Enriching Knowledge Distillation with Cross-Modal Teacher Fusion

Amir M. Mansourian Amir Mohammad Babaei Shohreh Kasaei
Image Processing Lab, Sharif University of Technology
Tehran, Iran
{amir.mansurian, amir.babaei79, kasaei}@sharif.edu

Abstract

*Multi-teacher knowledge distillation (KD), a more effective technique than traditional single-teacher methods, transfers knowledge from expert teachers to a compact student model using logit or feature matching. However, most existing approaches lack knowledge diversity, as they rely solely on unimodal visual information, overlooking the potential of cross-modal representations. In this work, we explore the use of CLIP's vision–language knowledge as a complementary source of supervision for KD, an area that remains largely underexplored. We propose a simple yet effective framework that fuses the logits and features of a conventional teacher with those from CLIP. By incorporating CLIP's multi-prompt textual guidance, the fused supervision captures both dataset-specific and semantically enriched visual cues. Beyond accuracy, analysis shows that the fused teacher yields more confident and reliable predictions, significantly increasing confident-correct cases while reducing confidently wrong ones. Moreover, fusion with CLIP refines the entire logit distribution, producing semantically meaningful probabilities for non-target classes, thereby improving inter-class consistency and distillation quality. Despite its simplicity, the proposed method, En**Rich**ing **K**nowledge **D**istillation (RichKD), consistently outperforms most of existing baselines across multiple benchmarks and exhibits stronger robustness under distribution shifts and input corruptions. Code is available at: https://github.com/IPL-sharif/RichKD*

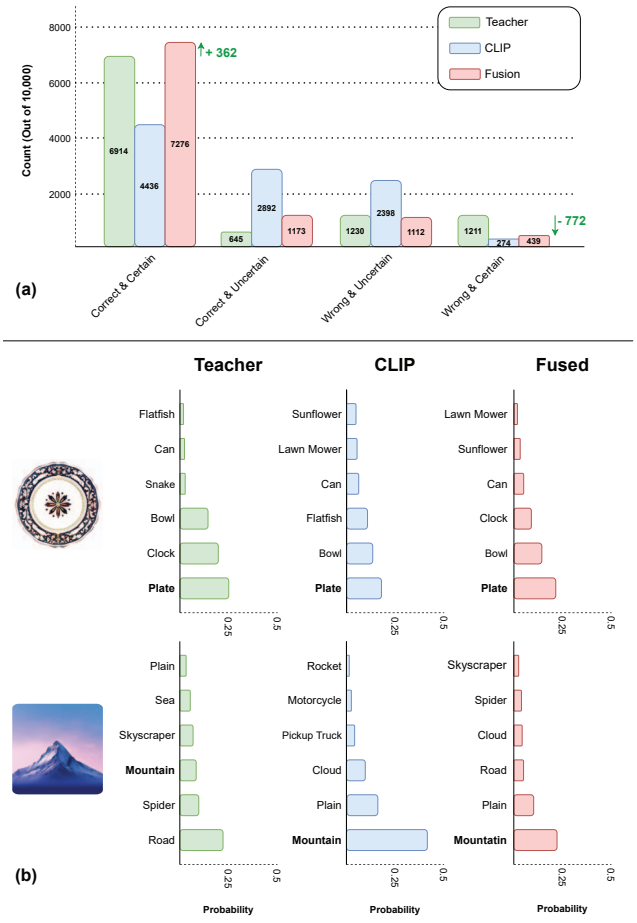

Figure 1. **Impact of cross-modal teacher fusion on CIFAR-100.** (a) Effect of perturbing the logits of the conventional teacher with CLIP's logits across four categories, considering cases where the teacher's predictions are correct/incorrect and certain/uncertain. (b) Effect of fusion with CLIP for two sample cases: when the teacher is incorrect, and when the teacher is correct but uncertain.

1. Introduction

Knowledge Distillation (KD) [9, 32] has become one of the most effective strategies for compressing deep neural networks, enabling lightweight student models to inherit the representational power of larger teachers. By transferring knowledge from a well-trained teacher through logit or feature alignment, KD has achieved success across image recognition, object detection, and many other domains. Despite these advances, most existing KD methods remain strictly *unimodal*, relying solely on visual cues learned from the target dataset. As a result, the distilled knowledge often inherits the teacher's task-specific biases and lacks the se-

mantic diversity necessary for broader generalization.

In contrast, large-scale vision–language models such as CLIP [38] capture rich cross-modal knowledge by jointly learning from hundreds of millions of image–text pairs. CLIP's training paradigm grounds visual concepts in linguistic semantics, enabling zero-shot classification and robust generalization across diverse datasets. However, while CLIP has revolutionized transfer and retrieval tasks, its potential as a *cross-modal teacher* for knowledge distillation remains largely underexplored [62]. Integrating CLIP's multimodal understanding into a KD framework presents a promising avenue for transferring both dataset-specific and semantically enriched knowledge to compact visual students.

This paper introduces En**Rich**ing **K**nowledge **D**istillation (RichKD), a novel cross-modal distillation framework that fuses the loits and features of a conventional teacher with those of a pre-trained CLIP model. Our key insight is that CLIP's predictions, **though not tailored to the target dataset**, encapsulate complementary information in their broader semantic space. By fusing CLIP's logits and features with those of the primary teacher, we effectively inject meaningful cross-modal perturbations into the supervision signal. This fusion acts as a soft ensemble of two diverse teachers; one grounded in dataset-specific discriminative patterns, and the other guided by semantic relationships learned from large-scale vision–language data. Figure 1(a) illustrates the number of certain/uncertain and correct/incorrect predictions for the conventional teacher, CLIP, and their fusion. The fusion improves overall accuracy, increasing confident correct predictions by 3.6% and reducing confident mistakes by 7.7%. Figure 1(b) presents two examples illustrating that perturbing the teacher's logits with CLIP's helps correct wrong predictions and refine the confidence distribution by better ranking non-target classes.

To enrich the multimodal supervision, we further employ a *multi-prompting strategy*, generating multiple textual templates per class through CLIP's text encoder and averaging the corresponding predictions. This multi-prompt ensemble captures different linguistic perspectives, smoothing the supervision and mitigating overfitting to specific textual formulations. The fused signals are then distilled into the student through a logit distillation loss.

In summary, this paper makes the following contributions:

- We investigate the integration of cross-modal knowledge into conventional knowledge distillation, an underexplored direction that leverages CLIP's vision–language representations to enrich the teacher–student learning process.
- We propose a simple yet effective *logit–feature fusion* mechanism that combines a dataset-specific teacher with CLIP's multi-prompt predictions to construct semantically diverse and informative supervisory signals.
- We provide both theoretical and empirical evidence that this cross-modal fusion enhances the diversity of teacher guidance, yielding improved performance and robustness compared to unimodal KD methods.

2. Related Work

In this section, literature related to this work, including KD and multi-teacher KD, is presented.

2.1. Knowledge Distillation

Knowledge distillation was popularized by KD [15], who showed that a compact student can learn from a larger teacher by matching softened teacher logits. Following KD, FitNet [39] proposed to distill intermediate features, while RKD [37] focused on transferring relational knowledge to the student. Subsequent works can be categorized into three groups: logit-based [7, 10, 12, 25, 33, 41, 50, 52, 61], feature-based [5, 11, 14, 28, 30, 58], and similarity-based [17, 27, 49, 51, 54, 55] distillation methods. Furthermore, some studies have sought to enhance the distillation process using various techniques, such as adaptive distillation [3, 18, 31, 35, 36, 63] or modifying the teacher's logits using label information [2, 22, 46, 53].

Although feature-based methods such as CRD [48] and SimKD [4] have achieved strong performance in KD, the dominant paradigm for classification remains logit distillation. Numerous logit-based methods have been proposed to improve the vanilla KD approach [6, 20, 26, 45, 61]. For example, DKD [61] decouples the KD loss into target and non-target class components; MLKD [20] introduces multi-level logit distillation; and CTKD [26] adopts a curriculum temperature strategy for logit distillation. More recently, NormKD [6] customizes the temperature for each sample, and LSKD [45] applies Z-score standardization to logits before the softmax operation.

2.2. Multi-Teacher Distillation

Using multiple teachers or auxiliary teacher assistants is a natural approach to provide richer supervision and to reduce the teacher–student capacity gap. Several strategies have been employed for this purpose, such as averaging logits [15], using voting strategies [57, 60], perturbing logits [16, 42], or assigning teachers that specialize in different subsets of the dataset [19]. TAKD [34] was one of the early works that introduced the concept of a teacher assistant, while CA-MKD [59] proposed a confidence-aware multi-teacher distillation framework, and DGKD [44] presented a densely guided multi-teacher approach. More recently, MLDF [19] proposed multi-level feature distillation, in which each teacher is specialized on a distinct dataset,

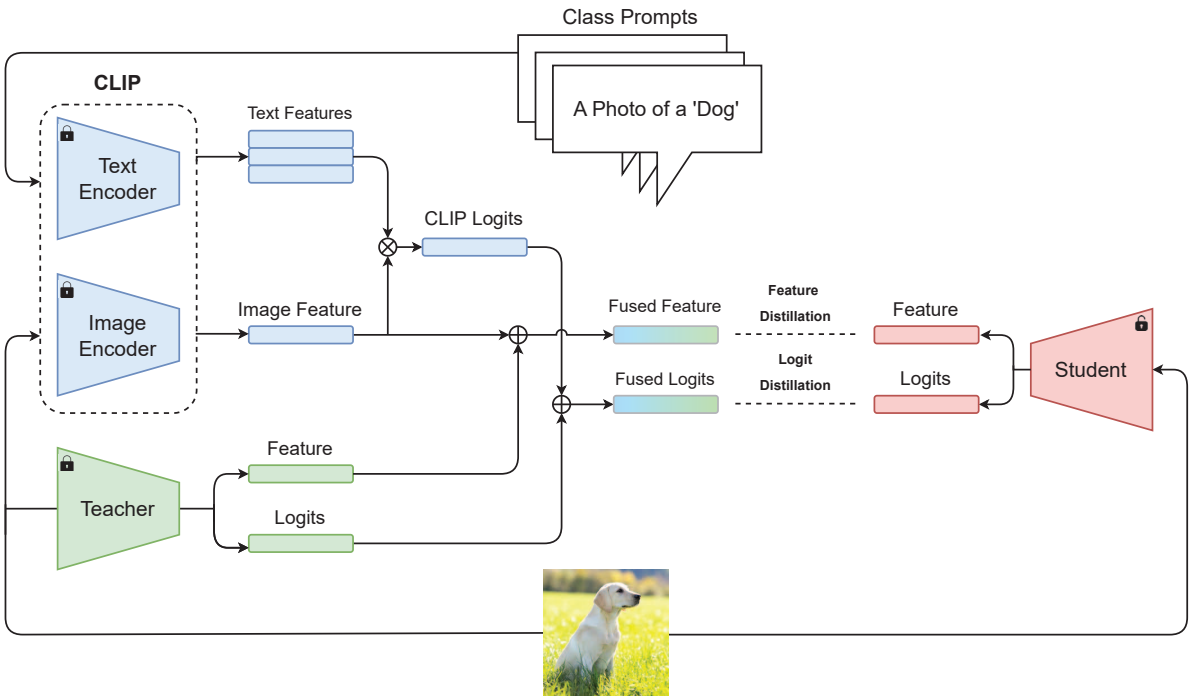

Figure 2. **Overall diagram of the proposed RichKD distillation method**. CLIP's logits and features are fused with those from the conventional teacher model. Feature and logit distillation losses are then defined between the fused representations and the student's corresponding features and logits. During training phase, the parameters of CLIP and the teacher model are frozen, and the student is trained using feature and logit distillation losses in addition to the cross-entropy loss. Inconsistencies in feature dimensions are addressed through a linear layer transformation.

and TeKAP [16] introduced an augmentation-based technique that generates multiple synthetic teacher representations by perturbing the knowledge of a single pretrained teacher.

Despite the advancements in multi-teacher distillation, existing methods still lack knowledge diversity, as they rely solely on unimodal visual information. Although works such as CLIP-KD [56] have explored distilling CLIP itself into smaller CLIP model, the use of vision–language models like CLIP as auxiliary teachers in traditional visual knowledge distillation remains largely underexplored. CIKD [62] was the first to investigate the use of CLIP for knowledge distillation, combining the text features from CLIP with the intermediate layer features of the student to generate new logit outputs.

Our work builds upon these lines of research by introducing a method to inject cross-modal diversity into the distillation process. Instead of training multiple task-specific teachers [19] or relying solely on perturbation-based approaches [16], we fuse the logits/features of conventional teacher with a frozen CLIP model to construct an implicit multi-teacher approach that integrates both dataset-specialized and language-grounded knowledge. Furthermore, unlike CIKD [62], which introduces additional loss terms based on CLIP's logits, the proposed method jointly utilizes both the logits and features of CLIP by fusing them with those of the original teacher.

3. Proposed Method

In this section, the proposed **RichKD** is presented, a framework that unifies a task-specific teacher T and the general-purpose CLIP teacher C into a single distillation process. RichKD leverages the complementary strengths of T's dataset-specific discrimination and C's broad semantic knowledge by fusing their logits and features through weighted combinations. Figure 2 shows the overall diagram of the proposed method. This fusion, inspired by ensemble learning theory, reduces individual biases and enhances the generalization of the distilled student.

3.1. Multi-Prompt CLIP Logits

Given an input image x, the conventional teacher produces a logit vector $z_T(x) \in \mathbb{R}^K$ for K classes. The CLIP model consists of an image encoder $C_{\text{img}}(\cdot)$ and a text encoder $C_{\text{text}}(\cdot)$. For each class $c \in \{1, \ldots, K\}$ and prompt p_m from a set of M textual templates $\mathcal{P} = \{p_1, \ldots, p_M\}$, the text description is formed

$$t_m(c) = p_m(\text{“class } c\text{”}), \tag{1}$$

and compute the text embedding $h_m(c) = C_{\text{text}}(t_m(c))$. The CLIP logit for class c under prompt p_m is

$$z_C^{(m)}(x)_c = \tau \cdot \frac{\langle C_{\text{img}}(x),\, h_m(c)\rangle}{\|C_{\text{img}}(x)\|_2 \, \|h_m(c)\|_2}, \tag{2}$$

where τ is a learned temperature scaling factor, $\langle \cdot, \cdot \rangle$ denotes the standard inner (dot) product, and $\| \cdot \|_2$ denotes the L2 norm. The final CLIP logit is the mean over prompts:

$$z_C(x) = \frac{1}{M} \sum_{m=1}^{M} z_C^{(m)}(x). \tag{3}$$

This averaging introduces diversity via multiple textual contexts, reducing the variance of individual prompt biases and yielding a smoother, more general prediction distribution.

3.2. Logit Fusion: Perturbed Supervision

The teacher and CLIP logits are fused as

$$z_F(x) = \alpha \, z_T(x) + (1 - \alpha) \, z_C(x), \tag{4}$$

where $\alpha \in [0, 1]$ controls the strength of the conventional teacher relative to CLIP. The student logits $z_S(x)$ are then trained to match the fused logits $z_F(x)$ through a softened Kullback–Leibler (KL) divergence loss:

$$\mathcal{L}_{\text{logit}} = \text{KL}\left(\sigma\left(\frac{z_F(x)}{T_{\text{temp}}} \right) \middle\| \, \sigma\left(\frac{z_S(x)}{T_{\text{temp}}} \right) \right), \tag{5}$$

where $\sigma(\cdot)$ denotes the softmax and T_{temp} is the temperature for smoothing.

This simple linear combination can be viewed as a stochastic perturbation of the teacher logits. Because $z_C(x)$ comes from a model trained on diverse open-domain data, it behaves as a low-frequency, semantically meaningful perturbation of $z_T(x)$:

$$\begin{aligned} z_F(x) &= z_T(x) + \epsilon(x), \\ \text{where} \quad \epsilon(x) &= (1 - \alpha) \, [z_C(x) - z_T(x)]. \end{aligned} \tag{6}$$

In general, the two teachers may have different biases with respect to the true target distribution, so $\mathbb{E}[\epsilon(x)]$ need not be zero. However, as long as the teachers are not perfectly aligned, the perturbation $\epsilon(x)$ has non-zero variance and encodes complementary information from CLIP. This variability acts as a semantically meaningful perturbation of $z_T(x)$, providing a richer and more regularized supervision signal for the student.

3.3. Feature Fusion

Beyond the output logits, intermediate representations from both teachers are also transferred. Let $f_T(x)$ and $f_C(x)$ denote the feature maps from the last hidden layer of the supervised teacher and CLIP image encoder, respectively. The fused feature is defined as

$$f_F(x) = \lambda \, f_T(x) + (1 - \lambda) \, f_C(x), \tag{7}$$

where $\lambda \in [0, 1]$ balances task-specific and general representations. The student feature $f_S(x)$ is trained to approximate $f_F(x)$ through a general feature distillation loss:

$$\mathcal{L}_{\text{feat}} = \mathcal{D}(f_S(x), f_F(x)) \,, \tag{8}$$

where $\mathcal{D}(\cdot, \cdot)$ denotes a generic similarity or alignment loss (e.g., cosine distance, attention transfer, or contrastive loss).

3.4. Overall Training Objective

The complete loss combines ground-truth supervision with fused logit and feature distillation:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(y, \sigma(z_S)) + \beta \, \mathcal{L}_{\text{logit}} + \gamma \, \mathcal{L}_{\text{feat}}, \tag{9}$$

where β and γ weight the respective distillation terms.

3.5. Theoretical Motivation

From a theoretical standpoint, fusing z_T and z_C can be interpreted as constructing an ensemble teacher $\mathcal{T}_F$ whose predictions are a convex combination of two heterogeneous teachers. This ensemble can be analyzed using a bias-variance view. Let Bias_T and Bias_C denote the biases of the task-specific teacher and CLIP teacher with respect to the true target, and let Var_T and Var_C be their prediction variances. The fused teacher has bias

$$\text{Bias}_F = \alpha \, \text{Bias}_T + (1 - \alpha) \, \text{Bias}_C. \tag{10}$$

and variance

$$\begin{aligned} \text{Var}[z_F] = \alpha^2 \text{Var}_T + (1 - \alpha)^2 \text{Var}_C \\ + 2\alpha(1 - \alpha) \, \text{Cov}[z_T, z_C]. \end{aligned} \tag{11}$$

The corresponding ensemble error can be written as

$$\mathcal{E}_F = \big(\text{Bias}_F\big)^2 + \text{Var}[z_F] + \sigma^2, \tag{12}$$

where σ^2 denotes irreducible noise. When the two teachers have *complementary* biases (e.g., a task-specific teacher that may overfit and a CLIP teacher that generalizes more broadly) and their prediction errors are not highly correlated, one can choose $\alpha \in (0, 1)$ such that $\mathcal{E}_F \leq \min\{\mathcal{E}_T, \mathcal{E}_C\}$. In the idealized case of two approximately unbiased teachers with similar variance and weakly correlated errors, the ensemble error $\mathcal{E}_F$ becomes strictly smaller than that of each individual teacher. This provides theoretical support for using the fused teacher $\mathcal{T}_F$ as a more informative and regularized supervision signal for the student compared to relying on a single teacher.

Table 1. Top-1 accuracy (%) of different knowledge distillation methods on the validation set of CIFAR-100. The teacher and student share the same architecture but differ in configuration. We apply our method to both logit-based and feature-based distillation approaches, and use Δ to indicate the performance gain. Values in blue represent minor improvements, while those in red indicate significant improvements of at least 0.15. The best and second-best results are indicated by **bold** and underlining font, respectively. "L" and "F" denote logit-based and feature-based methods, respectively.

Type	**ResNet32×4** **ResNet8×4**	**VGG13** **VGG8**	**WRN-40-2** **WRN-40-1**	**WRN-40-2** **WRN-16-2**	**ResNet56** **ResNet20**	**ResNet110** **ResNet32**	**ResNet110** **ResNet20**
Teacher	79.42	74.64	75.61	75.61	72.34	74.31	74.31
Student	72.50	70.36	71.98	73.26	69.06	71.14	69.06
FitNet [39]	73.50	71.02	72.24	73.58	69.21	71.06	68.99
AT [58]	73.44	71.43	72.77	74.08	70.55	72.31	70.65
RKD [37]	71.90	71.48	72.22	73.35	69.61	71.82	69.25
OFD [14]	74.95	73.95	74.33	75.24	70.98	73.23	71.29
SimKD [4]	**78.08**	**74.89**	74.53	75.53	71.05	73.92	71.06
LSKD [45]	76.62	74.36	74.37	76.11	71.43	**74.17**	71.48
KD [15]	73.33	72.98	73.54	74.92	70.66	73.08	70.67
+ RichKD (L)	75.32	73.68	74.17	76.29	71.38	73.86	71.09
Δ	1.99	0.70	0.63	1.37	0.72	0.78	0.39
CRD [48]	75.51	73.94	74.14	75.48	71.16	73.48	71.46
+ RichKD (F)	76.09	74.22	74.70	75.63	71.64	73.99	71.77
Δ	0.58	0.28	0.56	0.15	0.48	0.51	0.31
RichKD (L+F)	76.72	74.86	**74.73**	**76.35**	**72.12**	74.08	**72.11**

4. Experiments

In this section a complete discussion of the experiments, including datasets, baselines, implementation details, and results is provided. In addition, further experiments, ablation studies, and visualizations, such as class imbalance evaluation, Grad-CAM analyses, training time complexity, prompt templates, and experiments with transformer-based architectures, are also presented in the supplementary material.

Datasets. Experiments are conducted on the **CIFAR-100** [21] and **ImageNet** [40] datasets. The **CIFAR-100** dataset consists of 50,000 training and 10,000 validation images across 100 classes, with each image having a resolution of 32×32 pixels. The **ImageNet** dataset is a large-scale benchmark for image classification, containing approximately 1.28 million training and 50,000 validation images from 1,000 categories. To evaluate the generalization capability of the proposed method, additional experiments are performed on the corrupted versions of **CIFAR-100** dataset [13],designed to assess model robustness under distributional shifts.

Baselines. The effect of the proposed method is evaluated across multiple logit-based and feature -based distillation approaches, including **KD** [15], **FitNet** [39], **AT** [58], **RKD** [37], **OFD** [14], **SimKD** [4], **LSKD** [45], **CRD** [48], **DKD** [61], **MLKD** [20], **NormKD** [6], and **RLD** [46]. Comparisons are also made with various multi-teacher KD methods, including **TAKD** [34], **CA-MKD** [59], **DGKD** [44], **CIKD** [62], and **TeKAP** [16].

Implementation Details. The same experimental settings as previous works [5, 20, 45, 61] are followed. For experiments on **CIFAR-100**, the optimizer is set to SGD, and the number of training epochs is 240, except for **MLKD**, which is trained for 480 epochs. The initial learning rate is set to 0.01 for **MobileNets** and **ShuffleNets**, and 0.05 for other architectures, including **ResNets**, **WRNs**, and **VGGs**. All reported results are averaged over four independent trials. More detailed experimental configurations are provided in the supplementary materials.

All hyperparameters were fine-tuned to select optimal values. We set α in Eq. (4) to 0.7, T_{temp} in Eq. (5) to 3, λ in Eq. (7) to 0.7, β in Eq. (9) to 3, and γ in Eq. (9) to 0.8.

4.1. Main Results

Results on CIFAR-100. The proposed method is compared with several prominent feature- and logit-based approaches using different teacher and student architectures. Table 1 shows the results of our method when the teacher and student share a similar architecture, while Table 2 presents the results for cases where the teacher and student have distinct

Table 2. Top-1 accuracy (%) of different knowledge distillation methods on the validation set of CIFAR-100. The teacher and student have different architectures. We apply our method to both logit-based and feature-based distillation approaches, and use Δ to indicate the performance gain. Values in blue represent minor improvements, while those in red indicate significant improvements of at least 0.15. The best and second-best results are indicated by underlining and **bold** font, respectively. "L" and "F" denote logit-based and feature-based methods, respectively.

Type	**ResNet32×4 SHN-V2**	**ResNet32×4 SHN-V1**	**ResNet32×4 WRN-16-2**	**ResNet32×4 WRN-40-2**	**WRN-40-2 ResNet8×4**	**VGG13 MN-V2**	**ResNet50 MN-V2**
Teacher	79.42	79.42	79.42	79.42	75.61	74.64	79.34
Student	71.82	70.50	73.26	75.61	72.50	64.60	64.60
FitNet [39]	73.54	73.59	74.70	77.69	74.61	64.16	63.16
AT [58]	72.73	71.73	73.91	77.43	74.11	59.40	58.58
RKD [37]	73.21	72.28	74.86	77.82	75.26	64.52	64.43
OFD [14]	76.82	75.96	76.17	79.25	74.36	69.48	69.04
SimKD [4]	**78.39**	**76.31**	**77.17**	**79.29**	75.29	69.44	69.97
LSKD [45]	75.56	-	75.26	77.92	**77.11**	68.61	69.02
KD [15]	74.45	74.07	74.90	77.70	73.97	67.37	67.35
+ RichKD (L)	75.68	75.04	75.39	77.82	75.77	69.03	68.07
Δ	1.23	0.97	0.49	0.12	1.80	1.66	0.77
CRD [48]	75.65	75.11	75.65	78.15	75.24	69.73	69.11
+ RichKD (F)	76.49	75.65	76.53	78.67	76.28	69.87	69.81
Δ	0.84	0.54	0.88	0.52	1.04	0.14	0.70
RichKD (L+F)	76.95	76.13	76.58	78.76	76.31	**70.03**	**70.15**

Table 3. Top-1 accuracy (%) comparison with existing multi-teacher distillation methods on the CIFAR-100 validation set.

Method	**ResNet32×4 ResNet8×4**	**WRN_40_2 WRN_40_1**
Teacher	79.42	75.61
Student	72.50	71.98
TAKD [34]	73.93	73.83
CA-MKD [59]	75.90	74.56
DGKD [44]	75.31	74.23
CIKD [62]	74.79	74.32
TeKAP [16]	75.98	74.41
RichKD	**76.72**	**74.73**

architectures. It can be seen that RichKD can be combined with KD for logit distillation and CRD for feature distillation, and that their combination consistently achieves comparable or better results to existing methods on different architectures.

As RichKD employs an auxiliary teacher model, it inherently follows a multi-teacher distillation framework. Table 3 presents the results of our method in comparison with existing multi-teacher approaches. It can be seen that our method outperforms popular and recently proposed multi-teacher methods across two different teacher–student architecture settings. It should be noted that although RichKD adopts a multi-teacher framework by incorporating CLIP in addition to the primary teacher, CLIP has not been trained specifically on the target dataset and is merely utilized to inject general knowledge into the teacher.

Furthermore, Table 4 presents the results of combining our method with recent logit-based distillation approaches, namely DKD, MLKD, NormKD, and RLD, across two different teacher–student architecture settings. It can be seen that RichKD can be seamlessly integrated into each of these methods, further improving their performance.

Results on ImageNet. To validate the scalability of the proposed method, the results on the large-scale ImageNet dataset are reported. Table 5 presents the results of RichKD in comparison with the KD baseline and the recent TeKAP method, where ResNet-34 and ResNet-18 are used as the teacher and student models, respectively.

4.2. Generalization Comparison

To validate the effectiveness of the proposed method, generalization ability of students trained using vanilla KD and RichKD are compared. Table 6 presents the results on adversarially perturbed data generated by two well-known attack methods, FGSM [8] and PGD [29], with different noise

Table 4. Impact of integrating the proposed method with recent logit-based distillation methods in terms of Top-1 accuracy (%) on the CIFAR-100 validation set.

Method	**ResNet32×4 ResNet8×4**	**WRN_40_2 WRN_40_1**
Teacher	79.42	75.61
Student	72.50	71.98
DKD [61]	76.32	74.81
+ RichKD (L)	76.78 (+0.46)	75.49 (+0.68)
MLKD [20]	77.08	75.35
+ RichKD (L)	77.41 (+0.33)	75.82 (+0.47)
NormKD [6]	76.57	74.84
+ RichKD (L)	76.73 (+0.16)	74.99 (+0.15)
RLD [46]	76.11	74.58
+ RichKD (L)	76.45 (+0.34)	74.84 (+0.26)

Table 5. Scalability comparison on ImageNet. Lower is better (error %).

Set	**Teacher**	**Student**	**KD**	**TeKAP**	**RichKD (L)**
Top-1	26.69	30.25	29.59	29.33	**29.10**
Top-5	8.58	10.93	10.30	10.08	**9.85**

Table 6. Adversarial robustness of the proposed method. Top-1/Top-5 accuracy are reported for FGSM and PGD attacks with different attack parameters.

FGSM Attack			
Method	$\epsilon = 0.001$	$\epsilon = 0.005$	$\epsilon = 0.010$
KD	23.51 / 45.68	21.91 / 43.86	19.99 / 41.89
RichKD (L)	25.75 / 49.21	23.68 / 47.44	21.70 / 45.30
PGD Attack			
Method	$\epsilon = 0.001$	$\epsilon = 0.005$	$\epsilon = 0.010$
KD	22.13 / 44.43	15.25 / 38.57	10.61 / 32.60
RichKD (L)	24.30 / 48.03	17.72 / 41.55	12.61 / 35.66

parameters. Both the teacher and the student are trained on the clean CIFAR-100 dataset, using ResNet-32×4 and ResNet-8×4 architectures, respectively. The results demonstrate that our method achieves higher top-1 and top-5 accuracy, benefiting from the inclusion of CLIP, which provides general knowledge not limited to the target dataset images.

Furthermore, Table 7 shows the reulst of vanilla KD and RichKD on several corrupted versions of the CIFAR-100. All the models are trained on clean CIFAR-100, and teacher and student architecters are ResNet-32×4 and ResNet-8×4, respectively. Similar to Table 6, it can be observed that the student trained with RichKD is more robust to different types of corruption. The proposed method consistently improves the top-1 and top-5 accuracy of the baseline method and achieves a significant performance margin over the baseline for certain corruption types.

Table 7. Robustness comparison (accuracy %) on corrupted versions of the CIFAR-100 dataset.

Corruptions	**KD**		**RichKD (L)**	
	Top-1	**Top-5**	**Top-1**	**Top-5**
gaussian noise	14.42	34.70	15.02	35.87
motion blur	48.45	49.34	49.34	75.96
snow	48.22	74.93	51.03	77.76
jpeg compression	45.34	73.16	46.99	75.18
spatter	51.16	78.00	55.04	81.43
average	41.51	62.02	43.48	69.24

4.3. Ablation Studies

To further validate the effectiveness of the proposed method, a series of ablation studies are conducted. Since RichKD employs CLIP as an auxiliary teacher to perturb the logits of the primary teacher, the choice and performance of the CLIP model play an important role in our framework. Table 8 presents the results of our method using different variants of CLIP, compared to the vanilla KD method, across three different teacher–student architecture pairs. It can be seen that ViT-L/14 improves the distillation performance more than the other CLIP variants, as it employs a more powerful image encoder and contains a larger number of parameters.

Moreover, Figure 3 illustrates the effect of different prompting strategies on the distillation results. We evaluate three types of prompts for CLIP's text encoder: (1) single prompting, where a single template containing the class name is used; (2) multi-prompting, where multiple templates are used with the class name; and (3) complex prompting, where, in addition to multiple prompt templates, several synonyms of each class name are incorporated to further enrich the text embeddings. As shown in the figure, both multi-prompting and complex prompting improve the zero-shot performance of CLIP, which subsequently leads to better distillation results. The detailed prompt formulations are presented in the supplementary material.

4.4. Qualitative Results

In addition to the quantitative comparisons, qualitative results are also presented. Figure 4 shows the feature representations projected into a 2D space using t-SNE. The final feature layer of each model is used, taken before the classification head. As can be observed, the embeddings of different classes are more clearly separated in the feature space for the student model trained with RichKD compared to the

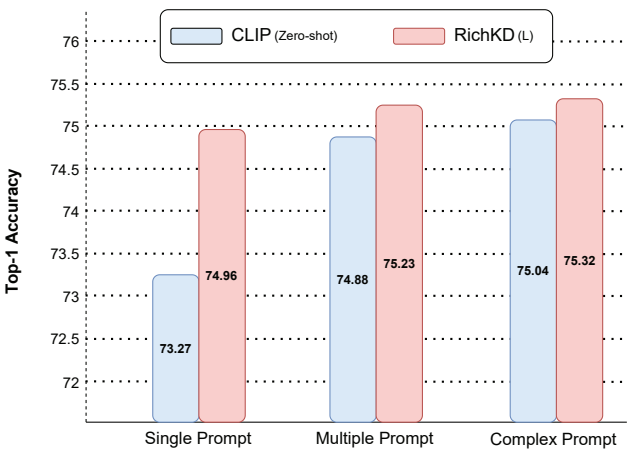

Figure 3. Impact of different types of prompting on CLIP's zero-shot performance and the student's performance. The teacher and student architectures are ResNet-32×4 and ResNet-8×4, respectively.

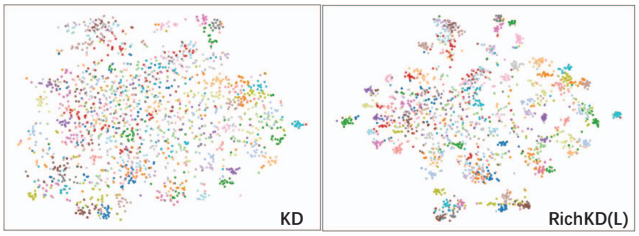

Figure 4. t-SNE visualization of features.

one trained without CLIP.

Table 8. Impact of different CLIP variants on distillation performance on the CIFAR-100 validation set; zero-shot accuracy for each CLIP model is shown in parentheses.

Method	CLIP Model	ResNet32×4 ResNet8×4	WRN-40-2 WRN-16-2	WRN-40-2 ResNet32×4
Teacher	—	79.42	75.61	75.61
Student	—	72.50	73.26	72.50
KD	—	73.33	74.92	73.97
RichKD (L)	RN101 (41.54)	74.89	75.03	75.34
	RN50x64 (52.15)	74.98	75.37	75.63
	ViT-B/32 (61.68)	74.75	75.79	75.49
	ViT-L/14 (73.27)	75.32	76.29	75.77

Furthermore, Figure 5 presents a comparison of inter-class correlations between vanilla KD and RichKD. It is evident that inter-class correlations are lower in RichKD, which can be attributed to the influence of CLIP's logits, which can alter the ranking of the top probabilities, thereby helping to reduce inter-class correlations and improve class separability. For both experiments, ResNet-32×4 and ResNet-8×4 were used as the teacher and student models, respectively.

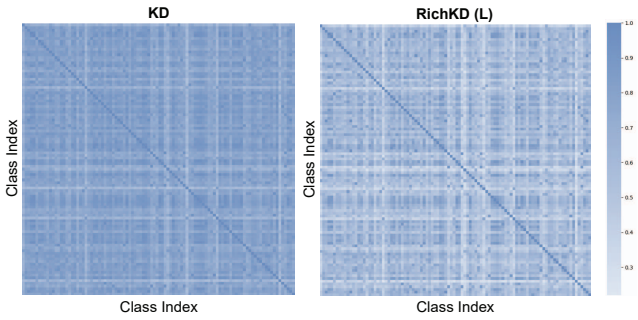

Figure 5. Inter-class correlation matrices on the CIFAR-100 dataset.

5. Discussion

Limitations. Using CLIP during training introduces a modest computational overhead due to additional forward passes. Although this occurs only during training, RichKD requires longer training time than standard KD. To reduce this cost, CLIP's logits and features are cached and reused across epochs, making the extra overhead negligible (see supplementary material for details). Moreover, while RichKD benefits from CLIP's broad visual–language knowledge, its effectiveness may diminish in domains underrepresented in CLIP's pretraining (e.g., medical imagery). In such cases, domain-specific CLIP variants can be seamlessly integrated into our framework.

Future Work. RichKD illustrates the promise of incorporating CLIP's cross-modal knowledge into knowledge distillation, yet several directions remain open. Our framework currently depends on hand-crafted prompts to query CLIP; integrating advanced techniques such as prompt tuning or dataset-specific textual descriptions could yield stronger supervision. Furthermore, since CLIP provides knowledge complementary to the in-domain teacher, developing adaptive strategies that selectively exploit CLIP's information per input sample is a promising avenue for future research toward more context-aware distillation.

6. Conclusion

In this work, we demonstrated that incorporating CLIP's cross-modal knowledge as an auxiliary teacher significantly enhanced conventional knowledge distillation. The proposed framework effectively combined dataset-specific and semantically enriched cues, leading to improved student performance, better robustness, and greater generalization. These findings underscored the potential of leveraging vision–language models to enrich traditional visual knowledge distillation approaches.

References

[1] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 2

[2] Qizhi Cao, Kaibing Zhang, Xin He, and Junge Shen. Be an excellent student: Review, preview, and correction. *IEEE Signal Processing Letters*, 30:1722–1726, 2023. 2

[3] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7028–7036, 2021. 2

[4] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11933–11942, 2022. 2, 5, 6

[5] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2021. 2, 5

[6] Zhihao Chi, Tu Zheng, Hengjia Li, Zheng Yang, Boxi Wu, Binbin Lin, and Deng Cai. Normkd: Normalized logits for knowledge distillation. *arXiv preprint arXiv:2308.00520*, 2023. 2, 5, 7

[7] Jiequan Cui, Zhuotao Tian, Zhisheng Zhong, Xiaojuan Qi, Bei Yu, and Hanwang Zhang. Decoupled kullback-leibler divergence loss. *Advances in Neural Information Processing Systems*, 37:74461–74486, 2024. 2

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 6

[9] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021. 1

[10] Jia Guo, Minghao Chen, Yao Hu, Chen Zhu, Xiaofei He, and Deng Cai. Reducing the teacher-student gap via spherical knowledge disitllation. *arXiv preprint arXiv:2010.07485*, 2020. 2

[11] Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin. Class attention transfer based knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11868–11877, 2023. 2

[12] Zhiwei Hao, Jianyuan Guo, Kai Han, Han Hu, Chang Xu, and Yunhe Wang. Revisit the power of vanilla knowledge distillation: from small scale to large scale. *Advances in Neural Information Processing Systems*, 36:10170–10183, 2023. 2

[13] Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. *arXiv preprint arXiv:1807.01697*, 2018. 5

[14] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1921–1930, 2019. 2, 5, 6

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 5, 6

[16] Md Imtiaz Hossain, Sharmen Akhter, Choong Seon Hong, and Eui-Nam Huh. Single teacher, multiple perspectives: Teacher knowledge augmentation for enhanced knowledge distillation. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 5, 6

[17] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35:33716–33727, 2022. 2

[18] Xiao Huang, Wu Chen, and Wei Zhou. Class-wise adaptive logits distillation with meta-learning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2

[19] Adrian Iordache, Bogdan Alexe, and Radu Tudor Ionescu. Multi-level feature distillation of joint teachers trained on distinct image datasets. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7133–7142. IEEE, 2025. 2, 3

[20] Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24276–24285, 2023. 2, 5, 7

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[22] Weichao Lan, Yiu-ming Cheung, Qing Xu, Buhua Liu, Zhikai Hu, Mengke Li, and Zhenghua Chen. Improve knowledge distillation via label revision and data selection. *IEEE Transactions on Cognitive and Developmental Systems*, 2025. 2

[23] Kehan Li, Runyi Yu, Zhennan Wang, Li Yuan, Guoli Song, and Jie Chen. Locality guidance for improving vision transformers on tiny datasets. In *European Conference on Computer Vision*, pages 110–127. Springer, 2022. 1, 2

[24] Lujun Li, Peijie Dong, Zimian Wei, and Ya Yang. Automated knowledge distillation via monte carlo tree search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17413–17424, 2023. 1, 2

[25] Tong Li, Long Liu, Kang Liu, Xin Wang, Bo Zhou, Hongguang Yang, and Kai Lu. Adaptive dual guidance knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18457–18465, 2025. 2

[26] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1504–1512, 2023. 2

[27] Li Liu, Qingle Huang, Sihao Lin, Hongwei Xie, Bing Wang, Xiaojun Chang, and Xiaodan Liang. Exploring inter-channel correlation for diversity-preserved knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8271–8280, 2021. 2

[28] Tao Liu, Chenshu Chen, Xi Yang, and Wenming Tan. Rethinking knowledge distillation with raw features for semantic segmentation. In *Proceedings of the IEEE/CVF Win-*

ter Conference on Applications of Computer Vision, pages 1155–1164, 2024. 2

[29] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 6

[30] Amir M Mansourian, Arya Jalali, Rozhan Ahmadi, and Shohreh Kasaei. Attention-guided feature distillation for semantic segmentation. *arXiv preprint arXiv:2403.05451*, 2024. 2

[31] Amir M Mansourian, Rozhan Ahamdi, and Shohreh Kasaei. Aicsd: Adaptive inter-class similarity distillation for semantic segmentation. *Multimedia Tools and Applications*, pages 1–20, 2025. 2

[32] Amir M Mansourian, Rozhan Ahmadi, Masoud Ghafouri, Amir Mohammad Babaei, Elaheh Badali Golezani, Zeynab Yasamani Ghamchi, Vida Ramezanian, Alireza Taherian, Kimia Dinashi, Amirali Miri, et al. A comprehensive survey on knowledge distillation. *arXiv preprint arXiv:2503.12067*, 2025. 1

[33] Roy Miles and Krystian Mikolajczyk. Understanding the role of the projector in knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4233–4241, 2024. 2

[34] 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*, pages 5191–5198, 2020. 2, 5, 6

[35] Jinhyuk Park and Albert No. Prune your model before distill it. In *European Conference on Computer Vision*, pages 120–136. Springer, 2022. 2

[36] Sangyong Park and Yong Seok Heo. Knowledge distillation for semantic segmentation using channel and spatial correlations and adaptive cross entropy. *Sensors*, 20(16):4616, 2020. 2

[37] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019. 2, 5, 6

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[39] 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. 2, 5, 6

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5

[41] Ioannis Sarridis, Christos Koutlis, Giorgos Kordopatis-Zilos, Ioannis Kompatsiaris, and Symeon Papadopoulos. Inditill: Information flow-preserving knowledge distillation for model compression. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9033–9042. IEEE, 2025. 2

[42] Bharat Bhusan Sau and Vineeth N Balasubramanian. Deep model compression: Distilling knowledge from noisy teachers. *arXiv preprint arXiv:1610.09650*, 2016. 2

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[44] Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. Densely guided knowledge distillation using multiple teacher assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9395–9404, 2021. 2, 5, 6

[45] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15731–15740, 2024. 2, 5, 6

[46] Wujie Sun, Defang Chen, Siwei Lyu, Genlang Chen, Chun Chen, and Can Wang. Knowledge distillation with refined logits. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1110–1119, 2025. 2, 5, 7

[47] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4

[48] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019. 2, 5, 6

[49] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1365–1374, 2019. 2

[50] Lu Wang, Liuchi Xu, Xiong Yang, Zhenhua Huang, and Jun Cheng. Debiased distillation for consistency regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7799–7807, 2025. 2

[51] Yukang Wang, Wei Zhou, Tao Jiang, Xiang Bai, and Yongchao Xu. Intra-class feature variation distillation for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 346–362. Springer, 2020. 2

[52] Shicai Wei, Chunbo Luo, and Yang Luo. Scaled decoupled distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15975–15983, 2024. 2

[53] Tiancheng Wen, Shenqi Lai, and Xueming Qian. Preparing lessons: Improve knowledge distillation with better supervision. *Neurocomputing*, 454:25–33, 2021. 2

[54] Xiaomeng Xin, Heping Song, and Jianping Gou. A new similarity-based relational knowledge distillation method.

In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3535–3539. IEEE, 2024. 2

[55] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12319–12328, 2022. 2

[56] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15952–15962, 2024. 3

[57] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1285–1294, 2017. 2

[58] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 2, 5, 6

[59] Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4498–4502. IEEE, 2022. 2, 5, 6

[60] Hailin Zhang, Defang Chen, and Can Wang. Adaptive multi-teacher knowledge distillation with meta-learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1943–1948. IEEE, 2023. 2

[61] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022. 2, 5, 7

[62] Jingtao Zhou, Hao Zheng, Wenkai Zhong, and Zhiqiang Bao. Improving knowledge distillation via cross-modal insights from clip. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2, 3, 5, 6

[63] Zaida Zhou, Chaoran Zhuge, Xinwei Guan, and Wen Liu. Channel distillation: Channel-wise attention for knowledge distillation. *arXiv preprint arXiv:2006.01683*, 2020. 2