# C$^2$KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation ——Appendix——

Fushuo Huo[1], Wenchao Xu[1*], Jingcai Guo[1] , Haozhao Wang[2], Song Guo[3]
[1]The Hong Kong Polytechnic University, [2]Huazhong University of Science and Technology,
[3]The Hong Kong University of Science and Technology

**Overview**

The appendix presents more dataset details, experimental results, settings, and analyses as follows:

**Appendix A** More details about multimodal datasets.

**Appendix B** More results on different architectures.

**Appendix C** More implementation details.

**Appendix D** More ablation studies.

**Appendix E** Training computation overhead analysis

**Appendix A: More details about multimodal datasets**

We follow [1, 6, 11] and conduct experiments on four multimodal datasets: (1) **CREMA-D** [3] is an audio-visual dataset for speech emotion recognition, which contains facial and vocal emotional expressions. The emotions has 6 categorizations: *angry*, *happy*, *sad*, *neutral*, *discarding*, and *fear*. The whole dataset consists of 7,442 video clips, divided into 6698 samples as a training set and 744 samples as a testing set. (2) **AVE** [13] is an audio-visual dataset for audio-visual event localization, in which there are 28 event classes and consists of 4,143 10-second video clips. Following [6, 11], we construct a labeled multimodal dataset by extracting the frames from event-localized video segments and capturing the audio clips within the same segments. The training, validation, and testing splits of the dataset follow [13]. (3) **VGGsound** [4] is a large-scale video dataset containing 309 classes covering a wide range of audio events in everyday life. All videos are captured from YouTube with audio-visual correspondence, and the source of the sound is visually evident. The duration of each video is 10 seconds, and the dataset partition is the same as in [4]. We randomly choose *50* class to conduct experiments. (4) **CrisisMMD** [2] is a multimodal crisis prediction dataset. It consists of 8079 aligned images and associated texts, where images and texts are collected from Twitter. The corpus is divided into eight humanitarian categories, including infrastructure and utility damage, vehicle damage, rescue, volunteering, or donation effort, injured or dead people, affected individuals, missing or found people, other relevant information, and not relevant or can't judge, with defaulted dataset splits.

*Corresponding author: wenchao.xu@polyu.edu.hk

**Appendix B: More results on different architectures**

Besides the results in Tables 3 & 4 in the main body, we conduct more experiments to further demonstrate the effectiveness of our method across diverse-capacities homogeneous and heterogeneous architectures. We compare C$^2$KD with vanilla KD [8], the state-of-the-art feature-based KD (Review [5]), online KD (SHAKE [9]), logits-based KD (NKD [14]). The results in Table 1 illustrate C$^2$KD can effectively transfer crossmodal knowledge across diverse-capacities homogeneous architectures (i.e., ResNet-18-ResNet-50) and heterogeneous architectures (i.e., BERT-ResNet-18 and BERT-ShuffleNet V2).

**Appendix C: More implementation details**

**C.1: Detailed preprocess strategy**

We follow [6, 11] and give the detailed preprocess strategy. For audio modality, we change the input channel from 3 to 1 as [4]. Audio data is transformed into a spectrogram of size 257×299 for CREMA-D, 257×1,004 for AVE, and 257×1,004 for VGGsound, respectively, with the window length of 512 and overlap of 353. For visual modality, the input channel is adjusted considering input frames [15]. Concretely, 3 frames are uniformly sampled from VGGsound, and 1 frame is extracted from AVE and CREMA-D. Standard augmentations are employed, including random cropping and flipping.

**C.2: More implementations of our method**

We initialize weights of the student model and proxies following [7]. All experiments are conducted with NVIDIA RTX3090 GPUs on CUDA 11.4 using the PyTorch framework. All results are the average of three different seeds, which are set to 1, 2, and 3, respectively.

**C.3: Detailed implementations of compared methods**

We imply traditional unimodal knowledge distillation with their defaulted settings. Previous logits-based KD methods can be seamlessly applied to the Cross-Modal Knowledge Distillation (CMKD) task. Due to the different spatial dimensions of multimodal inputs, the intermediate features have different spatial dimensions. Feature-based KD methods can't be directly applied to CMKD. To deal with this issue, we employ the bilinear interpolation opera-

| | CREMA-D [3] | | AVE [13] | | VGGsound [4] | | CrisisMMD [2] | |
|---|---|---|---|---|---|---|---|---|
| | Visual (A→ V) | Audio (V→ A) | Visual (A→V) | Audio (V→A) | Visual (A→V) | Audio (V→A) | Image (T→I) | Text (I→T) |
| | RN18 | RN50 | RN18 | RN50 | RN18 | RN50 | BERT | SNV2 |
| w/o KD | 58.1±0.33 | 57.9±0.19 | 31.6±0.18 | 53.7±0.16 | 38.7±0.16 | 60.1±0.18 | 66.7±0.22 | 68.0±0.12 |
| KD [8] | 57.1±0.57 | 54.1±0.43 | 32.6±0.62 | 48.5±0.35 | 39.0±0.46 | 57.8±0.51 | 66.2±0.38 | 68.4±0.22 |
| Review [5] | 59.4±0.52 | 56.9±0.62 | 32.0±0.53 | 51.3±0.57 | 38.5±0.53 | 58.7±0.60 | - | - |
| SHAKE [9] | 60.2±0.36 | 58.9±0.63 | 32.5±0.67 | 48.6±0.46 | 38.9±0.51 | 59.9±0.38 | 68.2±0.23 | 69.6±0.25 |
| NKD [14] | 60.5±0.62 | 56.9±0.43 | 33.0±0.36 | 52.5±0.36 | 39.2±0.67 | 59.6±0.54 | 67.3±0.31 | 68.6±0.25 |
| Ours | 63.1±0.25 | 62.1±0.37 | 35.0±0.21 | 55.3±0.12 | 41.0±0.22 | 62.0±0.23 | 68.9±0.12 | 70.0±0.09 |
| | RN50 | RN18 | RN50 | RN18 | RN50 | RN18 | BERT | RN18 |
| w/o KD | 59.7±0.20 | 56.3±0.22 | 32.7±0.25 | 52.8±0.11 | 39.3±0.13 | 59.4±0.16 | 66.7±0.22 | 68.1±0.13 |
| KD [8] | 58.2±0.53 | 54.0±0.36 | 33.0±0.43 | 46.9±0.42 | 38.9±0.52 | 56.4±0.61 | 66.2±0.42 | 68.5±0.21 |
| Review [5] | 60.4±0.58 | 55.9±0.39 | 32.7±.0.56 | 51.2±0.61 | 38.2±0.43 | 58.1±0.61 | - | - |
| SHAKE [9] | 60.5±0.53 | 59.0±0.48 | 33.4±0.53 | 47.5±0.43 | 38.6±0.41 | 59.8±0.49 | 68.0±0.19 | 69.8±0.23 |
| NKD [14] | 60.9±0.54 | 58.4±0.62 | 33.2±0.47 | 52.8±0.55 | 39.5±0.53 | 59.1±0.46 | 67.4±0.26 | 68.6±0.22 |
| Ours | 63.5±0.28 | 61.6±0.23 | 35.5±0.30 | 55.1±0.22 | 41.3±0.28 | 62.1±0.24 | 68.8±0.16 | 70.2±0.16 |

Table 1. **Comparison results on Visual-Audio and Image-Text datasets.** The metric is the top-1 accuracy (%). RN18: ResNet-18; RN50: ResNet-50; SNV2: ShuffleNet V2 [10].

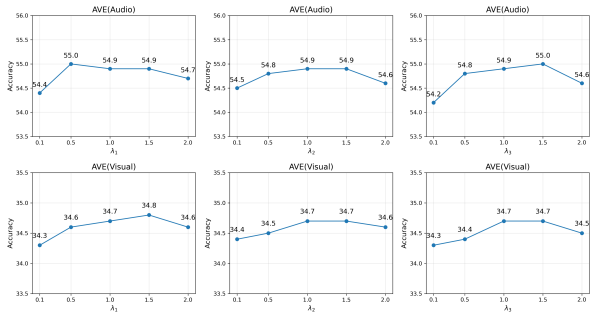


Figure 1. **Analysis of $\lambda_1$, $\lambda_2$, $\lambda_3$.** We conduct experiments on the AVE [13] dataset with ResNet-18 as the multimodal backbones.

tor to align the intermediate features of teacher and student. Besides, BERT has 12 layers while MobileNetV2 has 5 layers. We don't conduct feature-based KD on the CrisisMMD dataset because we can't choose which layers to be distilled based on their original implementations.

## Appendix D: More ablation studies

### D.1: Hyperparameter Analysis

We analyse the $\lambda_1$, $\lambda_2$, and $\lambda_3$ in Equation 6. $\lambda_1$ and $\lambda_2$ represent the weight of bidirectional distillation between the proxy and student/teacher and $\lambda_3$ denotes the weight of teacher proxy and student proxy. We vary one hyperparameter and leave the other unchanged. As shown in Figure 1, Our method is robust in terms of different hyperparameters. As our method can effectively transfer corssmodal information, the large and small values of $\lambda$ might hinder the knowledge transfer. Therefore, we adopt $\{\lambda_1 = \lambda_2 = \lambda_3 = 1\}$ in all experiments.

### D.2: Proxy Analysis

| Method | AVE [13] | | VGGsound [4] | |
|---|---|---|---|---|
| | Visual (A→V) | Audio (V→A) | Visual (A→V) | Audio (V→A) |
| w/o FA | 34.4±0.36 | 53.0±0.22 | 40.2±0.25 | 60.6±0.24 |
| w/ CFA | 34.7±0.18 | 55.0±0.20 | 40.8±0.28 | 62.0±0.21 |
| **Ours** | 34.7±0.23 | 54.9±0.16 | 40.9±0.31 | 61.9±0.27 |

Table 2. **Analysis of the structure of the proxy.** We conduct experiments on the AVE [13] and VGGsound [4] datasets with ResNet-18 as the multimodal backbones.

We provide the detailed analysis of the student and teacher proxies. The proxy consists of the feature adaptation layer and the linear classification head, as shown in Equation 5. The feature adaptation layer follows the feature-based KD methods [5, 12], consisting of 'Conv-BN-ReLU' block. Specifically, the kernel size of 'Conv' is set to $1\times1$, and input and output channel dimensions remain the same. Here, we analyse the structure of the proxy. We ablate the feature adaptation layer (w/o FA) and employ a complicated feature adaptation layer ('Conv-BN-Conv-BN-ReLU', i.e., w/ CFA). Table 2 illustrates that without the feature adaptation layer (w/o FA), the linear classification head can't effectively transfer crossmodal information, possibly due to the degradation of nonlinear ability. However, the complicated feature adaptation layer does not bring obvious improvement. Therefore, the feature adaptation layer and linear classification head constitute the proxy.

## Appendix E: Training computation overhead analysis

The extra training computation costs of C$^2$KD are twofold: student and teacher proxy and (partially) updating teacher. We measure the training time and GPU memory

| V→A | KD[8] | Review[5] | SHAKE[9] | NKD[14] | Ours† | Ours‡ |
|---|---|---|---|---|---|---|
| Mem. | 7492 | 8679 | 7736 | 7612 | 7635 | 11353 |
| Time | 55 | 67 | 60 | 53 | 54 | 66 |
| A→V | KD[8] | Review[5] | SHAKE[9] | NKD[14] | Ours† | Ours‡ |
| Mem. | 9670 | 10764 | 9854 | 9725 | 9787 | 11005 |
| Time | 56 | 68 | 62 | 55 | 53 | 68 |

Table 3. **Analysis of the training computation overhead.** We conduct experiments on the VGGsound [4] datasets with ResNet-18 as the multimodal backbones. Time: training times for one epoch in second; Mem.: peak GPU memory usages (MB); Ours‡: fully updating the teacher model; Ours†: partially finetuning the top 2 layers.

usages in Table 3. The proxies are lightweight and induce few costs. Although updating the whole teacher induces extra training computation costs, we partially tune the teacher model and also achieve superior performance and competitive training efficiency. Note that in the inference phase, the inference costs are the same for all distilled student models.

## References


[1] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. Multimodal categorization of crisis events in social media. In *CVPR*, 2020. 1

[2] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. *AAAI*, 2018. 1, 2

[3] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 2014. 1, 2

[4] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 1, 2, 3

[5] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, 2021. 1, 2, 3

[6] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *CVPR*, 2023. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. 1

[8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015. 1, 2, 3

[9] Lujun Li and Zhe Jin. Shadow knowledge distillation: Bridging offline and online knowledge transfer. In *NeurIPS*, 2022. 1, 2, 3

[10] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, 2018. 2

[11] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, 2022. 1

[12] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015. 2

[13] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. 1, 2

[14] Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, and Yu Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *ICCV*, 2023. 1, 2, 3

[15] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018. 1