Rethinking the Dark Knowledge and Kullback-Leibler Divergence Loss in Knowledge Distillation Under Capacity Mismatching

Yingchao Wang*, **Wenqi Niu, Xingshan Yao, Li You, Weilun Fei**

School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing, 100081, China
{yingchaowang; 3220255294; 3120245900; youli; wlfei}@bit.edu.cn

Abstract

Knowledge Distillation (KD) aims to transfer the dark knowledge that encodes inter-class similarity, semantic structure, and decision boundaries from a powerful teacher model to a compact student model by minimizing the Kullback-Leibler (KL) divergence between their output distributions. While effective, we demonstrate that KL-based KD is designed to match values precisely and does not explicitly constrain the relative relationships between classes. Meanwhile, we empirically find that vanilla KL-based KD suffers from gradient competition due to the zero-sum constraint in the softmax space, which may implicitly change the inter-class rank relationships learned by the student model, particularly under capacity mismatching. Therefore, we argue that the student model should learn not only the output values but also the relative ranking of classes. Accordingly, we propose a simple yet effective Relative Confidence Knowledge Distillation (RCKD) method that aligns the teacher's and student's relative confidence matrices via cosine similarity, achieving more efficient and robust distillation from a stronger teacher model. Extensive experiments demonstrate that RCKD consistently outperforms existing logit-based KD methods and exhibits strong adaptability across various teacher architectures and capacities.

Code — https://github.com/yingchao-wang/RCKD

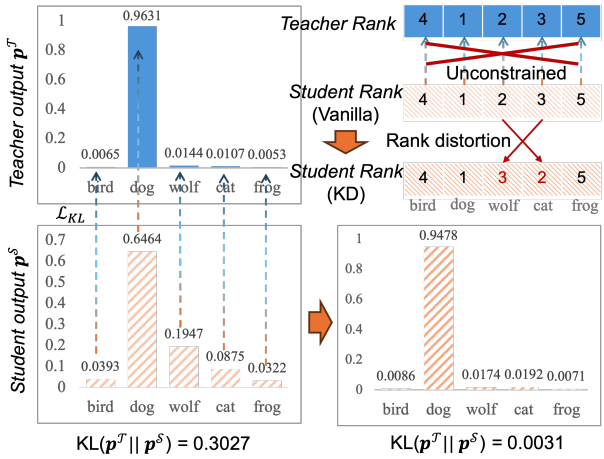

Figure 1: Vanilla KD enforces numerical consistency between the teacher and student output probabilities but does not explicitly constrain the preservation of the teacher's predicted class ranking. Moreover, due to the gradient competition inherent in the KL divergence, it implicitly leads the student model to learn a distorted rank structure.

Introduction

Knowledge Distillation (KD) (Hinton, Vinyals, and Dean 2015) holds the potential to transfer the dark knowledge acquired by a higher-capacity teacher model to a compact student model, achieving efficient model compression while maintaining commendable performance (Yang et al. 2022; Wang et al. 2024). Typically, this is achieved by minimizing the Kullback-Leibler (KL) divergence between the teacher's and student's softened output distributions. Intuitively, using a larger and stronger teacher model is expected to distill into a better-performing student model. However, previous studies (Cho and Hariharan 2019; Park et al. 2019; Mirzadeh et al. 2020; Son et al. 2021; Zhu and Wang 2021; Wang et al. 2022; Zhu et al. 2022; Huang et al. 2022; Rao et al. 2023; Liang et al. 2024; Yuan, Lang, and Quan 2024) have shown that this empiricism does not always hold. The student model distilled from a higher accuracy and larger-scale teacher model may perform worse due to the capacity mismatch. To address this issue, some studies (Zhu and Wang 2021; Mirzadeh et al. 2020; Son et al. 2021) focused on the innovation of KD architecture. For example, TAKD (Mirzadeh et al. 2020) was proposed to reduce the discrepancy between teacher and student by resorting to an additional teaching assistant of moderate model size. While some studies aimed to regularize the teacher's knowledge. For example, Wang et al. (2022) advocated that an intermediate checkpoint is more appropriate for distillation.

Different from the above studies, we revisit the fundamental reasons behind the poor performance of traditional KD under capacity mismatch, analyzing it from the perspectives of logit-level dark knowledge and the intrinsic properties of the KL divergence distillation loss. Specifically, we argue that the higher-order semantics of the teacher's output are presented in two complementary forms of knowledge:

*Corresponding author.

numerical knowledge, reflected in the absolute confidence values, and directional or rank knowledge, which encodes the relative ordering among class predictions. However, as shown in Figure 1, (*i*) we observe that KL-based distillation remains a pointwise alignment strategy that emphasizes individual class probability differences, while neglecting the underlying rank relations among classes. (*ii*) From the gradient perspective, we demonstrate that the KL loss induces a zero-sum constraint on student logits due to the normalization of the softmax. As a result, increasing the confidence in one class inherently suppresses others, leading to entangled and competitive gradient updates that can destabilize learning and impair the preservation of directional knowledge. For example, as shown in Figure 1, although the KL divergence value between teacher and student outputs decreases, the student model has learned ordinal relationships inconsistent with those of the teacher. This distortion potentially undermines the decision boundaries of the student model, thereby degrading its overall performance.

Therefore, we argue that the student model should be guided to learn not only the numerical knowledge encoded in the teacher's output, but also its directional (rank) knowledge. Importantly, when there exists a significant capacity gap between the teacher and the student, preserving directional knowledge becomes even more essential, as the student often lacks the capacity to reconstruct subtle relational structures solely through numerical imitation. Regarding this, we propose a simple yet effective Relative Confidence Knowledge Distillation (RCKD) method that directly constructs the distillation loss based on the Relative Confidence Matrix (RCM), which captures the pairwise numerical and directional knowledge between class probabilities.

The main contributions can be summarized as follows:

- We formally establish that gradient competition inherent in standard KL-divergence distillation loss provably distorts class ordinal relationships in student models under non-negligible teacher-student capacity gaps.
- We construct the relative confidence matrix that captures the pairwise log-ratio between class probabilities, thereby jointly characterizing the output probabilities' numerical and directional information.
- We propose RCKD, a novel relative confidence distillation method that employs cosine similarity to transfer dark knowledge represented as a relative confidence matrix from the teacher model.

Related Work

Logit-based Knowledge Distillation

Traditional logit-based KD methods aim to minimize the KL divergence between the teacher's and student's softened probability distributions (Hinton, Vinyals, and Dean 2015). Extensions such as temperature-based KD (Li et al. 2023; Sun et al. 2024) attempt to stabilize or calibrate the transferred probabilities. For example, Sun et al. (2024) introduced a Z-score pre-process of logit standardization before applying softmax and KL divergence loss. However, these approaches still implicitly mandate an exact match between teacher and student output. Conversely, several studies have argued that preserving the intrinsic rank relationships among predictions is more beneficial than strictly matching the teacher's output values. Huang et al.(2022) introduced a Pearson correlation-based loss to explicitly capture the inter-class dependencies encoded in the teacher's output. Similarly, Yin et al.(2024) replaced the conventional KL divergence with a loss function based on the Spearman rank coefficient, while Guan et al.(2025) proposed a plug-and-play loss utilizing the Kendall rank coefficient. Moreover, Fan et al.(2024) observed a strong positive correlation between the calibration level of the teacher model and the effectiveness of standard KD methods, and advocated for the use of calibration-insensitive alternatives, such as ranking-based losses (Huang et al. 2022). However, these studies have not provided an in-depth analysis of how KL divergence negatively affects rank relationship learning, particularly from the perspective of gradient optimization. Moreover, purely matching the rank relationships overlooks the informative value embedded in the magnitude of the teacher's output.

Capacity Mismatch in Knowledge Distillation

A well-recognized challenge in KD is the performance degradation that arises when there exists a large capacity gap between the teacher and student models. To bridge this gap, prior works have explored two main directions: architectural modification and teacher regularization. In the first direction, assistant-based frameworks such as TAKD (Mirzadeh et al. 2020), DGKD (Son et al. 2021), and NSKD (Liang et al. 2024) introduce intermediate models or auxiliary classifiers to ease the knowledge transition. Other approaches like SCKD (Zhu and Wang 2021) adaptively modulate the distillation process based on gradient similarity. However, these methods often rely on heuristic designs or manual intervention to select optimal intermediate models or tuning points. The second line of work focuses on regularizing the teacher's output to alleviate the overburdened student. Strategies include early stopping (Cho and Hariharan 2019), intermediate checkpointing (Wang et al. 2022), selective distillation (Zhu et al. 2022), and smoothness tuning via adapter modules (Rao et al. 2023). Additionally, SKD (Yuan, Lang, and Quan 2024) reformulates the teacher's outputs into simplified knowledge representations to facilitate learning. While these methods have achieved encouraging results, they primarily operate at the architectural or output calibration level. In contrast, our work revisits the capacity mismatch problem from the distillation loss perspective and reveals that the standard KL divergence can induce structural distortion of rank information due to its gradient competition, particularly under significant capacity gaps.

Motivation

KL Divergence Focuses on Numerical Alignment

In vanilla KD, the student model $\mathcal{S}$ is trained to align its predictive distribution with that of a pre-trained teacher model $\mathcal{T}$ by minimizing the KL divergence between their temperature (τ)-softened output probabilities. Specifically,

given a task with C classes, and the teacher's and student's softened probabilities $\boldsymbol{p}^{\mathcal{T}} = \mathrm{softmax}(\boldsymbol{z}^{\mathcal{T}}/\tau)$ and $\boldsymbol{p}^{\mathcal{S}} = \mathrm{softmax}(\boldsymbol{z}^{\mathcal{S}}/\tau)$, the distillation loss is formulated as:

$$\mathcal{L}_{\mathrm{KD}} = \mathrm{KL}\left(\boldsymbol{p}^{\mathcal{T}} \,\|\, \boldsymbol{p}^{\mathcal{S}}\right) = \sum_{c=1}^{C} p_c^{\mathcal{T}} \log \frac{p_c^{\mathcal{T}}}{p_c^{\mathcal{S}}}. \tag{1}$$

Differentiating the KL loss with respect to the student logits $z_k^{\mathcal{S}}$, we obtain:

$$\frac{\partial \mathcal{L}_{\mathrm{KD}}}{\partial z_k^{\mathcal{S}}} = \frac{1}{\tau} \cdot \left(p_k^{\mathcal{S}} - p_k^{\mathcal{T}}\right). \tag{2}$$

This shows that each gradient update is based solely on the pointwise discrepancy between the predicted probabilities of the student and the teacher for the same class. There is no explicit modeling of inter-class dependencies, such as relative ordering or confidence margins between class pairs.

To better understand this limitation, we further consider the distillation loss from the perspective of relative confidence structures. Let us define the pairwise relative confidence between classes j and k as:

$$\Delta_{j,k} = \log\left(\frac{p_j}{p_k}\right) = \frac{z_j - z_k}{\tau}. \tag{3}$$

This log-ratio encodes directional preferences between categories through its positive and negative values, capturing rank-relevant information beyond absolute probabilities. From this, we can decompose the KL loss into two components from the perspective of gradient optimization (Cui et al. 2024), as presented in Theorem 1.

Theorem 1 *From the perspective of gradient optimization, the KL Divergence distillation loss is equivalent to the following loss:*

$$\begin{aligned}\mathcal{L}_{\mathrm{KD}} = &\underbrace{\frac{\tau^2}{4}\sum_{j=1}^{C}\sum_{k=1}^{C}\left(\mathrm{S}(p_j^{\mathcal{T}} p_k^{\mathcal{T}}) \cdot (\Delta_{j,k}^{\mathcal{T}} - \mathrm{S}(\Delta_{j,k}^{\mathcal{S}}))^2\right)}_{\textit{Structure-preserving (weighted MSE)}} \\ &\underbrace{-\sum_{j=1}^{C}\mathrm{S}(p_j^{\mathcal{T}})\log p_j^{\mathcal{S}}}_{\textit{Cross-entropy}},\end{aligned} \tag{4}$$

where $\mathrm{S}(\cdot)$ *represents stop gradients operation.*

Interpretation. As shown in Theorem 1, the first term encourages the student to recover the teacher's pairwise relative confidence structure through a weighted Mean Squared Error (MSE), but it is disconnected from gradient flow due to the stop-gradient operator $\mathrm{S}(\cdot)$. Consequently, the actual optimization is dominated by the second term, which solely focuses on individual class-level probability matching. This decomposition confirms that although the teacher's internal relational knowledge (i.e., class rankings and margins) is implicitly embedded in its output distribution, the KL loss function does not directly encourage the student to recover such structure. Instead, it acts as a strict numerical matching objective, targeting exact alignment of probability values without regard for ordinal semantics.

Gradient Competition Dilemma

Since both $\boldsymbol{p}^{\mathcal{T}}$ and $\boldsymbol{p}^{\mathcal{S}}$ are valid probability distributions that sum to 1, we have:

$$\sum_{c=1}^{C}\frac{\partial \mathcal{L}_{\mathrm{KD}}}{\partial z_c^{\mathcal{S}}} = \frac{1}{\tau}\cdot\left(\sum_{c=1}^{C} p_c^{\mathcal{S}} - \sum_{c=1}^{C} p_c^{\mathcal{T}}\right) = 0. \tag{5}$$

This result reveals that the gradients of the KL loss over all logits must sum to zero, which imposes a zero-sum constraint. As a consequence, improving the prediction for one class must come at the cost of suppressing others, even if those classes are semantically unrelated. This leads to competitive gradient dynamics between classes. To further analyze the optimization landscape, we examine the second-order structure of the KL loss by computing the Hessian matrix with respect to the student logits:

$$\mathbf{H}_{\mathrm{KL}} = \nabla^2_{\boldsymbol{z}^{\mathcal{S}}}\mathcal{L}_{\mathrm{KD}} = \frac{1}{\tau^2}\left(\mathrm{diag}(\boldsymbol{p}^{\mathcal{S}}) - \boldsymbol{p}^{\mathcal{S}}(\boldsymbol{p}^{\mathcal{S}})^{\top}\right). \tag{6}$$

This Hessian has the following properties:

$$\frac{\partial^2 \mathcal{L}_{\mathrm{KD}}}{\partial z_j^{\mathcal{S}}\partial z_k^{\mathcal{S}}} = \begin{cases} \frac{1}{\tau^2}p_k^{\mathcal{S}}(1-p_k^{\mathcal{S}}), & \text{if } j = k \\ -\frac{1}{\tau^2}p_j^{\mathcal{S}}p_k^{\mathcal{S}}, & \text{if } j \neq k \end{cases}, \tag{7}$$

where all off-diagonal elements are strictly negative, indicating that any increase in one logit decreases the gradient w.r.t. others. This global coupling becomes particularly problematic when the student model has limited capacity and cannot precisely match the teacher's output. In such a case, the student model prioritizes matching classes exhibiting significant discrepancies between the student and teacher output probabilities. However, due to the gradient competition effect, this prioritized alignment inadvertently disturbs the optimization of classes whose probabilities are already close to the teacher's. As a result, the relative confidence ordering among these classes may become distorted, undermining the integrity of inter-class relationships.

To empirically verify the above analysis, we conduct Top-K distillation experiments on the CIFAR-100 dataset. As shown in Figure 2, the performance of the student model exhibits an inverted U-shaped trend as the value of Top-K increases, initially improving and then declining. A similar pattern is observed in the Spearman rank correlation between the student and teacher outputs. Specifically, when Top-K exceeds a certain threshold, the Spearman correlation begins to decrease, suggesting that inter-class interference and competition become increasingly prominent, which distorts the rank structure of the student's output.

In summary, KL divergence cannot inherently decouple numerical alignment from directional consistency, and may misguide the student to sacrifice rank semantics in favor of value matching, motivating the design of alternative objectives that enable more structure-aware and rank-consistent knowledge transfer.

Proposed Method

Relative Confidence Knowledge Distillation

As shown in Figure 3, RCKD explicitly models the structural information embedded in the teacher's output by constructing a relative confidence matrix (RCM), and transfers it to the student model via a cosine similarity loss.

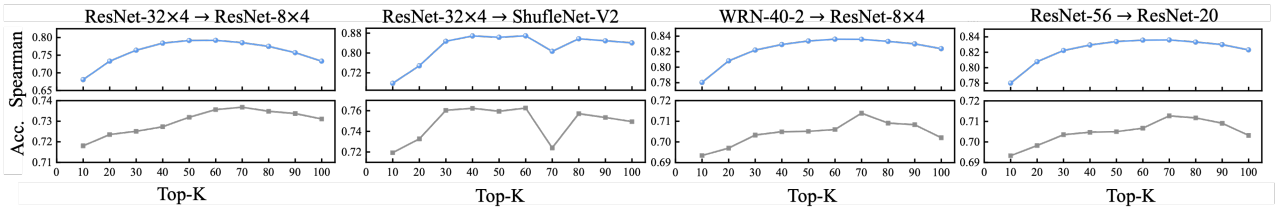

Figure 2: Top-1 accuracy and Spearman correlation between teacher and student under Top-K KD.

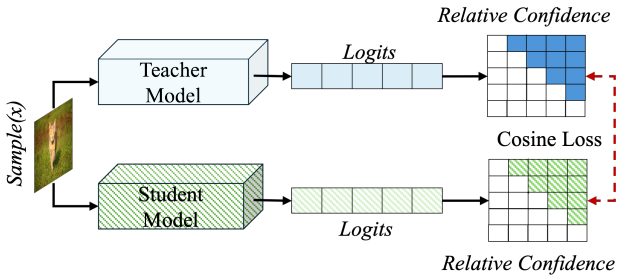

Figure 3: Framework of our proposed method RCKD.

Relative Confidence Matrix The RCM is defined as $\mathbf{\Delta} = [\Delta_{j,k}] \in \mathbb{R}^{C\times C}$, where each element $\Delta_{j,k}$ captures the relative confidence of class j over class k, encoding both the numerical magnitude and the directional preference in the model's prediction. In practice, the RCM can be efficiently constructed through a single broadcasted subtraction operation between expanded logit tensors. This operation is inherently parallelizable on GPUs and incurs negligible computational overhead, even as the number of classes C increases. However, the inherent symmetry of the RCM (i.e., $\Delta_{j,k} = -\Delta_{k,j}$, and the diagonal elements $\Delta_{j,j} = 0$.) induces parametric redundancy. Directly using the full RCM as the distillation knowledge inevitably introduces redundant storage and duplicated loss penalties, since each anti-symmetric pair is counted twice in the objective. Thus, we retain only the upper-triangular part of $\mathbf{\Delta}$ to obtain a minimal and non-redundant representation. Let $\mathcal{P} = \{(j,k) \mid 1 \le j < k \le C\}$ denote the set of all unordered class pairs and the total number of pairs is $|\mathcal{P}| = \binom{C}{2}$. For the teacher and student model, we flatten the upper-triangular entries into a vector: $\mathbf{v}^{\mathcal{T}} = [\Delta^{\mathcal{T}}_{j,k} \mid (j,k) \in \mathcal{P}] \in \mathbb{R}^{|\mathcal{P}|}$ and $\mathbf{v}^{\mathcal{S}} = [\Delta^{\mathcal{S}}_{j,k} \mid (j,k) \in \mathcal{P}] \in \mathbb{R}^{|\mathcal{P}|}$.

Cosine Similarity Distillation Loss Regarding the selection of the loss function, as shown in Theorem 1, the structure-preserving term aims to match the teacher model's RCM using a weighted MSE loss. However, this design suffers from inherent limitation: MSE is highly sensitive to the numerical scale. Given the prevalent scale discrepancy between teacher and student logits under capacity mismatching, the MSE loss is dominated by this scale difference. This fundamentally compels the student model to prioritize amplifying the global magnitude of its logits rather than precisely capturing the teacher's relative confidence structure. Therefore, we abandoned the weighted MSE matching in Theorem 1 and instead adopted the relaxed direction matching, that is, we constructed the loss using cosine similarity, ensuring the student focuses exclusively on replicating the relational structure encoded in $\mathbf{v}^{\mathcal{T}}$ rather than matching absolute scales:

$$\mathcal{L}_{\text{RCKD}} = 1 - \text{sim}\left(\mathbf{v}^{\mathcal{T}}, \mathbf{v}^{\mathcal{S}}\right) = 1 - \frac{\mathbf{v}^{\mathcal{T}} \cdot \mathbf{v}^{\mathcal{S}}}{\|\mathbf{v}^{\mathcal{T}}\| \cdot \|\mathbf{v}^{\mathcal{S}}\|}. \quad (8)$$

The overall training loss for the student model is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \beta \mathcal{L}_{\text{RCKD}}. \quad (9)$$

where $\mathcal{L}_{\text{CE}}$ denotes the standard cross-entropy loss.

Comparison with KL-based Distillation

Explicit Ordinal Structure Modeling Unlike KL divergence that implicitly encodes pairwise relationships, RCKD explicitly optimizes ordinal structures through the relative confidence vector $\mathbf{v}$. The cosine loss function $\mathcal{L}_{\text{RCKD}} = 1 - \cos(\mathbf{v}^{\mathcal{T}}, \mathbf{v}^{\mathcal{S}})$ encourages the student model to replicate the teacher's relative confidence direction. The gradient w.r.t. $\Delta^{\mathcal{S}}_{j,k}$ reveals:

$$\frac{\partial \mathcal{L}_{\text{RCKD}}}{\partial \Delta^{\mathcal{S}}_{j,k}} = \frac{1}{\|\mathbf{v}^{\mathcal{T}}\| \cdot \|\mathbf{v}^{\mathcal{S}}\|} \left(-\Delta^{\mathcal{T}}_{j,k} + \cos\theta \cdot \frac{\|\mathbf{v}^{\mathcal{T}}\|}{\|\mathbf{v}^{\mathcal{S}}\|} \Delta^{\mathcal{S}}_{j,k} \right), \quad (10)$$

where θ is the angle between vectors. This result shows that RCKD aligns the student logits to the teacher's ordinal structure in direction (via the $-\Delta^{\mathcal{T}}_{j,k}$ term), while automatically adjusting the scale based on their cosine similarity and norm ratio. Specifically, when $\cos\theta \approx 0$, the gradient is dominated by the alignment term $-\Delta^{\mathcal{T}}_{j,k}$, enforcing structural similarity. When $\cos\theta \to 1$, the gradient naturally encourages the student to match the teacher's magnitude up to a scaling factor $\|\mathbf{v}^{\mathcal{T}}\|/\|\mathbf{v}^{\mathcal{S}}\|$. The gradient direction always lies in the plane defined by $\mathbf{v}^{\mathcal{T}}$ and $\mathbf{v}^{\mathcal{S}}$, decoupling angular structure from global scale.

Absence of Gradient Competition In RCKD, the gradient w.r.t. student logit $z^{\mathcal{S}}_c$ is:

$$\frac{\partial \mathcal{L}_{\text{RCKD}}}{\partial z^{\mathcal{S}}_i} = \frac{1}{\tau} \left[\sum_{k:(i,k)\in\mathcal{P}} \frac{\partial \mathcal{L}}{\partial \Delta^{\mathcal{S}}_{i,k}} - \sum_{j:(j,i)\in\mathcal{P}} \frac{\partial \mathcal{L}}{\partial \Delta^{\mathcal{S}}_{j,i}} \right]. \quad (11)$$

Method		ResNet-32×4 ResNet-8×4	ResNet-56 ResNet-20	ResNet-110 ResNet-32	WRN-40-2 WRN-40-1	WRN-40-2 WRN-16-2	VGG-13 VGG-8
Strategy	Teacher (baseline)	79.42	72.37	74.31	75.61	75.61	74.64
	Student (baseline)	72.5	69.06	71.14	71.98	73.26	70.36
Feature-based	FitNet[2014]	73.5	69.21	71.06	72.24	73.58	71.02
	AT[2016]	73.44	70.55	72.31	72.77	74.08	71.43
	VID[2019]	73.09	70.38	72.61	73.3	74.11	71.23
	RKD[2019]	71.9	69.61	71.82	72.22	73.35	71.48
	OFD[2019]	74.95	70.98	73.23	74.33	75.24	73.95
	CRD[2020]	75.51	71.16	73.48	74.14	75.48	73.94
	ReviewKD[2021]	75.63	71.89	73.89	75.09	76.12	74.84
Logit-based	KD[2015]	73.33	70.66	73.08	73.54	74.98	72.98
	TAKD[2020]	73.81	70.83	73.37	73.78	75.12	73.23
	DKD[2022]	76.32	**71.97**	74.11	74.81	76.24	<u>74.68</u>
	DIST[2022]	76.16	71.55	73.55	74.42	75.29	73.74
	NKD[2023]	76.35	71.62	73.79	<u>74.43</u>	<u>76.37</u>	74.32
	LSKD[2024]	<u>76.62</u>	71.43	<u>74.17</u>	74.37	76.11	74.36
	RKKD[2025]	74.74	71.09	73.09	73.93	75.87	74.14
	RCKD(Ours)	**76.89**	<u>71.83</u>	**74.26**	**74.67**	**76.60**	**74.86**

Table 1: Comparison with different distillation methods on CIFAR-100 for teacher/student model combinations with the **same architecture**. Among logit-based methods, the best and second-best results are marked in **bold** and <u>underlined</u>, respectively.

Crucially:

$$\sum_{c=1}^{C} \frac{\partial \mathcal{L}_{\mathrm{RCKD}}}{\partial z_c^{\mathcal{S}}} = \frac{1}{\tau}\left[\sum_{i=1}^{C}\sum_{k:(i,k)\in\mathcal{P}} \frac{\partial \mathcal{L}}{\partial \Delta_{i,k}^{\mathcal{S}}} - \sum_{i=1}^{C}\sum_{j:(j,i)\in\mathcal{P}} \frac{\partial \mathcal{L}}{\partial \Delta_{j,i}^{\mathcal{S}}}\right] = 0. \tag{12}$$

Despite the zero-sum property of gradients, RCKD's "zero-gradient-sum" characteristic stems exclusively from the antisymmetric structure of relative confidence ($\Delta_{j,k} = -\Delta_{k,j}$), not from Softmax's normalization constraint ($\sum_c p_c = 1$). Moreover, the RCKD's Hessian can exhibit both positive and negative off-diagonal elements. This implies that RCKD avoids the softmax-induced probability mass competition among classes.

Experiment

Experimental Setup

We evaluated our method on CIFAR-100 (Krizhevsky 2009) and ImageNet (Deng et al. 2009) datasets. We compared our method with a broad spectrum of KD approaches, including logit-based methods (KD (Hinton, Vinyals, and Dean 2015), TAKD (Mirzadeh et al. 2020), DKD (Zhao et al. 2022), DIST (Huang et al. 2022), NKD (Yang et al. 2023), LSKD (Sun et al. 2024), RKKD (Guan et al. 2025)) and feature-based methods (FitNet (Romero et al. 2014), AT (Zagoruyko and Komodakis 2017), RKD (Park et al. 2019), OFD (Heo et al. 2019), CRD (Tian, Krishnan, and Isola 2020), ReviewKD (Chen et al. 2021)). For teacher and student models, we employed variants of VGG (Simonyan and Zisserman 2014), ResNet (He et al. 2016), and WideResNet (Zagoruyko and Komodakis 2016). Moreover, to assess performance on lightweight models, we also employed MobileNet-V2 (Sandler et al. 2018) and ShuffleNet (Zhang et al. 2018).

To ensure fair comparison, we followed standard protocols from prior works. All models were trained using SGD with momentum 0.9. On CIFAR-100, we trained for 240 epochs with batch size 64, weight decay 5×10^{-4}, and initial learning rate 0.05 (0.01 for MobileNet/ShuffleNet), decayed by 0.1 at epochs 150, 180, and 210. On ImageNet, we trained for 100 epochs with batch size 512, weight decay 1×10^{-4}, and an initial learning rate of 0.2, decayed by 10 at epochs 30, 60, and 90. We set β=5 for our method, with temperature τ=4 on CIFAR-100 and τ=1 on ImageNet. Experiments were run on NVIDIA 4090 GPUs (1× for CIFAR-100, 4× for ImageNet), and all results were averaged over three runs.

Comparison with State-of-the-Art Methods

Performance on CIFAR-100 Tables 1 and 2 present the empirical results on the CIFAR-100 dataset for homogeneous and heterogeneous architectures, respectively. Our proposed method, RCKD, demonstrates a significant performance advantage over conventional KD. Notably, while being a logit-based method, RCKD's performance is comparable to, and often exceeds, that of more complex feature-based distillation techniques. These improvements underscore the effectiveness and strong generalizability of our method.

Performance on ImageNet To further validate the scalability and efficacy of our approach, we conducted experiments on the large-scale ImageNet dataset. The results, detailed in Table 3, show that RCKD yields substantial improvements in both Top-1 and Top-5 accuracy over baseline KD and other state-of-the-art methods. These gains are

Method		ResNet-32×4 ShuffleNet-V1	WRN-40-2 ShuffleNet-V1	ResNet-32×4 ShuffleNet-V2	ResNet-50 MobileNet-V2	VGG-13 MobileNet-V2	WRN-40-2 ResNet-8×4
Strategy	Teacher (baseline)	79.42	75.61	79.42	79.34	74.64	75.61
	Student (baseline)	70.5	70.5	71.82	64.6	64.6	72.5
Feature-based	FitNet[2014]	73.59	73.73	73.54	63.16	64.16	74.61
	AT[2016]	71.73	73.32	72.73	58.58	59.4	74.11
	VID[2019]	73.38	73.61	73.57	65.79	65.56	74.65
	RKD[2019]	72.28	72.21	73.21	64.43	64.52	75.26
	OFD[2019]	75.98	75.85	76.82	69.04	69.48	74.36
	CRD[2020]	75.11	76.05	75.65	69.11	69.73	75.24
	ReviewKD[2021]	77.45	77.14	77.78	69.89	70.37	74.34
Logit-based	KD[2015]	74.07	74.83	74.45	67.35	67.37	73.79
	TAKD[2020]	74.53	75.34	72.12	68.02	67.91	74.03
	DKD[2022]	76.45	76.7	77.07	**70.35**	69.71	75.56
	DIST[2022]	75.23	75.73	76.35	69.14	68.48	75.67
	NKD[2023]	75.31	75.96	76.26	69.39	68.72	76.01
	LSKD[2024]	75.19	76.03	75.56	69.02	68.61	77.11
	RKKD[2025]	75.98	76.13	75.35	69.45	68.23	76.25
	RCKD(Ours)	**76.51**	**76.97**	**77.13**	69.67	**69.77**	**77.23**

Table 2: Comparison with different distillation methods on **CIFAR-100** for teacher/student model combinations with **different architectures**. Among logit-based methods, the best and second-best results are marked in **bold** and underlined, respectively.

Teacher→Student	Acc.	Baselines		Feature-based Methods				Logit-based Methods				
		Teacher	Student	AT	OFD	CRD	Review KD	KD	DKD	LSKD	RKKD	RCKD
ResNet-34 → ResNet-18	Top-1	73.31	69.75	70.69	70.81	71.17	71.61	71.03	71.70	71.42	71.27	**72.02**
	Top-5	91.42	89.07	90.01	89.98	90.13	90.51	90.05	90.41	90.29	90.16	**90.67**
ResNet-50 → MobileNet-V1	Top-1	76.16	68.87	69.56	71.25	71.37	72.56	70.50	72.05	72.18	71.54	**72.41**
	Top-5	92.86	88.76	89.33	90.34	90.41	91.00	89.80	91.05	90.80	**90.84**	90.82

Table 3: Performance comparison of different distillation methods on ImageNet. Among logit-based methods, the best and second-best results are marked in **bold** and underlined, respectively.

evident in both homogeneous and heterogeneous teacher-student pairings, highlighting RCKD's robustness and suitability for deployment in large-scale tasks.

Computational Efficiency

As illustrated in Figure 4, RCKD achieves competitive training efficiency while delivering the highest accuracy among all logit-based methods. Compared with feature-based approaches, RCKD requires only the construction of the relative confidence matrix, and introduces no auxiliary network modules or computation-intensive feature operations. This lightweight design enables RCKD to deliver significant performance gains with minimal additional training cost.

Loss Function Comparison

To evaluate the effect of different distillation objectives under the same RCM-based knowledge representation, we compare the conventional MSE loss with our cosine similarity loss on CIFAR-100. As shown in Table 4, cosine similarity consistently outperforms MSE across all configurations. This verifies that MSE is sensitive to logit-scale discrepancies induced by capacity mismatch, causing the student to emphasize magnitude fitting rather than structural alignment. In contrast, cosine similarity focuses purely on directional consistency, enabling more faithful recovery of the teacher's relational confidence patterns.

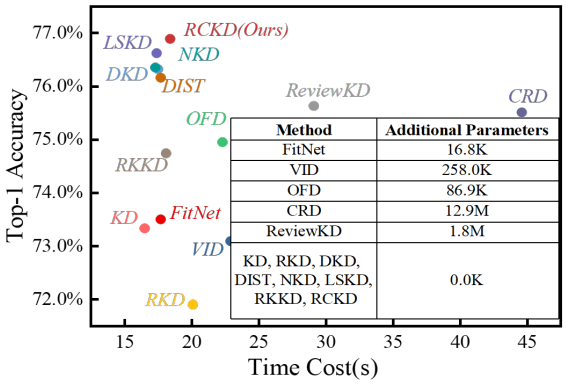

Figure 4: Comparison of different distillation methods in terms of training time, accuracy, and additional training parameters on the CIFAR-100 dataset, using the ResNet-32×4 (teacher) and ResNet-8×4 (student) model pair.

Architecture	Homogeneous Teacher-Student				Heterogeneous Teacher-Student			
Teacher	ResNet-32×4	ResNet-56	WRN-40-2	WRN-40-2	ResNet-32×4	WRN-40-2	WRN-40-2	VGG-13
Student	ResNet-8×4	ResNet-20	WRN-40-1	WRN-16-2	ShuffleNet-V2	ResNet-8×4	ShuffleNet-V1	MobileNet-V2
MSE	75.56	71.67	74.58	75.84	76.42	74.68	76.13	68.33
Cosine (ours)	**76.89**	**71.83**	**74.67**	**76.60**	**77.13**	**77.23**	**76.97**	**69.77**

Table 4: Comparison of distillation loss functions on CIFAR-100. The best results are marked in **bold**.

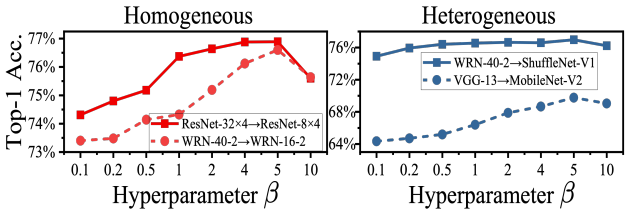

Figure 5: Different distillation coefficients on CIFAR-100.

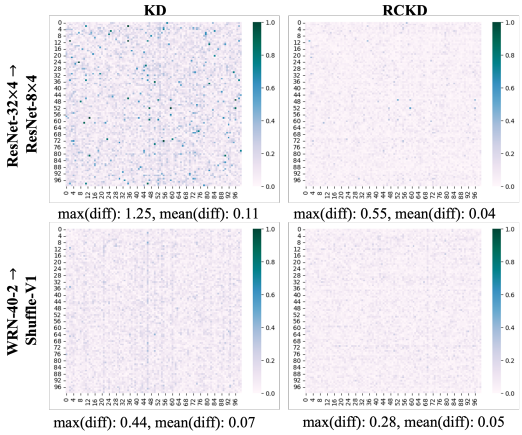

Figure 6: Correlation matrix of logits outputs for different teacher/student model combinations on the CIFAR-100.

Hyperparameter

To investigate the impact of the hyperparameter β on our method's performance, we conduct a sensitivity analysis on the CIFAR-100 dataset. We evaluate a range of β values, $\{0.1, 0.2, 0.5, 1, 2, 4, 5, 10\}$, across two distinct teacher-student configurations: homogeneous pair and heterogeneous pair. As illustrated in Figure 5, the results reveal that distillation performance is sensitive to the hyperparameter β. Both excessively small and large values of β degrade performance. Moreover, the optimal value of β remains largely consistent across both homogeneous and heterogeneous teacher-student architectures. Based on this analysis, we set $\beta = 5$ in our experiments.

Visualization

To qualitatively evaluate our method, we visualize the results on CIFAR-100 using two teacher-student pairs: ResNet-32×4→ResNet-8×4 and WRN-40-2→ShuffleNet-V1. Figure 6 presents the correlation matrices of teacher and student logits, where darker shades indicate greater discrepancies. The visualization demonstrates that our method (RCKD) effectively guides the student to produce logits that are more consistent with those of the teacher. On the other hand, we employed t-SNE to visualize the feature space learned by the student models. As illustrated in Figure 7, RCKD achieves markedly better class separability than the standard KD baseline, confirming that our approach enhances the discriminative power of the student's learned features.

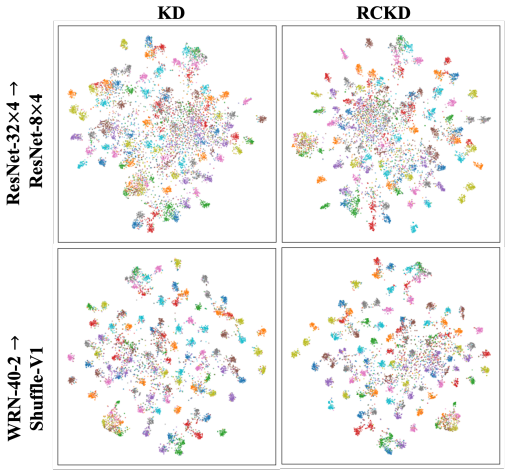

Figure 7: t-SNE visualization of students under different teacher-student pairs on CIFAR-100.

Conclusion

In this paper, we revisit the limitations of KL-divergence-based KD. We demonstrate that the pointwise numerical alignment enforced by KL divergence fails to preserve the teacher's relative class ranking, while its inherent gradient competition actively distorts ordinal relationships under capacity mismatch. To overcome these issues, we propose Relative Confidence Knowledge Distillation (RCKD), which explicitly transfers numerical and directional knowledge by aligning the student and teacher relative confidence matrices through cosine similarity. Extensive experiments validate that RCKD consistently outperforms state-of-the-art logit-based KD methods across diverse teacher-student architectures, with acceptable computational overhead. Future work includes extending RCKD to multimodal distillation and exploring learnable hyperparameters.

References

Chen, P.; Liu, S.; Zhao, H.; and Jia, J. 2021. Distilling Knowledge via Knowledge Review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5008–5017.

Cho, J. H.; and Hariharan, B. 2019. On the Efficacy of Knowledge Distillation. In *Proceedings of the IEEE/CVF international Conference on Computer Vision*, 4794–4802.

Cui, J.; Tian, Z.; Zhong, Z.; Qi, X.; Yu, B.; and Zhang, H. 2024. Decoupled Kullback-Leibler Divergence Loss. *Advances in Neural Information Processing Systems*, 37: 74461–74486.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Fan, W.-S.; Lu, S.; Li, X.-C.; Zhan, D.-C.; and Gan, L. 2024. Revisit the Essence of Distilling Knowledge through Calibration. In *International Conference on Machine Learning*, 12882–12894.

Guan, Y.; Cheng, R.; Liu, K.; and Yuan, C. 2025. Enhancing Logits Distillation with Plug and Play Kendall's τ Ranking Loss. In *Forty-second International Conference on Machine Learning*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Heo, B.; Kim, J.; Yun, S.; Park, H.; Kwak, N.; and Choi, J. Y. 2019. A Comprehensive Overhaul of Feature Distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1921–1930.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*.

Huang, T.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2022. Knowledge Distillation from a Stronger Teacher. *Advances in Neural Information Processing Systems*, 35: 33716–33727.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Li, Z.; Li, X.; Yang, L.; Zhao, B.; Song, R.; Luo, L.; Li, J.; and Yang, J. 2023. Curriculum Temperature for Knowledge Distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1504–1512.

Liang, P.; Zhang, W.; Wang, J.; and Guo, Y. 2024. Neighbor Self-Knowledge Distillation. *Information Sciences*, 654: 119859.

Mirzadeh, S. I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; and Ghasemzadeh, H. 2020. Improved Knowledge Distillation via Teacher Assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5191–5198.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3967–3976.

Rao, J.; Meng, X.; Ding, L.; Qi, S.; Liu, X.; Zhang, M.; and Tao, D. 2023. Parameter-Efficient and Student-Friendly Knowledge Distillation. *IEEE Transactions on Multimedia*, 26: 4230–4241.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. FitNets: Hints for Thin Deep Nets. *arXiv preprint arXiv:1412.6550*.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Simonyan, K.; and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*.

Son, W.; Na, J.; Choi, J.; and Hwang, W. 2021. Densely Guided Knowledge Distillation Using Multiple Teacher Assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9395–9404.

Sun, S.; Ren, W.; Li, J.; Wang, R.; and Cao, X. 2024. Logit Standardization in Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15731–15740.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Representation Distillation. In *International Conference on Learning Representations*.

Wang, C.; Yang, Q.; Huang, R.; Song, S.; and Huang, G. 2022. Efficient Knowledge Distillation from Model Checkpoints. *Advances in Neural Information Processing Systems*, 35: 607–619.

Wang, Y.; Yang, C.; Lan, S.; Zhu, L.; and Zhang, Y. 2024. End-Edge-Cloud Collaborative Computing for Deep Learning: A Comprehensive Survey. *IEEE Communications Surveys & Tutorials*, 26(4): 2647–2683.

Yang, C.; Wang, Y.; Lan, S.; Wang, L.; Shen, W.; and Huang, G. Q. 2022. Cloud-Edge-Device Collaboration Mechanisms of Deep Learning Models for Smart Robots in Mass Personalization. *Robotics and Computer-Integrated Manufacturing*, 77: 102351.

Yang, Z.; Zeng, A.; Li, Z.; Zhang, T.; Yuan, C.; and Li, Y. 2023. From Knowledge Distillation to Self-Knowledge Distillation: A Unified Approach with Normalized Loss and Customized Soft Labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17185–17194.

Yin, S.; Xiao, Z.; Song, M.; and Long, J. 2024. Adversarial Distillation Based on Slack Matching and Attribution Region Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24605–24614.

Yuan, M.; Lang, B.; and Quan, F. 2024. Student-Friendly Knowledge Distillation. *Knowledge-Based Systems*, 296: 111915.

Zagoruyko, S.; and Komodakis, N. 2016. Wide Residual Networks. In *BMVC*.

Zagoruyko, S.; and Komodakis, N. 2017. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *International Conference on Learning Representations*.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6848–6856.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11953–11962.

Zhu, Y.; Liu, N.; Xu, Z.; Liu, X.; Meng, W.; Wang, L.; Ou, Z.; and Tang, J. 2022. Teach Less, Learn More: On the Undistillable Classes in Knowledge Distillation. *Advances in Neural Information Processing Systems*, 35: 32011–32024.

Zhu, Y.; and Wang, Y. 2021. Student Customized Knowledge Distillation: Bridging the Gap between Student and Teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5057–5066.