emotion recognition. *IEEE Transactions on Affective Computing*, 2025.

[Wu *et al.*, 2023] Yi-Chiao Wu, Li-Wen Chiu, Chun-Chih Lai, Bing-Fei Wu, and Sunny SJ Lin. Recognizing, fast and slow: Complex emotion recognition with facial expression detection and remote physiological measurement. *IEEE Transactions on Affective Computing*, 14(4):3177–3190, 2023.

[Xiao *et al.*, 2020] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[Zadeh *et al.*, 2017] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[Zhang *et al.*, 2022] Yue Zhang, Wanying Ding, Ran Xu, Xiaohua Hu, and Lud De Raedt. Visual emotion representation learning via emotion-aware pre-training. In *IJCAI*, pages 1679–1685, 2022.

[Zhang *et al.*, 2024] Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27456–27466, 2024.

[Zhao *et al.*, 2018] Sicheng Zhao, Guiguang Ding, Jungong Han, and Yue Gao. Personality-aware personalized emotion recognition from physiological signals. In *IJCAI*, pages 1660–1667, 2018.