2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.01515

C^{2}KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation

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

Abstract

Existing Knowledge Distillation (KD) methods typically focus on transferring knowledge from a large-capacity teacher to a low-capacity student model, achieving substantial success in unimodal knowledge transfer. However, existing methods can hardly be extended to Cross-Modal Knowledge Distillation (CMKD), where the knowledge is transferred from a teacher modality to a different student modality, with inference only on the distilled student modality. We empirically reveal that the modality gap, i.e., modality imbalance and soft label misalignment, incurs the ineffectiveness of traditional KD in CMKD. As a solution, we propose a novel ***C****ustomized* ***C****rossmodal* ***K****nowledge* ***D****istillation (C^{2}KD). Specifically, to alleviate the modality gap, the pre-trained teacher performs bidirectional distillation with the student to provide customized knowledge. The On-the-Fly Selection Distillation(OFSD) strategy is applied to selectively filter out the samples with misaligned soft labels, where we distill cross-modal knowledge from non-target classes to avoid the modality imbalance issue. To further provide receptive cross-modal knowledge, proxy student and teacher, inheriting unimodal and cross-modal knowledge, is formulated to progressively transfer cross-modal knowledge through bidirectional distillation. Experimental results on audio-visual, image-text, and RGB-depth datasets demonstrate that our method can effectively transfer knowledge across modalities, achieving superior performance against traditional KD by a large margin.*

1. Introduction

Knowledge Distillation (KD) is an effective approach to transfer knowledge from the large-capacity teacher model to the low-capacity student model during training [12, 40]. During the KD process, the student is trained to mimic the teacher's output via the distillation loss. KD methods can be divided into two main categories: logits-based and feature-based methods. The former minimizes the discrepancy between soft labels of the teacher model and the student model [16, 17, 53], and the latter distills knowledge from intermediate feature layers [7, 15, 18].

	AVE[37]		**VGGSound**[5]	
Method	**Visual** **(A→V)**	**Audio** **(V→A)**	**Visual** **(A→V)**	**Audio** **(V→A)**
w/o KD	31.6±0.18	52.8±0.11	38.7±0.16	59.4±0.16
KD [16]	32.3±0.35	46.6±0.24	38.5±0.50	56.3±0.46
Review [7]	32.1±0.63	50.6±0.31	38.2±0.47	57.9±0.33
DML [53]	31.8±0.41	48.0±1.31	38.7±0.86	58.2±1.01
SHAKE [25]	32.2±0.59	47.3±0.72	38.3±0.41	59.5±0.34
DKD [54]	32.6±0.65	48.6±1.02	38.1±0.43	57.2±0.86
DIST [17]	29.8±0.61	49.3±0.52	38.5±0.39	58.9±0.45
NKD [47]	32.9±0.32	52.2±0.62	39.2±0.52	59.3±0.40
Ours	34.7±0.23	54.9±0.16	40.9±0.31	61.9±0.27

Table 1. **Performances of traditional KD in CMKD.** The results of distilled student modality infer only on the student modality. **A→V**: Audio teacher modality distills visual student modality; **V→A**: Visual teacher modality distills audio student modality.

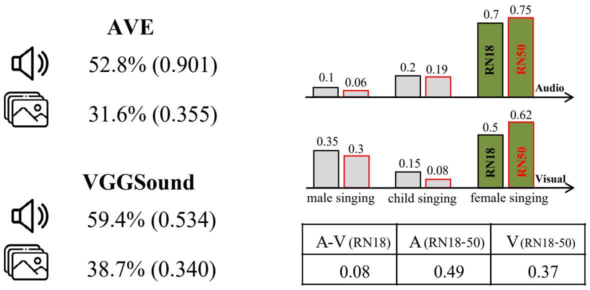
Despite the success of traditional KD methods in single modality scenario, extending these methods to address the Cross-Modal Knowledge Distillation (CMKD) tasks remains a critical challenge. The CMKD task involves knowledge transfer from one modality to another during the distillation phase, with inference *only* on the *distilled student modality*, which is crucial especially in computation-constrained and sensor-failure scenarios. As demonstrated in Table 1, unimodal KD methods struggle to transfer knowledge from the low-accuracy visual modality to the high-accuracy audio modality, while the visual modality has only marginal gains from the audio modality. Based on the above analysis, a pivotal and fundamental question arises: *Can we effectively transfer arbitrary unimodal information to another modality?*

To answer this question, we conduct empirical analysis to investigate why traditional KD methods fail in CMKD from the logits-based perspective, which can be attributed to the inter-modality gap that inducing *modality imbalance* and *soft label misalignment*, as illustrated in Figure 1.

For the first factor, we define *modality imbalance*, akin to [11, 30], as the performance disparities between modalities. We quantitatively calculate the top-1 accuracy (followed

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

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.01515



A-V (RN18)	A (RN18-50)	V (RN18-50)
0.08	0.49	0.37

(a) modality imbalance **(b) soft label misalignment**

Figure 1. **The Modality Gap of CMKD. (a)** Top-1 accuracy (followed by average prediction probability of target classes) of each modality. Both modalities utilize ResNet-18 as the backbone. **(b)** Up: Example of three-class classification. Down: Kendall Rank Correlation [20] of soft labels across modalities in VGGSound. A: audio; V: visual; RN: ResNet.

by the average prediction probability of target classes) after training on the corresponding single modality. Figure 1(a) shows that the audio modality outperforms the visual modality in AVE and VGGsound datasets, and there are significant gaps in the average prediction probability of the target class, particularly in AVE. Merely distilling knowledge from the visual to the audio modality could potentially yield adverse effects, as shown in Table 1 (Audio columns).

For the second factor, we define *soft label* as the output distributions from the teacher network, following [16, 26]. The soft labels contain meaningful information on similarity among various classes. However, inter-modality gap leads to severe soft label misalignment between teacher and student modalities. Take three-class classification as an example (Figure 1(b) Up). Although both Audio and Visual modalities branches successfully predict the target class of 'female singing', the non-target soft labels are rank-distorted, where the *audio accent* of 'child singing' is more closely related to 'female singing', while the *visual appearance* of 'male singing' is more closely resembles 'female singing'. Direct transferring soft label information across modalities is unreasonable, which could explain why distilling the audio modality to the visual modality does not yield significant improvements. To quantitatively validate soft label misalignment, we further calculate the average Kendall Rank Correlation (KRC) [20] of soft labels in Figure 1(b) Down. A higher KRC indicates better rank correlation. The table indicates the KRC of multimodal soft label (i.e., A-V(RN-18)) is significantly lower than that of a single modality with diverse-capacity networks (i.e., A(RN18-50) and V(RN18-50)), indicating the presence of misalignment of multimodal soft labels.

To address the above issues in CMKD, we propose Customized cross-modal Knowledge Distillation (C^{2}KD). Concretely, instead of using the pre-trained teacher to provide supervision signals to the student, we bidirectionally update to customize both the pre-trained teacher and student via On-the-Fly Selection Distillation (OFSD) strategy, where OFSD selectively distill receptive soft labels according to the Kendall Rank Correlation, and cross-modal knowledge is transferred from non-target classes to avoid the modality imbalance issue. Furthermore, Proxy student and teacher, inheriting unimodal and cross-modal knowledge, is formulated to progressively transfer cross-modal knowledge in the bidirectional distillation form.

Our main contributions can be summarized as follows.

- We empirically analyze the factors for the failure of unimodal KD in CMKD, which can be attributed to the modality imbalance and soft label misalignment.
- To address these issues, we propose a novel method named C^{2}KD. Specifically, OFSD produces selected crossmodel non-target class knowledge through on-the-fly bidirectionally distilling both student and teacher. Moreover, Proxy student and teacher are built to progressively transfer receptive knowledge across modalities. The proposed strategies are plug-and-play, enhancing traditional KD methods in CMKD.
- We conduct experiments on sparse and dense prediction tasks, including audio-visual, image-text, and RGB-Depth datasets. Diverse capacities and homogeneous/heterogeneous architectures are also considered. Extensive experiments validate C^{2}KD can transfer cross-modal knowledge from *arbitrary* modality to another.

2. Related Work

2.1. Unimodal Knowledge Distillation

Unimodal Knowledge Distillation (KD) transfers the knowledge of a pretrained teacher to a student by minimizing the discrepancies between output logits or intermediate features between student and teacher. Previous KD methods primarily concentrate on inheriting knowledge from the large-capacity teacher. Pioneering work [16] regularizes Kullback–Leibler (KL) divergence between student and teacher soft labels. CRD [38] develops contrastive-based objectives for knowledge transferring. SCKD [55] automatically adjusts the KD process according to the distillation gradient similarity. Yang et al. [45] utilize the teacher's pre-trained classifier to regularize the student's penultimate layer feature. Zhu et al. [56] identify and discard the undistillable classes from the large teacher model based on the validation set. DKD [54] decouple KD into target class and non-target class knowledge distillation to balance learning effectiveness and flexibility. Review [7] proposes the review mechanism to utilize knowledge of teacher's multi-level features. RKD [28] regularizes the student with distance-wise and angle-wise structural relations to replace KL loss. DIST [17] further proposes a novel correlation-based loss to capture the inter-class and intra-class relations. L2D [46] extends relation-based distillation into multi-label classification. These KD methods focus on unimodal KD and learn to inherit knowledge from a fixed teacher. However, for CMKD, the modality gap (shown in Figure 1) impedes knowledge transfers across modalities. We *argue* that teacher modality should be optimized with feedback supervision of student modality to produce *receptive knowledge*. Previous online knowledge distillation methods [8, 22, 25, 53] update teacher model to adapt student in unimodal scenarios. Specifically, DML [53] simply applies KD losses mutually, treating each other as teachers. ONE [22] further exploits gated ensemble logits of multiple training networks. AFD [8] proposes online feature alignments via adversarial training. The recently proposed SHAKE [25] bridges offline and online KD by transferring knowledge through extra shadow heads. However, these methods primarily consider differences in network capacity and suffer from the modality gap in cross-modal KD, as the results of DML and SHAKE are shown in Table 1.

2.2. Cross-modal Knowledge Distillation

With the rising prevalence of multi-modal sensors, traditional KD methods have been extended to achieve knowledge transfer across multimodal data, thereby enhancing downstream tasks [13, 23, 32, 40, 41, 50, 52]. *However*, previous methods typically utilize *high-accuracy* or *well-labeled* modality as the teacher to transfer knowledge to low-accuracy or unlabeled modality [40]. For example, [13] leverage a large labeled modality as the supervisory signal for a new unlabeled paired modality. [32] transfers knowledge among the missing and available modalities via GANs. [41] adapts a multimodal network to the unlabeled modality by inheriting knowledge from the well-trained unimodal teacher. [23] proposes a decomposed cross-modal distillation method to enhance RGB-based detector by transferring knowledge of the optical flow modality. [50] distills ImageNet pre-trained visual modality to audio modality for indoor dense prediction. Recently, Xue et al. [42] first perform an in-depth investigation on CMKD and propose the modality focusing hypothesis (MFH), suggesting that modality-general decisive features are crucial determinants of CMKD efficacy. [42] contributes to MFH but doesn't develop unified solutions. In this paper, we further quantitatively analyze the challenges of CMKD (the modality gap, i.e., *modality imbalance* and *soft label misalignment*) and propose effective solutions to address these issues.

2.3. Multimodal Learning

Multimodal learning [3] is a crucial area of research, given ubiquitous multimodal information. Multimodal learning [10, 11, 19, 30, 51] involves training and inference on multiple aligned modalities. To integrate multimodal information, [30] modulates the gradient of each modality to alleviate under-optimized uni-modal representation. [11] introduces class prototypes that direct modality optimization. [10] analyzes Modality Laziness in multimodal learning and proposes unimodal ensemble and unimodal distillation strategies based on the distribution of uni-modal and paired features. [51] devises a novel quality-aware multimodal fusion method, leveraging energy-based uncertainty to characterize each modality quality. Different from multimodal learning, CMKD trains on multimodal data but *infers only* on a single modality, which has significant practical implications for downstream applications.

3. Method

3.1. Cross-Modal KD Effectiveness Analysis

First, we revisit traditional KD in cross-modal scenario. Given multimodal training data ([X_1, X_2], Y) containing multimodal samples X_1 and X_2 and labels Y. Let f_T and f_S be the output logits of the teacher T and student S. The corresponding prediction probabilities are obtained using the softmax function (σ): $p_S = \sigma(f_S)$ and $p_T = \sigma(f_T)$. Typical KD trains the student network as follows:

$$L_{KD} = \mathcal{H}(p_S, Y) + \lambda \mathcal{D}(p_S, p_T) \tag{1}$$

where $\mathcal{H}$ is the supervision loss function (typical Cross-Entropy (CE) loss), $\mathcal{D}$ is the KD loss to minimize the discrepancy of output distribution between teachers and students, commonly achieved using Kullback–Leibler (KL) divergence [16], and λ is a balancing parameter for these two terms. Pioneering work [42] proposes the Modality Focusing Hypothesis (MFH) and claims that modality-general decisive features are crucial for transferring knowledge across modalities during the distillation phase. In this work, we provide another fine-grained perspective to investigate the efficacy of CMKD: the modality gap, which refers to the *modality imbalance* in target-class logits and *soft label misalignment*, incurs the failure of CMKD.

Regarding *modality imbalance*, as depicted in Figure 1(a), the prediction possibility of the target class exhibits significant variations across modalities. If simply let student modality (audio) imitate teacher modality (vision), audio will inevitably reduce prediction confidence [39] and conflict *one-hot* label (Y). To validate our claim, we follow DKD [54] and decouple the KD loss into Target Class (TC) and Non-target Class (NC) KD:

$$\begin{aligned} \mathcal{D}(f_S, f_T) &= \alpha \underbrace{[p_T^t log(\frac{p_T^t}{p_S^t}) + p_T^{\backslash t} log(\frac{p_T^{\backslash t}}{p_S^{\backslash t}})]}_{\text{TCKD}} \\ &= \beta \underbrace{\sum_{i=1, i\neq t}^{C} \hat{p}_T^i log(\frac{\hat{p}_T^i}{\hat{p}_S^i})}_{\text{NCKD}} \end{aligned} \tag{2}$$

where α and β are hyperparameters. p^t denotes the target class probability: $p^t = \exp(f^t)/\sum_{j=1}^{C}\exp(f^j)$, $p^{\backslash t}$ represents the probability of all the other non-target classes $p^{\backslash t} = \sum_{k=1,k\neq t}^{C}\exp(f^i)/\sum_{j=1}^{C}\exp(f^j)$, and $\hat{p}^i$ means the probability among non-target classes: $\hat{p}^i = \exp(f^i)/\sum_{j=1,j\neq t}^{C}\exp(f^j)$. Here, C is the number of classes. When *only* applying TCKD in CMKD, as shown in Table 2, the performance of distilled audio modality severely degrades 4.8% and 3.6%, respectively, while the distilled visual modality is not clearly enhanced. Therefore, *modality imbalance* hinders the efficiency of CMKD, particularly when transferring knowledge from a low-accuracy modality to a high-accuracy modality.

To analyze *soft label misalignment*, we only conduct NCKD (Equation 2) to exclude the influence of modality imbalance. As depicted in Table 2, the low-accuracy visual teacher modality degrades the performance of the high-accuracy audio student modality. *Notably*, distilling high-accuracy audio information into the low-accuracy visual modality only results in marginal gains in the AVE, while surprisingly exhibiting a degradation in the VGGsound. [16, 26, 49, 54] investigate the mechanism of logit distillation, as soft logits provide reliable similarity information between categories. The privileged similarity information brings fine-grained supervision compared to a one-hot label. However, in the context of CMKD, the category similarities between different modalities are varied and even *conflicting*. An intuitive example is the three-class classification example in Figure 1(b) Up, where the unreliable similarity information of non-target classes across the modalities is contradictory. Directly minimizing cross-modal distributions leads to performance degradation. To quantitatively evaluate the misalignment of soft labels, we employ the Kendall Rank Correlation (KRC) [20] metric to measure the rank correlation. Specifically, given teacher and student output logits f_T and f_S, the KRC between f_T and f_S can be explicitly computed as follows:

$$\mathrm{KRC} = \frac{2}{C(C-1)}\sum_{i<j}\mathrm{sign}(f_T^i - f_T^j)\mathrm{sign}(f_S^i - f_S^j) \qquad (3)$$

As depicted in Figure 1(b) Down, the KRC between multimodal networks is significantly lower than that observed in unimodal networks with different capacities. We argue that the misalignment of rank correlation is another reason for the failure of CMKD. To validate our argument, we filter out multimodal samples with $\mathrm{KRC} < 0$ (+KRC), indicating that the count of misaligned soft label pairs is larger than aligned ones. Additionally, we randomly filter out the same number of samples (+Random). From Table 2, we can see that both visual and audio modalities are improved when guided by the KRC metric, whereas randomly filtering out samples has almost no effect.

	AVE [37]				VGGsound [5]			
Method	**Visual (A→V)**		**Audio (V→A)**		**Visual (A→V)**		**Audio (V→A)**	
w/o KD	31.6		52.8		38.7		59.4	
proba.	0.355		0.901		0.340		0.534	
w/ KD	32.3	↑0.7	46.6	↓6.2	38.5	↓0.2	56.3	↓3.1
+Random	32.1	-0.2	46.8	+0.2	38.2	-0.3	56.4	+0.1
+KRC	32.9	+0.6	47.9	+1.3	39.2	+0.7	57.4	+1.1
TCKD	31.8	↑0.2	48.0	↓4.8	37.9	↓0.8	55.8	↓3.6
NCKD	31.9	↑0.3	50.1	↓2.7	38.5	↓0.2	57.5	↓1.9
+Random	31.5	-0.4	50.2	+0.1	38.5	-	57.6	+0.1
+KRC	33.1	+1.2	51.0	+0.9	39.6	+1.1	58.3	+0.8
DKD [54]	32.6	↑1.0	48.6	↓4.2	38.1	↓0.6	57.2	↓2.2

Table 2. **Efficacy Analysis on modality imbalance and soft label misalignment.** *proba.* represents average prediction probability of target class. DKD is with defaulted $\{\alpha = 1, \beta = 8\}$ (Eq. 2).

3.2. Customized Cross-modal KD

Based on the aforementioned analysis, we propose a simple yet effective method named Customized Cross-modal Knowledge Distillation (C^{2}KD) to transfer cross-modal knowledge to an arbitrary single modality. To bridge the modality gap, we argue that both student and teacher should be tuned with the bidirectional distillation from each other, in this way, teacher modality could provide receptive information for student modality. Meanwhile, the soft label misalignment samples should be filtered out otherwise induce conflicting information. Therefore, we propose the On-the-Fly Selection Distillation (OFSD) strategy to exclude non-distillable samples and inherit knowledge from non-target classes. Furthermore, dual proxies with the bidirectional distillation strategy are introduced to progressively transfer cross-modality knowledge. The evolution of our proposed framework is depicted in Figure 2.

Formulation of C^{2}KD. As illustrated in Figure 2(d), C^{2}KD proposes the OFSD strategy to dynamically select receptive knowledge. This strategy involves distilling knowledge from non-target classes and innovatively employing the Kendall Rank Correlation (KRC) [20] metric to filter out samples with rank-distorted soft labels. Given the output logits f_T and f_S from the teacher and student modalities, the sample selection strategy is as follows:

$$\eta = \begin{cases} 1, & \mathrm{KRC}(f_T, f_S) > \omega \\ 0, & otherwise \end{cases} \qquad (4)$$

The KRC is as Equation 3, $\eta \in \{0, 1\}$ is OFSD filter, and ω is the threshold.

Moreover, we additionally build dual proxies to progressively produce soft labels. Formally, the output features (F) obtained from the backbone (B) are fed to the original classification head and the proposed proxy as follows:

$$\begin{aligned} f_m &= fc_m^{cls}(\mathrm{GAP}(B_m(F_m))), m \in \{\mathrm{T}, \mathrm{S}\} \\ f_m^{pro} &= fc_m^{cls(pro)}(\mathcal{A}[\mathrm{GAP}(B_m(F_m))]), m \in \{\mathrm{T}, \mathrm{S}\} \end{aligned} \qquad (5)$$

where GAP and fc^{cls} refer to global average pooling and classification head. $\mathcal{A}$ represents feature adaptation layer,

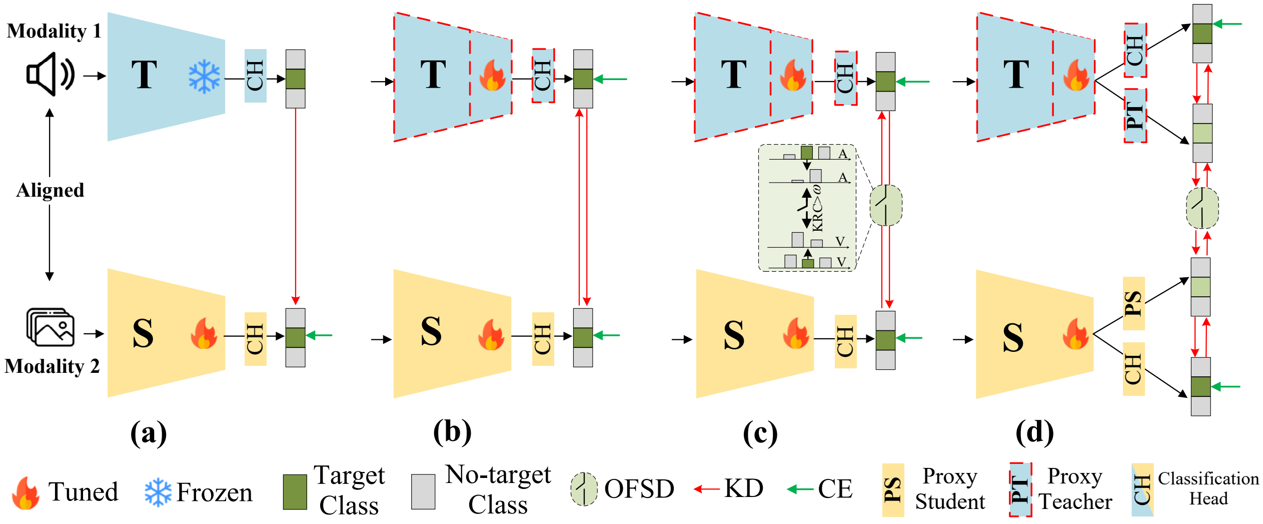


Figure 2. **Evolution of our Customized Cross-modal Knowledge Distillation (C^2KD) method.** (a) Traditional KD [16] with output logits from the fixed teacher. (b) We (partially) tune the teacher with the bidirectional distillation to provide customized teacher knowledge. (c) To bridge the modality gap of CMKD analyzed in Section 3.1, On-the-Fly Selection Distillation (OFSD) is proposed to filter out samples with distorted rank correlations and perform KD on non-target classes. (d) Additionally, we introduce proxy teacher and proxy student as bridges to progressively transfer receptive cross-modal knowledge.

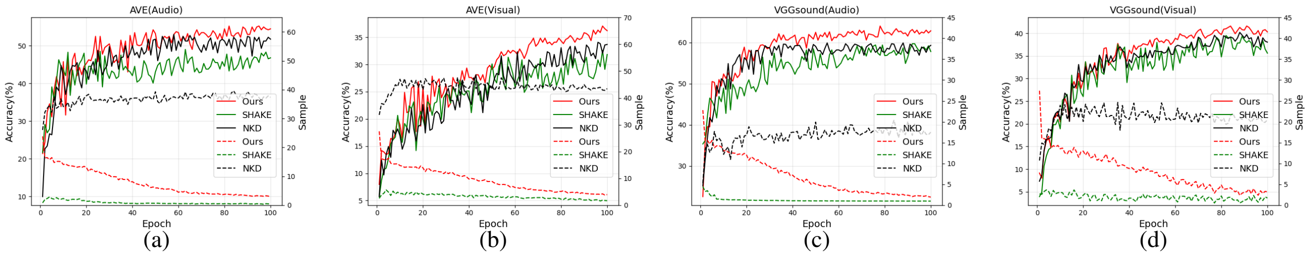
akin to [25, 33], consisting of the 'Conv-BN-ReLU' block. To further produce customized knowledge, both student and teacher proxies serve as *bridges* and get bidirectional distillation from both uni-modality and cross modality. In summary, the total loss function can be expressed as follows:

$$\begin{aligned} L_{all} =& \mathcal{H}(\sigma(f_S), Y) + \mathcal{H}(\sigma(f_T), Y) \\ &+ \lambda_1 \mathcal{D}(\sigma(f_T), \sigma(f_T^{pro})) + \lambda_1 \mathcal{D}(\sigma(f_T^{pro}), \sigma(f_T)) \\ &+ \lambda_2 \mathcal{D}(\sigma(f_S), \sigma(f_S^{pro})) + \lambda_2 \mathcal{D}(\sigma(f_S^{pro}), \sigma(f_S)) \\ &+ \lambda_3 \eta \mathcal{D}(\sigma(\hat{f}_S^{pro(i)}), \sigma(\hat{f}_T^{pro(i)})) \\ &+ \lambda_3 \eta \mathcal{D}(\sigma(\hat{f}_T^{pro(i)}), \sigma(\hat{f}_S^{pro(i)})) \end{aligned} \quad (6)$$

where λ_1, λ_2, and λ_3 are balancing parameters and $i \neq t$. $\mathcal{H}$ and $\mathcal{D}$ represent supervision and KD loss, respectively. We simply set $\{\lambda_1 = \lambda_2 = \lambda_3 = 1\}$ in all experiments.

Understanding CMKD training dynamics. We visualize the training dynamics of CMKD and compare it with SHAKE [25] and NKD [47] to demonstrate the CMKD progress. Figure 3 shows the test accuracy and the average number of samples with KRC $<\omega$ ($\omega = 0$) during the training process. As the advanced online KD, SHAKE gets the reverse cross-modal feedback supervision without discrimination. However, SHAKE suffers from severe instability of training, possibly due to conflicting cross-modal information. Meanwhile, the sample number of KRC $<\omega$ drops to close to 0 within initial epochs, which represents the teacher modality is influenced by the student modality and might lose teacher modality information. In contrast, NKD minimizes the distance between student modality logits and teacher modality logits. The teacher model of NKD is not updated to cater to student modality, so the sample number of KRC $<\omega$ is large, and NKD also falls into the unstable training process. As for ours, we selectively inherit cross-modal knowledge based on KRC and progressively update the teacher model through proxies to obtain receptive knowledge. During the distillation progress, the rank-distorted samples gradually reduce, and our method only filters out the non-distillable samples.

Comparisons with other distance metrics. We select other distance metrics to verify the effectiveness of KRC defined in Equation 4. Concretely, we choose the cosine similarity (Cos), gradient cosine similarity (GradCos), and Pearson correlation coefficient [29] (Pearson) as alternatives. Cosine similarity and Pearson correlation coefficient are used to measure the distance between teacher and student logits, ranging from -1 to 1. They can be formulated as: $\amalg(f_T, f_S) > \omega$, $\amalg \in \{\mathrm{Cos}; \mathrm{Pearson}\}$. Similar to [48, 55], Gradient cosine similarity regards CMKD (Equation 1) as two tasks: cross-modal distillation ($L_{cmkd} = \mathcal{D}(p_S, p_T)$) and unimodal task ($L_{task} = \mathcal{H}(p_S, Y)$) and calculates the gradient cosine similarity between these two tasks as: $\mathrm{Cos}(\nabla_\theta L_{cmkd}, \nabla_\theta L_{task}) > \omega$. It's worth noting that these three metrics consider *both* rank and intensity between cross-modal logits, while KRC *only* concerns about rank-distorted ones. As shown in Table 4, KRC makes the best performance among these metrics. Although inferior to KRC metric, other metrics with proper ω perform better than without sample selection (w/o Selection) strategy. The results validate the necessity of filtering out samples with misaligned soft labels.



(a) (b) (c) (d)

Figure 3. **Training dynamics analysis.** The solid lines correspond to the test accuracy, and the dotted lines indicate the average number of samples with KRC $< \omega$ each data batch during the training process. Here we set $\{\omega = 0, \text{batchsize} = 64\}$.

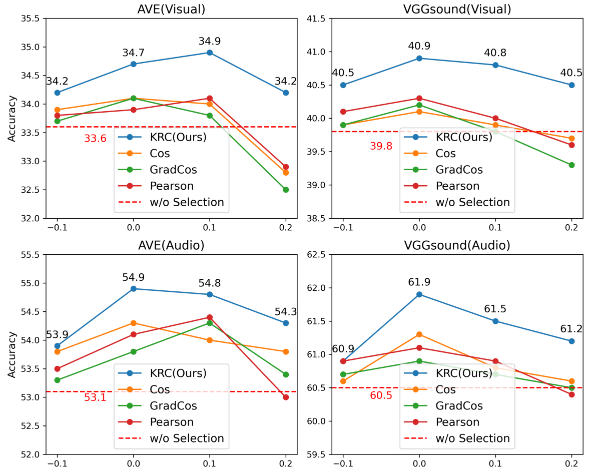


Figure 4. **Comparisons of different distance metrics.** The X-axis represents the value of ω.

4. Experiments

We conduct extensive experiments to validate the effectiveness of our method. First, we compare our method with KD methods regarding multimodal classification tasks. Also, we apply our method to the multimodal semantic segmentation. Then, we perform ablation and sensitivity analysis.

4.1. Multimodal Classification

We follow [1, 11, 30] and conduct experiments on four visual-audio and image-text datasets: (1) **CREMA-D** [4] is an audio-visual dataset for speech emotion recognition, with 6 categorizations. (2) **AVE** [37] is an audio-visual dataset for audio-visual event localization, in which there are 28 event classes. (3) **VGGsound** [5] is a large-scale video dataset containing 309 classes covering daily life activities. We randomly choose *50* class to conduct experiments. (4) **CrisisMMD** [2] is a multimodal crisis prediction dataset and is divided into eight humanitarian categories. Details of these four datasets are in the Appendix A.

Implementation. For *visual-audio* datasets, the preprocess strategy follows [11, 30] and detailed implementations are in the Appendix C. We train the network for 100 epochs with 1e-2 initial learning rate and decay follow the 'poly' policy with the power of 0.9. We use SGD with 0.9 momentum and default hyperparameters as the optimizer. For the *image-text* dataset, we use the same training strategies and adopt $\omega = 0$ across all experiments. Here, following [1, 30], we adopt the same ResNet-18 [14] as the backbone for visual and audio modality, and BERT-base [9] for text and MobileNetV2 [34] and image feature extractors, respectively. More experiments on diverse-capacities homogeneous and heterogeneous architectures are in the Appendix B. All results are the average of three different seeds.

Comparison Results. In Table 3, we compare our method to some advanced KD methods with the same training settings. Details of implementations of compared methods are in the Appendix C. We can learn from Table 3 that our proposed method, C^{2}KD, consistently outperforms other KD methods across four datasets. Existing KD methods can not effectively distill one modality information to another modality, especially for the datasets with the significant modality imbalance issue like AVE and VGGsound. Concretely, feature-based KD (FitNet[33], Review[7]) methods fail in CMKD because of significant feature divergence (see Section 5). Online KD (DML [53] and SHAKE [25]) methods update teacher models and achieve better cross-modal knowledge transfer ability, compared with the baseline [16]. Due to soft label misalignment between modalities, the relation-based method (RKD [28]) degrades severely in CMKD. Recent advanced logits-based methods (DKD [54], DIST [17], and NKD [47]) significantly outperform the vanilla KL loss by proposing the relaxed KD function and logits decoupling strategies. However, these methods fail to transfer knowledge from low-accuracy to high-accuracy modality, impeding their practical deployments in CMKD.

4.2. Multimodal Semantic Segmentation

We also extend C^{2}KD to the multimodal semantic segmentation, a challenging dense prediction task. Concretely, following [42], we conduct experiments on the NYU-Depth V2 dataset [36]. NYU-Depth V2 contains 1,449 aligned RGB and depth pairs with 40 category labels, of which 795 pairs are used for training, and 654 pairs are used for testing.

Implementation. Both teacher and student networks deploy the DeepLab V3+ [6] architecture with diverse back-

	CREMA-D		AVE		VGGsound		CrisisMMD	
Method	**Visual**	**Audio**	**Visual**	**Audio**	**Visual**	**Audio**	**Image**	**Text**
w/o KD	58.1±0.33	56.3±0.22	31.6±0.18	52.8±0.11	38.7±0.16	59.4±0.16	66.7±0.22	68.1±0.21
FitNet[33]	56.4±0.47	52.9±0.32	29.6±0.63	48.0±0.81	37.9±0.39	57.1±0.79	-	-
Review[7]	59.6±0.45	55.7±0.36	32.1±0.63	50.6±0.31	38.2±0.47	57.9±0.33	-	-
KD[16]	57.4±0.92	53.4±0.85	32.3±0.35	46.6±0.24	38.5±0.50	56.3±0.46	66.3±0.24	68.4±0.12
DML[53]	60.3±1.60	56.4±0.55	31.8±0.41	48.0±1.31	38.7±0.86	58.2±1.01	67.9±0.18	69.6±0.24
SHAKE[25]	60.0±0.35	58.6±0.61	32.2±0.59	47.3±0.72	38.3±0.41	59.5±0.34	68.1±0.16	69.7±0.26
RKD[28]	48.3±0.68	51.9±1.36	28.2±0.71	44.5±0.73	33.4±0.49	41.5±1.36	67.0±0.23	67.4±0.21
DKD [54]	60.4±0.82	55.1±0.65	32.6±0.65	48.6±1.02	38.1±0.43	57.2±0.86	68.0±0.17	69.2±0.23
DIST [17]	61.1±1.82	57.9±0.57	29.8±0.61	49.3±0.29	38.5±0.39	58.9±0.45	68.3±0.21	67.8±0.18
NKD [47]	60.6±0.64	56.1±0.68	32.9±0.32	52.2±0.62	39.2±0.52	59.3±0.40	67.2±0.26	68.5±0.16
Ours†	62.4±0.24	60.5±0.37	34.2±0.28	54.5±0.22	40.8±0.23	61.6±0.34	68.2±0.09	69.8±0.16
Ours‡	**62.8**±0.28	**61.4**±0.44	**34.7**±0.23	**54.9**±0.16	**40.9**±0.31	**61.9**±0.27	**68.8**±0.15	**70.1**±0.12

Table 3. **Comparison results on Visual-Audio and Image-Text datasets.** The metric is the top-1 accuracy (%). Ours‡ means fully updating the teacher model, and Ours† means partially finetuning the top 2 layers. The best is in **bold**, and the second is underlined.

	RGB RN18	**Depth** RN18	**RGB** RN18	**Depth** MNV2	**RGB** MNV2	**Depth** RN18
w/o KD	36.1	30.5	36.1	31.2	36.3	30.5
KD [16]	35.8	30.9	36.2	31.9	36.5	31.8
SHAKE [25]	37.1	31.2	37.0	32.7	37.1	32.9
DIST [17]	36.9	32.0	36.5	32.9	36.8	33.1
NKD [47]	36.5	30.8	36.4	32.2	36.4	32.7
CIRKD [44]	37.3	32.6	36.9	32.7	36.7	33.4
Ours	37.5	32.5	37.2	32.8	37.4	33.1
Ours+[17]	**38.1**	**33.2**	**37.7**	**33.5**	**37.9**	**33.7**

Table 4. **Comparison results on RGB-Depth semantic segmentation dataset.** The metric denotes the mean Intersection over Union (mIoU: %). Ours+[17] means we replace KL loss with the advanced DIST loss. RN18: ResNet-18; MNV2: MobileNetV2.

bones. The training settings follow [44] that we adopt SGD as the optimizer with a momentum of 0.9, a batch size of 16, an initial learning rate of 0.02, and ImageNet pre-trained weights. The total training iterations is 40K, decayed by the 'poly' policy with the power of 0.9. Experiments on homogeneous/heterogeneous backbones, including ResNet-18/ResNet-18 and ResNet-18/MobileNetV2 pairs, are conducted to validate our method. All results are the average of three different seeds.

Comparison Results. We compare our methods with advanced traditional KD methods (KD [16], SHAKE [25], DIST [17], and NKD [47]) as well as the semantic segmentation KD method (CIRKD [44]). The results compared with previous methods are summarized in Table 4. We can see that previous KD methods do not perform well in CMKD, especially in transferring low-accuracy modality information to high-accuracy modality. Our method can significantly improve the distilled performance of arbitrary single modality. For instance, ours consistently surpasses the advanced CIRKD in transferring depth information to RGB modality. Besides, when replacing KL loss with DIST loss [17], our method affords clear improvements.

4.3. Ablation and Sensitivity Analysis

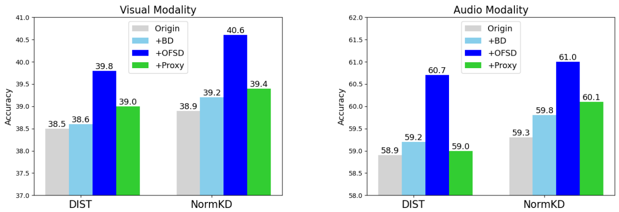
Effectiveness and generalizability of each module. We analyze how the proposed modules improve CMKD. Table 5 reports the results of ablation studies on AVE, VGGsound, and CrisisMMD with the same backbones. The configurations of (a), (b), (c), and (d) correspond to the evolution steps shown in Figure 2. Compared to the vanilla KD (i.e., (a)), the Bidirectional Distillation (BD) updates the teacher model (i.e., (b)) to mitigate the model gap. Furthermore, to validate the effectiveness of OFSD, we decouple OFSD into the On-the-Fly Selection (OFS) strategy and Non-Target (NT) classes distillation approach. We can learn that both OFS and NT benefit CMKD, and the combination of both brings significant improvement compared to (b). The proxy teacher and student circumvent the direct imitation of cross-modal logits, serving as bridges for inheriting unimodal and cross-modal knowledge and facilitating the transfer of integrated knowledge through bidirectional distillation. The progressive KD strategy further improves the CMKD results. The structure of proxies adheres to [25, 33]. We provide analyses about the proxies and computation overhead in the Appendix D&E.

	Proxy	OFS	NT	BD	AVE		VGGsound		CrisisMMD	
					Visual	**Audio**	**Visual**	**Audio**	**Image**	**Text**
(e)					31.6	52.8	38.7	59.4	66.7	68.1
(a)					32.3	46.6	38.5	56.3	66.3	69.2
(b)				✓	32.7	47.9	38.8	57.6	67.3	69.2
(c)		✓	✓	✓	34.6	54.3	40.4	61.5	68.5	69.8
-			✓	✓	33.2	52.9	39.4	60.3	67.9	68.9
-		✓		✓	34.4	52.5	40.0	59.9	68.0	69.5
(d)	✓	✓	✓	✓	34.7	54.9	40.9	61.9	68.8	70.1

Table 5. **Ablation studies on each module.** (a), (b), (c), and (d) represent the evolution steps of C^{2}KD (Figure 2). (e) indicates the results without KD. The metric is the top-1 accuracy (%).

To ascertain the generalizability of each component, we incorporate the proposed plug-and-play modules into advanced KD methods (i.e., DIST [17] and NKD [47]). We can learn from Figure 5 that the proposed modules consistently improve the performances of traditional KD methods in the CMKD task, especially for the OFSD strategy.

Necessity of cross-modal KD. Considering the challenges of cross-modal KD, a question may arise: Do we really need CMKD rather than fully explore self-knowledge? Self-knowledge distillation (Self-KD) techniques [24, 26, 35, 43, 47] have been proposed to utilize the information within



(a) T: audio; S: visual (b) T: visual; S: audio

Figure 5. **Generalizability of each module.** We conduct experiments on VGGsound dataset in terms of DIST [17] and NKD [47].

	AVE		VGGsound		CrisisMMD	
Method	**Visual**	**Audio**	**Visual**	**Audio**	**Image**	**Text**
w/o KD	31.6	52.8	38.7	59.4	66.1	68.1
DLB[35]	32.6	53.3	39.1	60.2	66.9	68.6
ZipfKD[26]	33.3	53.5	40.2	60.3	67.4	68.9
USKD[47]	33.1	53.2	40.0	60.1	67.1	69.0
Ours	34.7	54.9	40.9	61.9	68.8	70.1

Table 6. **Comparison results of different Self-KD methods.**

the student model to facilitate its learning process. Specially, DLB [35] leverages the soft targets generated in the last mini-batch backup for training consistency and stability. ZipfKD [26] and USKD [47] generate soft labels following the Zipf's law distribution [27]. We conduct experiments on these advanced Self-KD methods to validate the necessity of CMKD. From Table 6, we can learn that although Self-KD improves the performance of each modality, our method consistently outperforms Self-KD methods by a clear margin. The results indicate the necessity of cross-modal KD.

Parameter sensitivity. Here, we conduct a sensitivity study on KRC threshold ω. Results are in Figure 4. Large ω filters out more samples, which might hinder cross-modal knowledge transfer, while low ω preserves more samples, which might contain rank-distorted samples that induce adverse effects. We heuristically set ω to 0 and achieve balanced results. Note that, as shown in Tables 3 and 5, even the worst accuracy of varying ω is still competitive with the baselines, we think the studies show the necessity of on-the-fly filtering out rank-distorted samples based on KRC. More analyses about λ_1, λ_2, and λ_3 are in the appendix.

5. Discussion

Feature-based CMKD Analysis. We analyze the modality gap of CMKD in the *logits-based* perspective and propose C^{2}KD to mitigate the issues. Furthermore, we provide the analysis of the challenge of CMKD from the *feature-based* perspective. We adopt the Center Kernel Alignment (CKA) [21], a feature similarity metric that measures input similarity with different dimensions. As shown in Figure 6, compared to unimodal features, cross-modal features have significant feature divergence, and directly using feature-based distillation methods for CMKD is unreasonable. We leave the exploration of feature-based CMKD as the future work.

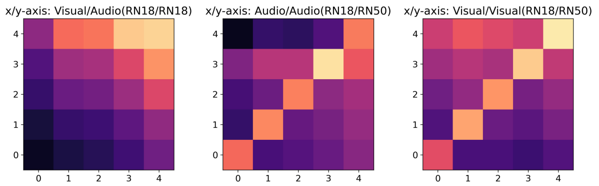


Figure 6. **The CKA score of intermediate features on AVE.**

	AVE				VGGsound			
Method	**V**(S)	T	**A**(S)	T	**V**(S)	T	**A**(S)	T
NKD [47]	32.9	*52.8*	52.2	*31.6*	39.2	*59.4*	59.3	*38.7*
+Cat	34.0	*59.6*	52.0	*60.2*	39.9	*62.9*	59.9	*63.2*
+FiLM [31]	33.2	*57.4*	51.7	*57.6*	39.0	*62.1*	58.6	*62.8*
+OGM [30]	33.6	*60.9*	52.7	*61.6*	40.2	*65.2*	59.8	*64.3*
Ours	34.7	*54.2*	54.9	*34.6*	40.9	*59.0*	61.9	*41.6*

Table 7. **Comparison results of different multimodal teachers.** The *italic* numbers mean teachers' accuracy. S: student; T: teacher.

Multimodal Teacher Efficacy Analysis. We provide analysis of the efficacy of multimodal teacher in CMKD. Concretely, the multimodal teacher is formulated by fusing the teacher and the student modalities with the supervision loss. Following the multimodal learning [10, 11, 30], we adopt the fusion strategies including Concatenation (Cat), FiLM [31], and OGM [30]. Pretrained teachers are updated for the better information fusion. Table 7 indicates multimodal teachers generate high-accuracy soft labels, while don't necessarily improve the distilled modality, especially for the high-accuracy modality (i.e., audio). Our customized teacher integrates receptive corssmodal information and ensures effective knowledge transfer. Inspired by [42], developing the multimodal learning method that contains more modality-general decisive information is a possible solution. We leave this intriguing challenge to future work.

6. Conclusion

In this paper, we conduct thorough investigation toward the efficacy of CMKD and reveal that the modality imbalance and soft label misalignment induced by the intermodality gap are the main factors for the failure of CMKD. Based on our analyses, we propose a simple yet effective method, **C**ustomized **C**rossmodal **K**nowledge **D**istillation (C^{2}KD). Specifically, we propose On-the-Fly Sample Selection (OFSD) strategy to filter out rank-distorted samples based on the KRC metric and distill knowledge from non-target classes. Meanwhile, the pre-trained teacher conducts bidirectional distillation with the student. Proxy student and teacher, inheriting unimodal and cross-modal knowledge, progressively transfer cross-modal knowledge. Extensive experiments demonstrate the effectiveness of our method.

7. Acknowledgments

The work described in this paper was substantially supported by two grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU15222621, PolyU15225023).

References

[1] Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. Multimodal categorization of crisis events in social media. In *CVPR*, 2020. 6

[2] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. *AAAI*, 2018. 6

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 2019. 3

[4] 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. 6

[5] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 1, 4, 6

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 6

[7] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, 2021. 1, 2, 6, 7

[8] Inseop Chung, Seonguk Park, Jangho Kim, and Nojun Kwak. Feature-map-level online adversarial knowledge distillation. In *ICML*, 2020. 3

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv:2305.15712*, 2018. 6

[10] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *ICML*, 2023. 3, 8

[11] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *CVPR*, 2023. 1, 3, 6, 8

[12] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *IJCV*, 2021. 1

[13] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *CVPR*, 2016. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[15] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, 2019. 1

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015. 1, 2, 3, 4, 5, 6, 7

[17] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *NeurIPS*, 2022. 1, 2, 6, 7, 8

[18] Tao Huang, Yuan Zhang, Mingkai Zheng, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge Diffusion for Distillation. *arXiv:2305.15712*, 2023. 1

[19] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). In *NeurIPS*, 2021. 3

[20] M. G. Kendall. Rank correlation methods. In *Griffin, Oxford, England, 1948*. 2, 4

[21] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. 8

[22] Xu Lan, Xiatian Zhu, and Shaogang Gong. Knowledge distillation by on-the-fly native ensemble. In *NeurIPS*, 2018. 3

[23] Pilhyeon Lee, Taeoh Kim, Minho Shim, Dongyoon Wee, and Hyeran Byun. Decomposed cross-modal distillation for rgb-based temporal action detection. In *CVPR*, 2023. 3

[24] Lujun Li. Self-regulated feature learning via teacher-free feature distillation. In *ECCV*, 2022. 7

[25] Lujun Li and Zhe Jin. Shadow knowledge distillation: Bridging offline and online knowledge transfer. In *NeurIPS*, 2022. 1, 3, 5, 6, 7

[26] Jiajun Liang, Linze Li, Zhaodong Bing, Borui Zhao, Yao Tang, Bo Lin, and Haoqiang Fan. Efficient one pass self-distillation with zipf's label smoothing. In *ECCV*, 2022. 2, 4, 7, 8

[27] David M. W. Powers. Applications and explanations of zipf's law. In *New methods in language processing and computational natural language learning*, 1998. 8

[28] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019. 2, 6, 7

[29] Karl Pearson. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. In *Philosophical Transactions of the Royal Society of London*, 1896. 5

[30] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, 2022. 1, 3, 6, 8

[31] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. *AAAI*, 2018. 8

[32] Siddharth Roheda, Benjamin S. Riggan, Hamid Krim, and Liyi Dai. Cross-modality distillation: A case for conditional generative adversarial networks. In *ICASSP*, 2018. 3

[33] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015. 5, 6, 7

[34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 6

[35] Yiqing Shen, Liwu Xu, Yuzhe Yang, Yaqian Li, and Yandong Guo. Self-distillation from the last mini-batch for consistency regularization. In *CVPR*, 2022. 7, 8

[36] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 6

[37] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. 1, 4, 6

[38] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020. 2

[39] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 3

[40] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE TPAMI*, 2022. 1, 3

[41] Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. Multimodal knowledge expansion. In *ICCV*, 2021. 3

[42] Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. In *ICLR*, 2023. 3, 6, 8

[43] Chuanguang Yang, Zhulin An, Helong Zhou, Linhang Cai, Xiang Zhi, Jiwen Wu, Yongjun Xu, and Qian Zhang. Mixskd: Self-knowledge distillation from mixup for image recognition. In *ECCV*, 2022. 7

[44] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *CVPR*, 2022. 7

[45] Jing Yang, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. Knowledge distillation via softmax regression representation learning. In *ICLR*, 2021. 2

[46] Penghui Yang, Ming-Kun Xie, Chen-Chen Zong, Lei Feng, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Multi-label knowledge distillation. In *ICCV*, 2023. 3

[47] 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, 5, 6, 7, 8

[48] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *NeurIPS*, 2020. 5

[49] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *CVPR*, 2020. 4

[50] Heeseung Yun, Joonil Na, and Gunhee Kim. Dense 2d-3d indoor prediction with sound via aligned cross-modal distillation. In *ICCV*, 2023. 3

[51] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. *ICML*, 2023. 3

[52] Tianlu Zhang, Hongyuan Guo, Qiang Jiao, Qiang Zhang, and Jungong Han. Efficient rgb-t tracking via cross-modality distillation. In *CVPR*, 2023. 3

[53] Ying Zhang, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, 2018. 1, 3, 6, 7

[54] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, 2022. 1, 2, 3, 4, 6, 7

[55] Yichen Zhu and Yi Wang. Student customized knowledge distillation: Bridging the gap between student and teacher. In *ICCV*, 2021. 2, 5

[56] Yichen Zhu, Ning Liu, Zhiyuan Xu, Xin Liu, Weibin Meng, Louis Wang, Zhicai Ou, and Jian Tang. Teach less, learn more: On the undistillable classes in knowledge distillation. In *NeurIPS*, 2022. 2