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

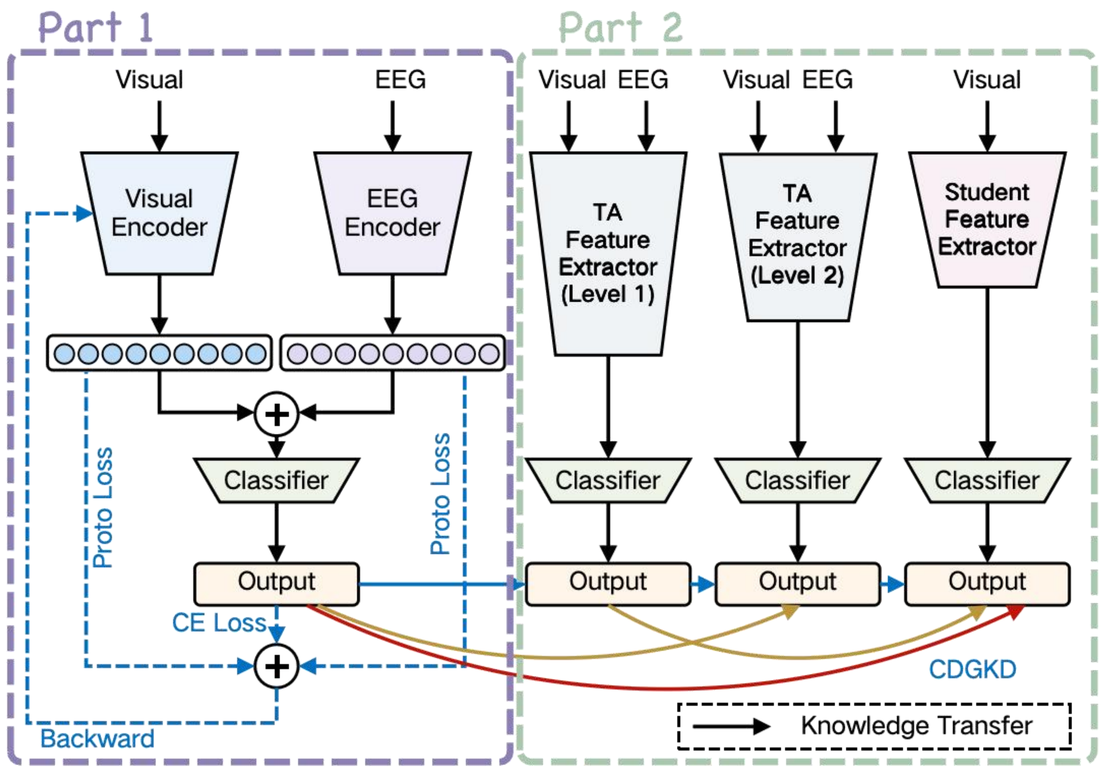
Shuang Wu[1], Heng Liang[2], Yong Zhang[3], Yanlin Chen[4], Ziyu Jia[5]

[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

Introduction

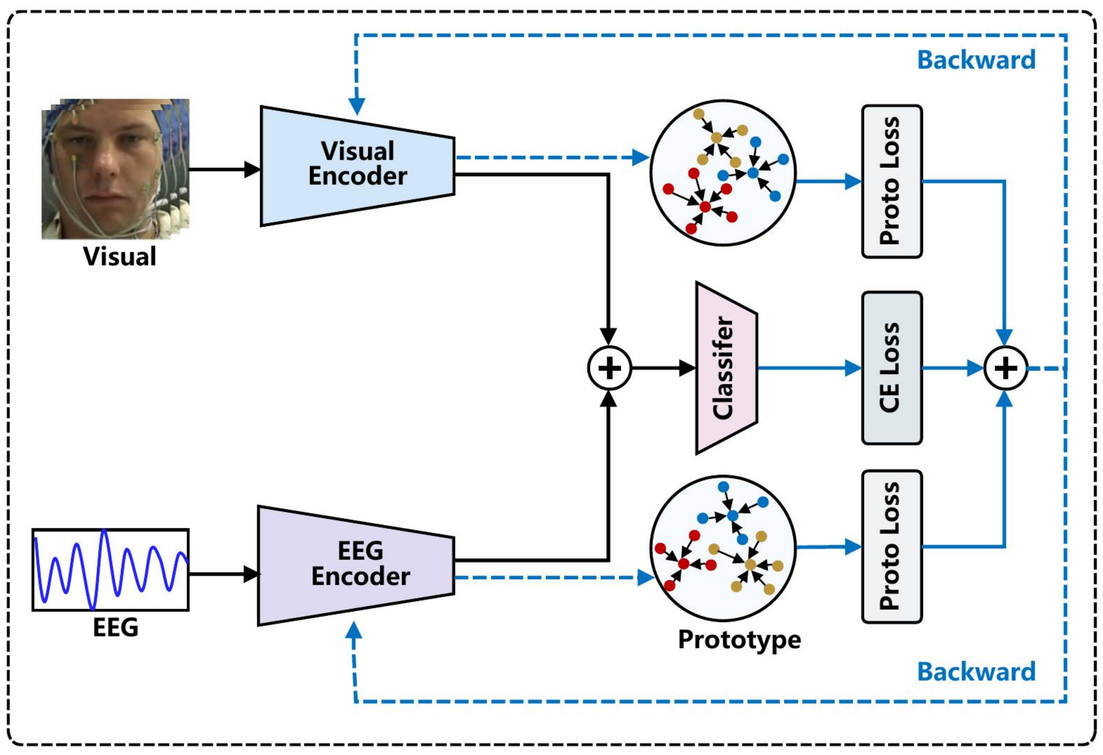
- **Goal**: Enhance the performance of a unimodal (visual-only) emotion recognition network by transferring knowledge from a powerful multimodal (Visual + EEG) teacher network.
- **Problem**: In many real-world applications, physiological signals like EEG are difficult and expensive to acquire compared to visual data, limiting the use of multimodal systems.
- **Challenge 1**: Modality Imbalance. Multimodal learning is often inefficient because feature differences cause stronger modalities (e.g., visual) to suppress weaker ones (e.g., EEG), limiting the model's overall performance.
- **Challenge 2**: Teacher-Student Gap. Direct knowledge transfer from a complex multimodal teacher to a simple unimodal student is often ineffective due to the large structural gap. Step-by-step transfer using assistants can lead to error accumulation.

Overall of our approach



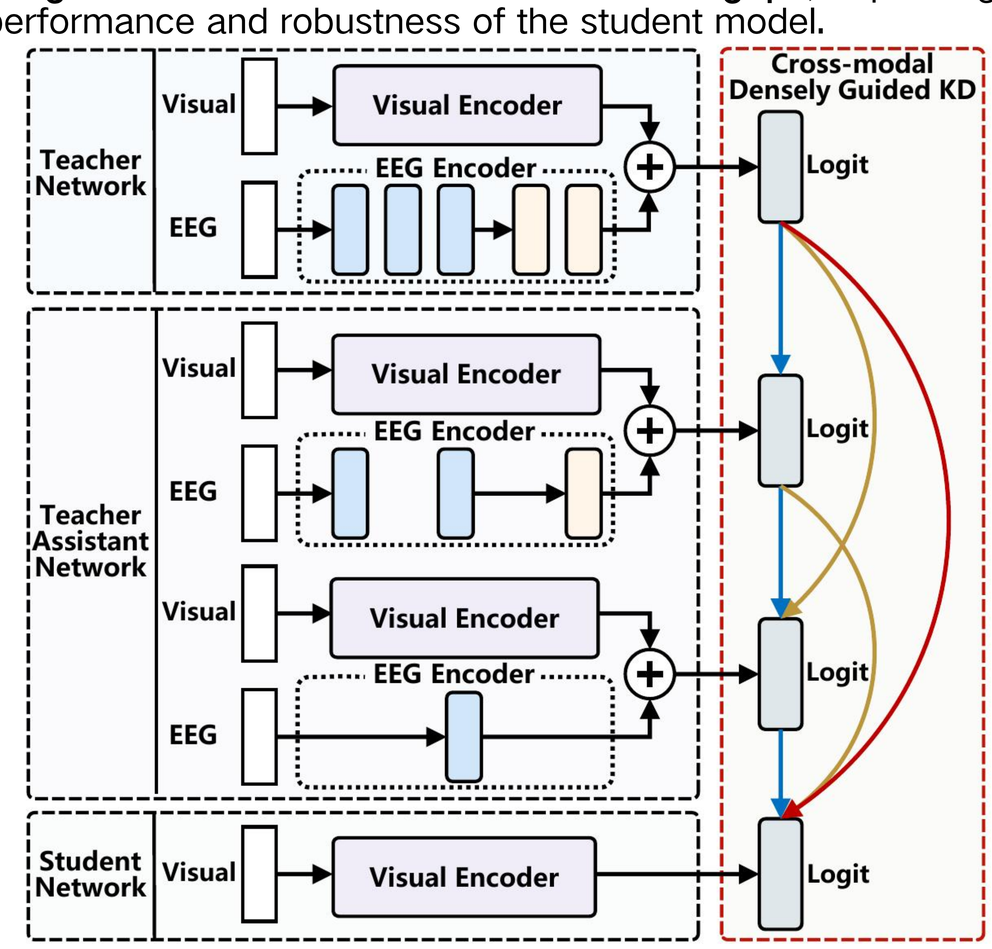
Part 1: Prototype-based modality rebalancing

- **Compute emotion-class prototypes** for each modality.
- Use **prototype loss** to encourage faster convergence.
- **Adaptively increase loss weight** for slower modalities during training, balancing multimodal learning without suppressing stronger ones.



Part 2: Cross-modal densely guided knowledge distillation (CDGKD)

- **Construct a multi-level distillation framework** using a series of TA networks.
- **Provide dense, stochastic supervision** from the teacher and all TAs to guide the unimodal student.
- **Mitigate error accumulation and structural gaps**, improving the performance and robustness of the student model.



Results

Datasets: DEAP and **MAHNOB-HCI**, both containing synchronized EEG and facial video recordings along with self-reported valence and arousal labels.

We compare our CDGKD-based student network against existing cross-modal knowledge distillation methods.

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**

We further demonstrate the effectiveness of our modality rebalancing strategy in enhancing multimodal feature fusion.

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**

Conclusion

- We propose a **cross-modal distillation framework** that transfers multimodal (EEG + visual) knowledge to a unimodal visual network.
- A **modality rebalancing strategy** and multi-level distillation help bridge structural gaps and improve transfer efficiency.
- Our method boosts emotion recognition on DEAP and MAHNOB-HCI, offering a novel solution for modality-constrained settings.