ENHANCING LOGITS DISTILLATION WITH PLUG&PLAY KENDALL'S τ RANKING LOSS

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation typically employs the Kullback-Leibler (KL) divergence to constrain the output of the student model to precisely match the soft labels provided by the teacher model. However, the optimization process of KL divergence is challenging for the student and prone to suboptimal points. Also, we demonstrate that the gradients provided by KL divergence depend on channel scale and thus tend to overlook low-probability channels. The mismatch in low-probability channels also results in the neglect of inter-class relationship information, making it difficult for the student to further enhance performance. To address this issue, we propose an auxiliary ranking loss based on Kendall's τ Coefficient, which can be plug-and-play in any logit-based distillation method, providing inter-class relationship information and balancing the attention to low-probability channels. We show that the proposed ranking loss is less affected by channel scale, and its optimization objective is consistent with that of KL divergence. Extensive experiments on CIFAR-100, ImageNet, and COCO datasets, as well as various CNN and ViT teacher-student architecture combinations, demonstrate that the proposed ranking loss can be plug-and-play on various baselines and enhance their performance.

1 INTRODUCTION

The recent advancements in deep neural networks (DNN) have significantly enhanced performance in the field of computer vision. However, the heightened computational and storage costs associated with complex networks limite their applicability. To address this, knowledge distillation (KD) has been proposed to obtain performant lightweight models. Typically, knowledge distillation involves a well-trained heavy teacher model and an untrained lightweight student model. The same data is fed to both the teacher and the student, with the teacher's outputs serving as soft labels for training the student. By constraining the student to produce predictions that match the soft labels, the knowledge in the teacher model is transferred to the lightweight student model.

Most logit-based KD methods adhere to the paradigm introduced by Hinton (Hinton et al., 2015), employing the Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951) to align the logits output by the student and teacher models:

$$\mathcal{L}_{KL}(q^t \parallel q^s) = \sum_{i=1}^{C} q_i^t \cdot log(\frac{q_i^t}{q_i^s}). \tag{1}$$

Where $q_t, q_s \in \mathbb{R}^{1 \times C}$ is the prediction vectors of teacher and student. The optimization goal of KL divergence is to achieve identical outputs between the student and teacher, thereby transferring as much knowledge as possible from the teacher to the student. However, the optimization process of KL divergence is not easy, as it is prone to suboptimal points, which can hinder further improvement in student performance. As illustrated in Fig. 2, compared to Student 2, Student 1 exhibits a smaller KL divergence with the teacher; however, Student 2 achieves the correct classification result consistent with the teacher, while Student 1 does not. This indicates that the optimization direction of KL divergence sometimes diverges from the task objective, leading students to suboptimal points.

Intuitively, as shown in Eq. 1, the weight of matching a channel in KL divergence is the probability of the teacher in that channel. This indicates that KL divergence tends to overlook the matching of low-probability channels. In practice, most channels in a single prediction have lower probabilities,

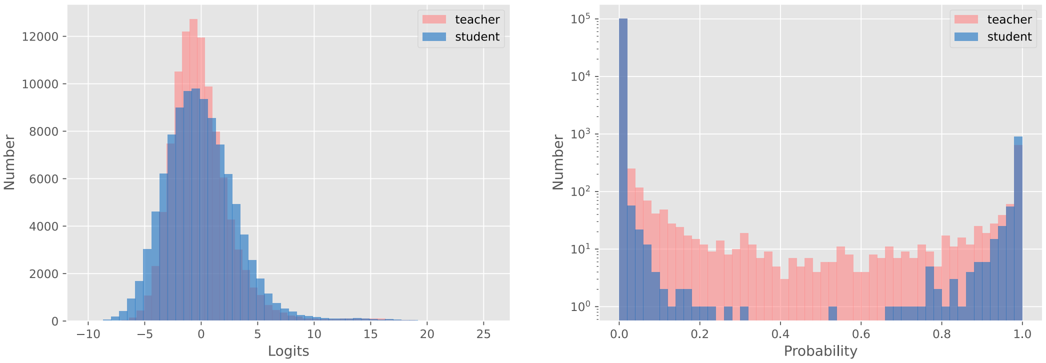


Figure 1: Logits Value Distributions. **Left:** The original logits output by teacher and student. **Right:** The probability output by teacher and student.

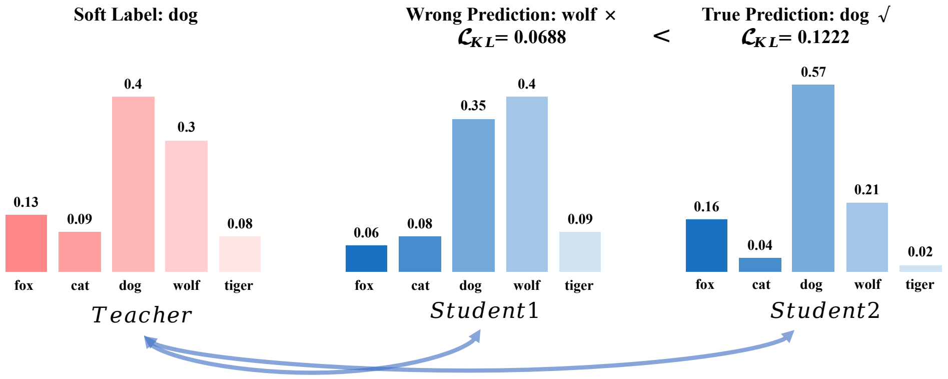


Figure 2: Suboptimal Case of the KL Divergence.

as illustrated in Fig. 1. Neglecting these channels affects the student's further learning from the teacher for two main reasons:

R1: The neglect of low-probability channels leads to lower matching degrees for some channels. Due to their low weights, these channels receive smaller gradients during optimization, making it harder to learn knowledge.
R2: The neglect of low-probability channels results in the missing of some inter-class relationships from the teacher. For example, in Fig. 2, Student 1 may not learn the distinction between the cat and fox classes.

Considering these challenges, we aim to find an auxiliary method that mitigates the issues caused by channel scale differences, learns inter-class relationship information, and avoids the suboptimal points that KL divergence might encounter while maintaining the optimization objective of KL divergence. To address this problem, we propose a ranking-based plug-and-play auxiliary loss. The benefits of imposing ranking constraints are as follows:

B1: The gradients provided by ranking loss are less affected by channel scale.
B2: The channel ranking provides inter-class relationship knowledge.
B3: By constraining the rank of target channels, ranking loss helps to avoid suboptimal solutions.

Therefore, we propose to constrain the channel ranking similarity between student and teacher. We construct a plug-and-play ranking loss function based on Kendall's τ Coefficient. This ranking loss can supplement the attention to smaller channels in the logits and provide inter-class relationship information. We demonstrate that the gradients provided by KL divergence are related to channel scale, whereas the proposed ranking loss is not. We also show that the optimization objective of the proposed ranking loss is consistent with that of KL divergence, and visualize the loss landscape to illustrate that the ranking loss helps avoid suboptimal points in the early stage and does not alter the optimization objective in the end. Extensive experiments on CNN and ViT show that the proposed plug-and-play ranking loss can enhance the performance of various logit-based methods. In conclusion, our contributions are as follows:

- We introduce a plug-and-play ranking loss for assisting knowledge distillation tasks, addressing the issues of KL divergence's neglect of low-probability channels and its tendency to fall into suboptimal points, while also learning inter-class relationship information to further enhance student performance.
- We demonstrate that the gradients provided by KL divergence are related to channel scale, whereas the proposed ranking loss is not. We also prove and visualize that the optimization objective of the proposed ranking loss is consistent with that of KL divergence and helps avoid the suboptimal points of KL divergence.
- Extensive experiments are conducted on a variety of CNN and ViT teacher-student architectures using the CIFAR-100, ImageNet, and MS-COCO datasets. Our findings confirm the widespread effectiveness of the proposed ranking loss in various distillation tasks, and its role as a plug-and-play auxiliary function provides substantial support for the training of distillation tasks.

2 RELATED WORKS

Knowledge Distillation. Knowledge distillation, initially proposed by Hinton (Hinton et al., 2015), serves as a method for model compression and acceleration, aiming to transfer knowledge from a heavy teacher model to a lightweight student model. By feeding the same samples, the teacher can produce soft labels, and training the student with these soft labels allows the transfer of knowledge from the teacher model to the student. Knowledge distillation tasks can be mainly divided into two categories: feature-based distillation (Chen et al., 2022; Zhang & Ma, 2021; Yang et al., 2022b; Guo et al., 2023; Park et al., 2019) and logit-based distillation (Hinton et al., 2015; Li et al., 2023; Zhao et al., 2022; Jin et al., 2023; Sun et al., 2024; Chi et al., 2023; Wen et al., 2021). Feature-based distillation additionally utilizes the model's intermediate features, providing more information to the student and enabling the student to learn at the feature level from the teacher, often resulting in better performance. Considering safety and privacy, the intermediate outputs of the model are often not obtainable; hence, logit-based methods that only use the model outputs for distillation offer better versatility and robustness. Our proposed method can act as a plug-and-play module added to logit-based methods, offering higher flexibility and further enhancing the performance of logit-based methods.

Recent knowledge distillation methods have found that overly strict constraints can sometimes hinder the student's transfer of knowledge from the teacher's soft labels. For instance, (Cao et al., 2022) discovered that differences in feature sizes in feature-based methods could limit the student's learning; while (Sun et al., 2024) found that using the same temperature for both teacher and student in logit-based methods could affect further improvements in student performance. To reduce the learning difficulty for the student, (Sun et al., 2024; Cao et al., 2022) provided methods to ease the logit matching difficulty between student and teacher, yet we find that methods providing additional guidance to the student have not been fully explored.

Ranking Loss in Knowledge Distillation. For knowledge distillation, ranking loss is a relaxed constraint that can provide rich inter-class information. It was first applied in the distillation of recommendation systems (Reddi et al., 2021; Tang & Wang, 2018; Choi et al., 2021; Qin et al., 2023; Yang et al., 2022a), (Li et al., 2022) explored the application of ranking loss in object detection tasks, and (Gao et al., 2020) discussed the role of ranking in the distillation of language model tasks. However, the exploration of ranking loss in logit-based image classification task distillation is not yet comprehensive. We find that the KL divergence used for distillation in classification tasks tends to overlook information from smaller-valued channels and may lead to suboptimal results. Our proposed method leverages ranking loss to balance the model's attention to larger and smaller-valued channels, while also using inter-class relationships to help the model avoid suboptimal outcomes.

3 PRELIMINARY

Most logit-based distillation methods adhere to the original KD proposed by Hinton (Hinton et al., 2015), which transfers knowledge by matching the logit outputs of the student and teacher. This setting is more generalizable, allowing for distillation solely through outputs when the internal structures of the student and teacher are invisible. For a given dataset $\mathcal{D}$, assuming there are C categories and N samples, we possess a teacher model f_t and a student model f_s. For a given sample $\mathcal{I} \in \mathcal{D}$, we can obtain the outputs of the teacher and student model, denoted as $z^t = f_t(\mathcal{I}), z^s = f_s(\mathcal{I})$ where $z^t, z^s \in \mathbb{R}^{1\times C}$. Through a softmax function with temperature, the outputs are processed into the prediction vectors $q_t, q_s \in \mathbb{R}^{1\times C}$ finally:

$$q_i^t = \frac{exp(z_i^t/\mathcal{T})}{\sum_{j=1}^{C} exp(z_j^t/\mathcal{T})} \tag{2}$$

$$q_i^t = \frac{exp(z_i^s/\mathcal{T})}{\sum_{j=1}^{C} exp(z_j^s/\mathcal{T})} \tag{3}$$

where $\mathcal{T}$ is a temperature parameter, z_i represents the logit value of the i-th channel of the model output, q_i represents the predicted probability for the target being the i-th class. The KL divergence loss function used to constrain the student and teacher logits is of the following form:

$$\mathcal{L}_{KL}(q^t \parallel q^s) = \sum_{i=1}^{C} q_i^t \cdot log(\frac{q_i^t}{q_i^s}) \tag{4}$$

It can be observed that in Eq. 4, the importance of the match between the student and teacher at the i-th channel is influenced by the coefficient q_i^t. This implies that channels with smaller logit values receive less attention, leading to the KL divergence's disregard for smaller channels.

4 RANKING LOSS BASED ON KENDALL'S τ COEFFICIENT

In this section, we introduce the plug-and-play ranking loss function designed to constrain the ranking of channels in the logits output by the student model. With the aid of the ranking loss, the knowledge distillation task can balance the overemphasis on larger logit values and the neglect of smaller ones as measured by the KL divergence. Additionally, the ranking loss imposes a constraint on the leading channel value, which helps to correct the optimization direction and avoid suboptimal solutions. In Section4.1, we will introduce the ranking loss function and employ Kendall's τ coefficient to compute the ordinal consistency between the teacher and student logits. In Section 4.2, we discuss how ranking loss benefits optimizing logits distillation with KL divergence. Furthermore, we discuss the differentiable form of Kendall's τ coefficient and, based on this, design three distinct forms of ranking loss functions in Appendix A.4.

4.1 DIFFERENTIABLE KENDALL'S τ COEFFICIENT

In order to make the loss function pay more attention to the information provided by smaller channels as well as to help correct the suboptimal problem in Figure 1, we introduce the ranking loss based on Kendall's τ coefficient. By pairing each of the C channels, we can obtain $\frac{C(C-1)}{2}$ pairs, whose Kendall's τ coefficient is expressed as follows:

$$\tau = \frac{P_c - P_d}{\frac{1}{2}C(C-1)} \tag{5}$$

where P_c represents concordant pairs and P_d represents discordant pairs. Kendall's τ coefficient provides an expression of ordinal similarity. For the teacher and student logits, a channel pair (i, j) is considered concordant if the signs of $(z_i^t - z_j^t)$ for the teacher and$(z_i^s - z_j^s)$ for the student are the same; otherwise, the pair is discordant. We substitute the logits of the teacher and student in the following manner:

$$\tau = \frac{\sum_i \sum_{j<i} sgn(z_i^t - z_j^t) \cdot sgn(z_i^s - z_j^s)}{\frac{1}{2}C(C-1)} \tag{6}$$

where $sgn()$ represents sign function. While we have quantified the ordinal similarity between the teacher and student logits, the aforementioned formula is non-smooth. To utilize the ordinal similarity for gradient computation, we approximate the sign function with the $tanh$ function to convert it into a differentiable form:

$$\begin{aligned}\tau_d &= \frac{\sum_i \sum_{j<i} tanh(k \cdot (z_i^t - z_j^t)) \cdot tanh(k \cdot (z_i^s - z_j^s))}{\frac{1}{2}C(C-1)} \\ &= \frac{2}{C(C-1)} \cdot \sum_i \sum_{j<i} \left(1 - \frac{2}{1+e^{2\cdot(z_i^t - z_j^t)\cdot k}}\right) \cdot \left(1 - \frac{2}{1+e^{2\cdot(z_i^s - z_j^s)\cdot k}}\right)\end{aligned} \tag{7}$$

where k is a parameter that controls the steepness of the function; a larger k causes the $tanh$ function to more closely approximate the sign function. By negating the similarity measure in Eq. 7, it can serve as a loss function to enforce the consistency of the logits order between the teacher and student:

$$\mathcal{L}_{RK} = -\tau_d = -\frac{2}{C(C-1)} \cdot \sum_i \sum_{j<i} \left(1 - \frac{2}{1+e^{2\cdot(z_i^t - z_j^t)\cdot k}}\right) \cdot \left(1 - \frac{2}{1+e^{2\cdot(z_i^s - z_j^s)\cdot k}}\right) \tag{8}$$

The overall loss function is formulated as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{KL} + \beta\mathcal{L}_{CE} + \gamma\mathcal{L}_{RK} \tag{9}$$

where $\mathcal{L}_{KL}$ denotes KL divergence loss, $\mathcal{L}_{CE}$ denotes Cross-Entropy loss, $\mathcal{L}_{RK}$ denotes our ranking loss, and α, β, γ are hyper-parameters.

4.2 How Ranking Loss Benefits Optimizing Logits Distillation with KL Divergence

4.2.1 From the Perspective of Gradient: Ranking Loss Cares about Smaller Channels

In this section, we analyze why ranking loss cares about smaller channels. The gradient of KL divergence and ranking loss is written as follows. The specific calculation process of the gradient can be found in the appendix A.6:

$$\frac{\partial \mathcal{L}_{\text{KD}}}{\partial z_i^s} = -T\left(q_i^t - q_i^s\right). \tag{10}$$

$$\frac{\partial \mathcal{L}_{RK}}{\partial z_i^s} = -\frac{k}{C(C-1)} \sum_{j\neq i} \left[1 - \tanh^2\left(k(z_i^s - z_j^s)\right)\right] \tanh\left(k(z_i^t - z_j^t)\right) \tag{11}$$

The gradient scale of the KL divergence is influenced by the scales of the teacher and student models. Consequently, for outputs from smaller channels, when the scales of the teacher and student are similar, the KL divergence's gradient for these components becomes negligible, leading to the neglect of information from smaller channels. Although temperature scaling is typically employed to address this issue, it uniformly amplifies the gradients of all logits, thereby continuing to overlook the outputs of smaller channels. In contrast, the gradients produced by ranking loss will be less affected by the scale of the teacher's logits. The gradient of a logit channel primarily depends on the difference between its rank and the target rank, effectively harnessing the knowledge from smaller channels. Therefore. In the early stage of training, ranking loss helps the model quickly learn the overall ranking of the teacher model, which helps the model converge to a better initial solution faster.

4.2.2 From the Perspective of Optimal Solution Domain: Ranking Loss won't Interfere with KL Ddivergence at the End

In this section, we analyze the optimal solution domain for Kullback-Leibler (KL) divergence and ranking loss to show how ranking loss affects the optimization process of KL divergence. For KL divergence, the optimal solution domain is achieved when the distribution of the student model aligns perfectly with that of the teacher model, which is equivalent to a linear mapping on the logits.:

$$q_i^t = q_i^s \quad \forall i \quad \Rightarrow \quad z_i^t = z_i^s + c \quad \forall i \tag{12}$$

p_i^t and p_i^s respectively represent the probability outputs of the teacher and the student for the i-th class, and c represents any constant number. For the ranking loss, the optimal solution domain is such that for any two indices, the order of the logits output by the teacher and the student is consistent:

$$\begin{aligned} & sgn(z_i^t - z_j^t) \cdot sgn(z_i^s - z_j^s) = 1 \quad \forall i, j \\ \iff & (z_i^t - z_j^t) \cdot (z_i^s - z_j^s) > 0 \quad \forall i, j \\ \iff & z_i^t = F(z_i^s) \quad \forall i \quad where \quad F'(x) > 0 \end{aligned} \tag{13}$$

It implies that the optimal solution domain for ranking loss is quite lenient and easily attainable, encompassing the optimal solution domain for KL divergence. Therefore, in the later stage of training, ranking loss will not hinder the optimization of KL divergence, which makes our ranking loss can perfectly be used as an auxiliary loss.

4.2.3 FROM THE PERSPECTIVE OF CLASSIFICATION: RANKING LOSS HELPS STUDENT CLASSIFY CORRECTLY

As illustrated in Figure 1, using KL loss may lead to a smaller loss yet result in incorrect classification. The tendency to fall into local optima will hinder the optimization process in distillation. Although the use of cross-entropy loss can mitigate this issue, it does not incorporate information from the teacher model. Consequently, in experiments, cross-entropy loss is often assigned a very small weight, which also leads to a small gradient. For instance, KD (Hinton et al., 2015)$\mathcal{L}_{\text{CE}} : \mathcal{L}_{\text{KL}} = 0.1 : 0.9$; DKD (Zhao et al., 2022) $\mathcal{L}_{\text{CE}} : \mathcal{L}_{\text{TCKL}} : \mathcal{L}_{\text{NCKL}} = 1 : 1 : 8$; MLKD (Jin et al., 2023) $\mathcal{L}_{\text{CE}} : \mathcal{L}_{\text{KL}} = 0.1 : 9$; LSKD (Sun et al., 2024) $\mathcal{L}_{\text{CE}} : \mathcal{L}_{\text{KL}} = 0.1 : 9$. As a result, cases with lower KL loss but incorrect classification still frequently occur. In contrast, by aligning the rank between student and teacher, ranking loss helps KL divergence avoid such suboptimal situations. Also, the ranking loss can incorporate information from the teacher model. In this way, ranking loss helps students classify correctly, avoiding the suboptimal case in logit distillation. Furthermore, a model with better generalization should have a more reasonable rank of logits. For instance, recognizing that a tiger is more similar to a cat than to a fish. By learning such inter-class relationships, the student can improve classification performance and enhance its representational capacity.

5 EXPERIMENT

Datasets. In order to validate the efficacy and robustness of our proposed method, we conduct widely experiments on three datasets. 1) CIFAR-100 (Krizhevsky et al., 2009) is a significant dataset for image classification, comprising 100 categories, with 50,000 training images and 10,000 test images. 2) ImageNet (Russakovsky et al., 2015) is a large-scale dataset utilized for image classification, comprising 1,000 categories, with approximately 1.28 million training images and 50,000 test images. 3) MS-COCO (Lin et al., 2014) is a mainstream dataset for object detection comprising 80 categories, with 118,000 training images and 5,000 test images.

Baselines. As a plug-and-play loss, we apply the proposed ranking loss to various logit-based methods, including KD (Hinton et al., 2015), CTKD (Li et al., 2023), DKD (Zhao et al., 2022), and MLKD (Jin et al., 2023), to verify whether it can bring performance gains to knowledge distillation methods. We also compare it with various feature-based methods, including FitNet (Adriana et al., 2015), CRD (Tian et al., 2019), and ReviewKD (Chen et al., 2021). Additionally, we compare it with other auxiliary losses and modified KL divergence methods, including DIST (Huang et al., 2022) and LSKD (Sun et al., 2024).

Implementation Details. To ensure the robustness of the proposed plug-and-play loss without introducing excessive configurations, we maintain the same experimental settings as the baselines used (KD+Ours and KD share the same experimental setups for example). We set the batch size to 64 for CIFAR-100, 512 for ImageNet and 8 for COCO. We employ SGD (Sutskever et al., 2013) as the optimizer, with the number of epochs and learning rate settings consistent with the comparative baselines. The hyper-parameters α, β in Eq. 7 are set to be the same as the compared baselines to maintain fairness, and γ are set equal to α. We utilize 1 NVIDIA GeForce RTX 4090 to train models on CIFAR-100 and 4 NVIDIA GeForce RTX 4090 for training on ImageNet. The algorithm of our method can be found in Appendix A.5.

Table 1: CIFAR-100 Heterogeneous Architecture Results. The Top-1 Accuracy (%) is reported as the evaluation metric. The teacher and student have heterogeneous architectures. We incorporate the proposed ranking loss into the existing logit-based methods, with the performance gains indicated in parentheses. The best and second best results are emphasized in **bold** and underlined.

KD						
KD	Teacher	ResNet32×4	ResNet32×4	ResNet50	ResNet32×4	WRN-40-2
		79.42	79.42	79.34	79.42	75.61
	Student	WRN-16-2	WRN-40-2	MN-V2	SHN-V1	SHN-V1
		73.26	75.61	64.60	70.50	70.50
	FitNet (Adriana et al., 2015)	74.70	77.69	63.16	73.59	73.73
	CRD (Tian et al., 2019)	75.65	78.15	69.11	75.11	76.05
	ReviewKD (Chen et al., 2021)	76.11	78.96	69.89	77.45	77.14
	DIST (Huang et al., 2022)	75.58	78.02	68.66	76.34	76.00
	LSKD (Sun et al., 2024)	**77.53**	79.66	71.19	76.48	76.93
	KD (Hinton et al., 2015)	74.9	77.7	67.35	74.07	74.83
	KD+Ours	75.18(+0.28)	78.50(+0.80)	70.45(+3.10)	75.98(+1.91)	76.13(+1.30)
	CTKD (Li et al., 2023)	74.57	77.66	68.67	74.48	75.61
	CTKD+Ours	75.71(+1.14)	78.61(+0.95)	70.18(+1.51)	76.67(+2.19)	76.80(+1.19)
	DKD (Zhao et al., 2022)	75.7	78.46	70.35	76.35	76.33
	DKD+Ours	75.99(+0.29)	78.75(+0.29)	70.90(+0.55)	77.36(+1.01)	76.43(+0.10)
	MLKD (Jin et al., 2023)	76.52	79.26	71.04	77.18	77.44
	MLKD+Ours	76.83(+0.31)	**79.86(+0.60)**	**71.66(+0.62)**	**77.63(+0.45)**	**77.87(+0.43)**

Table 2: CIFAR-100 Homogenous Architecture Results. The Top-1 Accuracy (%) is reported as the evaluation metric. The teacher and student have homogeneous architectures. We incorporate the proposed ranking loss into the existing logit-based methods, with the performance gains indicated in parentheses. The best and second best results are emphasized in **bold** and underlined.

KD					
KD	Teacher	ResNet32×4	VGG13	WRN-40-2	ResNet110
		79.42	74.64	75.61	74.31
	Student	ResNet8×4	VGG8	WRN-40-1	ResNet20
		72.50	70.36	71.98	69.06
	FitNet (Adriana et al., 2015)	73.50	71.02	72.24	68.99
	CRD (Tian et al., 2019)	75.51	73.94	74.14	71.46
	ReviewKD (Chen et al., 2021)	75.63	74.84	75.09	71.34
	DIST (Huang et al., 2022)	76.31	73.80	74.73	71.40
	LSKD (Sun et al., 2024)	**78.28**	75.22	75.56	72.27
	KD (Hinton et al., 2015)	73.33	72.98	73.54	70.67
	KD+Ours	74.74(+1.41)	74.14(+1.16)	74.49(+0.95)	71.09(+0.42)
	CTKD (Li et al., 2023)	73.39	73.52	73.93	70.99
	CTKD+Ours	75.59(+2.2)	74.76(+1.24)	74.86(+0.93)	71.08(+0.09)
	DKD (Zhao et al., 2022)	76.32	74.68	74.81	71.06
	DKD+Ours	76.61(+0.29)	75.1(+0.42)	74.94 (+0.13)	71.84(+0.78)
	MLKD (Jin et al., 2023)	77.08	75.18	75.35	71.89
	MLKD+Ours	77.25(+0.17)	**75.35(+0.17)**	**76.08(+0.73)**	**72.35(+0.46)**

5.1 MAIN RESULT

CIFAR-100 Results. In our study, we conduct a comparative analysis of Knowledge Distillation (KD) outcomes across various teacher-student (He et al., 2016; Simonyan & Zisserman, 2014; Zagoruyko & Komodakis, 2016; Zhang et al., 2018; Howard et al., 2017; Sandler et al., 2018) configurations. While Tab. 1 presents cases where the teacher and student models share the heterogeneous architecture, Tab. 2 illustrates instances of homogenous structures. Furthermore, as a plug-and-play method, we applied ranking loss in multiple logit-based distillation techniques. The incorporation of ranking loss resulted in an average improvement of 1.83% in vanilla KD. In the context of existing state-of-the-art (SOTA) logits-based methods, including DKD, CTKD, and MLKD, significant gains are made.

ImageNet Results. The results on ImageNet of KD in terms of top-1 and top-5 accuracy are compared in Tab.3. Our proposed method can achieve consistent improvement on the large-scale dataset as well.

Table 3: The top-1 and top-5 accuracy (%) on the ImageNet validation set. The teacher and student are ResNet50 and MN-V1

	AT	OFD	CRD	KD	KD+RKKD
Top-1	69.56	71.25	71.37	70.50	**71.54(+1.04)**
Top-5	89.33	90.34	90.41	89.80	**90.84(+1.04)**

5.2 EXTENSIONS

KD for Transformer. To validate the effectiveness of our plug-and-play ranking loss on ViT and to assess its performance when facing larger teacher-student structural differences, we conduct experiments using ViT students. The experimental results indicate that the incorporation of ranking loss yields significant improvements over the conventional knowledge distillation, as shown in Tab. 4. This denotes that our proposed method is also applicable to distillation challenges predicated on transformer architectures. The implementation details of the Transformer experiments are attached in the Appendix.

Table 4: KD for Transformer. The Top-1 Accuracy (%) on the validation set of CIFAR-100. The Teacher model is ResNet56.

Student	DeiT-Tiny 65.08	T2T-ViT-7 69.37	PiT-Tiny 73.58	PVT-Tiny 69.22
KD	71.11	71.72	74.03	72.46
KD+RKKD	73.25(+2.14)	72.49(+0.77)	74.33(+0.30)	73.42(+0.96)

KD for Object Detection. To further verify the generality of our proposed method in downstream tasks, we also conduct experiments under the setting of object detection. The results show that our proposed ranking loss can improve the performance of knowledge distillation method in object detection and achieve better performance than the same period of the feature-based object detection method, which is shown in Tab. 5.

Table 5: KD for Object Detection. All experiments are conducted on COCO2017 with the teacher as ResNet50 and the student as MobileNet-V2.

Method	AP	AP50	AP75
KD (Hinton et al., 2015)	30.13	50.28	31.35
FitNet (Adriana et al., 2015)	30.20	49.80	31.69
FGFI (Wang et al., 2019)	31.16	50.68	32.92
KD+RKKD	31.99(+1.86)	53.80(+3.52)	33.37(+2.02)

Ablation Study. We conduct extensive ablation studies to investigate the effectiveness of ranking loss under various settings of k and coefficients. Tab. 7 presents the distillation outcomes across different k configurations. It is observed that a larger k value yields superior performance, suggesting that the differential form of ranking loss more closely approximates the Kendall τ coefficient, thereby imparting stronger ranking knowledge. Tab. 6 delineates the performance of ranking loss under varying coefficients. This setting effectively aligns the classification outcomes of the teacher and student models, thus significantly enhancing the distillation effect.

More Experiments. Additional experiments and discussions, including *Ablation of Temperature*, *Ablation of Normalization*, and *Different Forms of Ranking Loss*, are provided in the Appendix. Please refer to the Sec. A for further details.

5.3 ANALYSIS

Accuracy & Loss Curves with Ranking Loss. The accuracy and loss curves of KD and KD+Ours, as shown in Fig. 3, demonstrate how ranking loss aids in optimization. The middle figure shows that the precise alignment of KL divergence also makes channel ranking more ordered, but adding ranking loss achieves a more consistent ranking more quickly. The right figure shows that in the early stages, ranking loss accelerates the reduction of KL loss and reduces its oscillation in suboptimal

Table 6: Ablation of Weight. The Top-1 Accuracy (%) on the validation set of CIFAR-100.

Teacher	ResNet32×4 79.42	WRN-40-2 75.61	ResNet32×4 79.42	ResNet50 79.34
Student	ResNet8×4 72.50	WRN-40-1 71.98	SHN-V2 71.28	MN-V2 64.60
KD(Baseline)	73.33	73.54	74.45	67.35
$\gamma = 0.1$	74.15	74.15	75.52	69.25
$\gamma = 0.5$	74.84	74.07	76.34	69.81
$\gamma = 0.9$	74.74	74.49	76.58	70.45
$\gamma = 2$	74.62	**74.65**	**77.07**	69.97
$\gamma = 4$	**75.07**	73.60	77.05	**70.59**
$\gamma = 6$	74.78	73.09	76.90	69.58

Table 7: Ablation of k. The Top-1 Accuracy (%) on the validation set of CIFAR-100.

Teacher	ResNet32×4 79.42	WRN-40-2 75.61	ResNet32×4 79.42	ResNet50 79.34
Student	ResNet8×4 72.50	WRN-40-1 71.98	SHN-V2 71.28	MN-V2 64.60
KD(Baseline)	73.33	73.54	74.45	67.35
$k = 0.1$	73.13	73.75	75.47	68.9
$k = 0.5$	74.36	74.17	75.73	68.87
$k = 1$	74.74	74.49	76.58	70.45
$k = 2$	74.79	**74.87**	77.06	70.53
$k = 4$	75.56	74.48	**77.21**	**70.81**
$k = 6$	**75.74**	74.11	76.11	70.32

regions. In the later stages, ranking loss does not interfere with KL divergence and ultimately reaches a better position. The left figure shows that with ranking loss, student achieves leading accuracy at all stages. This indicates that the addition of the ranking loss helps the model converge and achieve better generalization and performance.

Visualization of Loss Landscape. To further investigate the role of ranking loss in the distillation process, we visualized the loss landscapes (Li et al., 2018) of student models with and without the application of ranking loss during Knowledge Distillation (KD), as depicted in Fig. 5. It is evident that the student models distilled with ranking loss exhibit a markedly flatter loss landscape and fewer local optima compared to those without it. We hypothesize that during the optimization process, ranking loss can filter out certain local optima that, despite presenting a better overall loss performance, yield poorer classification outcomes. Consequently, ranking loss can effectively enhance the generalization performance of student models.

Top-k & Min-k Ranking. To further validate the beneficial knowledge present in smaller channels, we conducted comparative experiments using the top 10%, top 30%, top 50%, and min 10%, min 30%, min 50% channels. The experiments were performed across four combinations of homogeneous and heterogeneous teacher-student pairs, and the results are presented in Fig. 5. We observed that

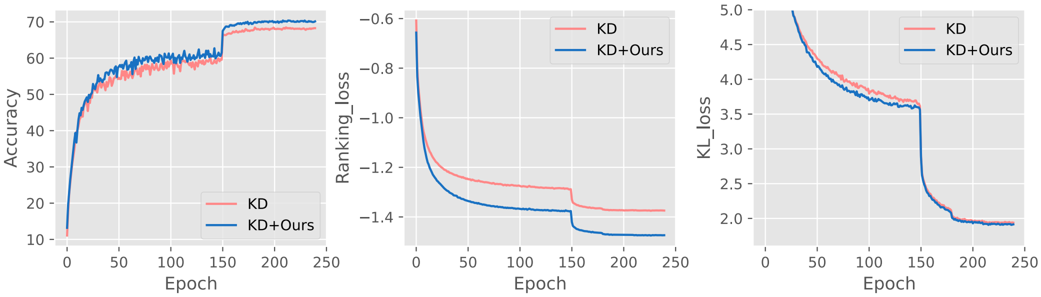


Figure 3: Accuracy & Loss Curves with Ranking Loss. **Left:** Top-1 Test Accuracy (%) Curve. **Middle:** Loss Curve of Ranking Loss. **Right:** Loss Curve of KL Divergence.

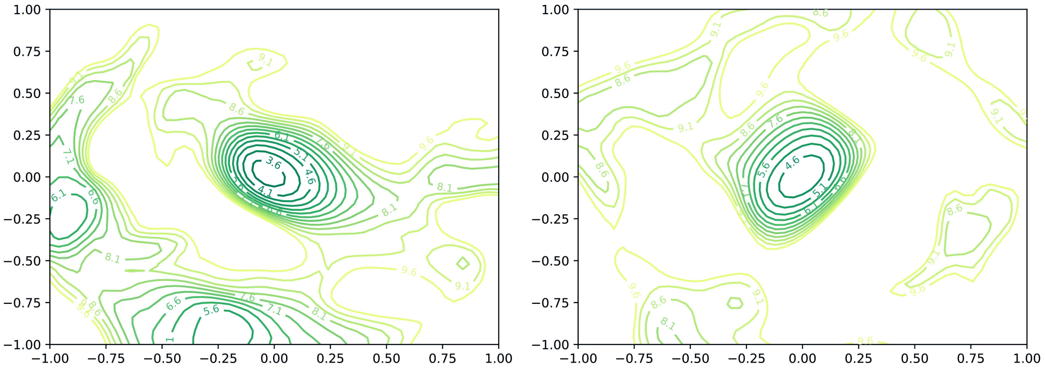
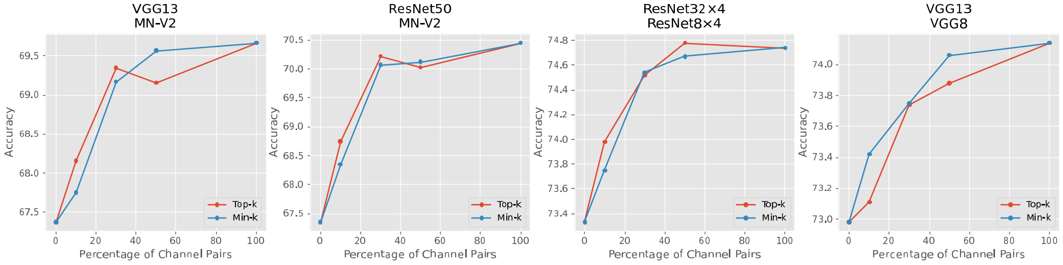


Figure 4: Loss Landscape. **Left:** The left landscape shows the suboptimal solutions in the distillation task. **Right:** After adding our ranking loss, the suboptimal solution is significantly reduced, as shown on the right.



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

Figure 5: Ranking Loss using Top-K and Min-K channels. Top-1 Accuracy (%) of Top-k/Min-k Ranking Loss. 0 and 100 on the x-axis represent KD and KD+Rank, respectively.

using the min-k channels achieves results similar to those obtained with top-k channels (as shown in Fig. 5(a),(b), and (c)). Additionally, in some cases, min-k channels provide even more beneficial information to aid student learning (as demonstrated in Fig. 5(d)), which indicates that the smaller channels also contain rich knowledge.

6 CONCLUSION

In this paper, we investigate the optimization process of logit distillation and identify that the Kullback-Leibler divergence tends to overlook the knowledge embedded in the smaller channels of the output. Moreover, KL divergence does not guarantee alignment between the classification results of the student and teacher models. To address this issue, we introduce a plug-and-play ranking loss based on Kendall's τ Coefficient that encourages the student model to pay more attention to the knowledge contained within the low-probability channels, while also enforcing alignment with the teacher's predictive outcomes. We provide a theoretical analysis demonstrating that the gradients of the ranking loss are less affected by channel scale and that its optimization objective is consistent with that of KL divergence, making it an effective auxiliary loss for distillation. Extensive experiments validate that our approach significantly enhances the distillation performance across various datasets and teacher-student architectures.

REPRODUCIBILITY STATEMENT

The details of datasets, model architectures, hyper-parameters, and evaluation metrics are described in subsection 5, the algorithm can be found in Appdenix A.5. Our code is attached to the Supplementary Material.

REFERENCES

Romero Adriana, Ballas Nicolas, K Samira Ebrahimi, Chassang Antoine, Gatta Carlo, and Bengio Yoshua. Fitnets: Hints for thin deep nets. *Proc. ICLR*, 2(3):1, 2015.

Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *Advances in Neural Information Processing Systems*, 35:15394–15406, 2022.

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*, pp. 11933–11942, 2022.

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*, pp. 5008–5017, 2021.

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.

Jaekeol Choi, Euna Jung, Jangwon Suh, and Wonjong Rhee. Improving bi-encoder document ranking models with two rankers and multi-teacher distillation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 2192–2196, 2021.

Luyu Gao, Zhuyun Dai, and Jamie Callan. Understanding bert rankers under distillation. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pp. 149–152, 2020.

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*, pp. 11868–11877, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

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.

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

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

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Gang Li, Xiang Li, Yujie Wang, Shanshan Zhang, Yichao Wu, and Ding Liang. Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1306–1313, 2022.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

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*, volume 37, pp. 1504–1512, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

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*, pp. 3967–3976, 2019.

Zhen Qin, Rolf Jagerman, Rama Kumar Pasumarthi, Honglei Zhuang, He Zhang, Aijun Bai, Kai Hui, Le Yan, and Xuanhui Wang. Rd-suite: A benchmark for ranking distillation. *Advances in Neural Information Processing Systems*, 36, 2023.

Sashank Reddi, Rama Kumar Pasumarthi, Aditya Menon, Ankit Singh Rawat, Felix Yu, Seungyeon Kim, Andreas Veit, and Sanjiv Kumar. Rankdistil: Knowledge distillation for ranking. In *International Conference on Artificial Intelligence and Statistics*, pp. 2368–2376. PMLR, 2021.

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*, 115:211–252, 2015.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. *arXiv preprint arXiv:2403.01427*, 2024.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. PMLR, 2013.

Jiaxi Tang and Ke Wang. Ranking distillation: Learning compact ranking models with high performance for recommender system. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2289–2298, 2018.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4933–4942, 2019.

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

Shuo Yang, Sujay Sanghavi, Holakou Rahmanian, Jan Bakus, and Vishwanathan SVN. Toward understanding privileged features distillation in learning-to-rank. *Advances in Neural Information Processing Systems*, 35:26658–26670, 2022a.

Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*, pp. 53–69. Springer, 2022b.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2021.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6848–6856, 2018.

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*, pp. 11953–11962, 2022.

Kaipeng Zheng, Huishuai Zhang, and Weiran Huang. Diffkendall: A novel approach for few-shot learning with differentiable kendall's rank correlation. *Advances in Neural Information Processing Systems*, 36:49403–49415, 2023.

A APPENDIX

A.1 IMPLEMENTATION DETAILS FOR TRANSFORMER

We use the AdamW optimizer and train for 300 epochs with an initial learning rate of 5e-4 and a weight decay of 0.05. The minimum learning rate is 5e-6, and the patch size is 16. We set α = 1, β = 1, γ = 0.5, and batch size is 128. The GPU we used is a single RTX4090.

A.2 ABLATION OF TEMPERATURE

There has been extensive and detailed research on the temperature of the KL loss (e.g., CTKD(Li et al., 2023), MLKD(Jin et al., 2023), LSKD(Sun et al., 2024)), achieving significant results. Unlike these studies, our approach aims to move away from focusing on the KL divergence and instead explore plug-and-play auxiliary losses to guide the KL divergence. Therefore, in our experiments, the temperature and other parameters of the KL divergence are kept the same as those in the respective baselines (with dynamic/multi-level temperatures used in CTKD+Ours and MLKD+Ours). The proposed ranking loss does not have a temperature parameter because the temperature does not affect the order of logits. Instead, the control over the distribution can be achieved by manipulating the steepness parameter k of the sign function, and ablation experiments for k are shown in Tab. 7. Meanwhile, although the temperature helps control the logit distribution, additional measures are needed to prevent the KL divergence from falling into suboptimal. We conducted an ablation study on the temperature, as shown in Tab. **??**. The results demonstrate that ranking loss as a plug and play loss can bring further improvements under multiple temperature settings.

Table 8: Ablation of Temperature. All experiments are conducted on CIFAR-100 with the teacher as ResNet32×4 and the student as ResNet8×4.

Tempetature	T = 4	T = 5	T = 6	T = 10
KD	73.33	73.39	73.43	73.55
KD+Ours	73.56(+0.23)	73.49(+0.10)	74.36(+0.0.93)	74.04(+0.49)

A.3 ABLATION OF NORMALIZATION

Due to the random initialization of student, the output logits will be too variant at the start of optimization and occasionally lead to gradient explosions. Therefore, we add normalization to the ranking loss to stabilize the ranking loss optimization at the very beginning. We also supplemented a set of small-scale ablation experiments that showed that the improvement in ranking performance did not come from the normalization added to the ranking loss, as shown in Tab. 9 below:

Table 9: Ablation of Normalization.

Teacher	ResNet32×4 79.42	WRN-40-2 75.61
Student	SHN-V1 70.30	WRN-40-1 71.98
KD	74.07	73.54
KD+Ours w/o Norm	76.38(+2.31)	74.07(+0.53)
KD+Ours w Norm	75.98(+1.91)	74.49(+0.95)

A.4 DIFFERENT FORMS OF RANKING LOSS

In the initial application of the ranking loss(Zheng et al., 2023), it is necessary to compute the gradients for two input vectors separately. However, In the scenario of distillation, the soft labels from the teacher model do not require gradients. Therefore, we propose three variants of the diff-Kendall ranking loss suitable for distillation scenarios, aimed at further exploring the role of ranking loss in distillation. Since the sample pair $(z_i^t - z_j^t)$ itself has a sign,, we can derive an equivalent form from Eq.6:

$$\tau = \frac{\sum_i \sum_{j<i} sgn((z_i^t - z_j^t) \cdot (z_i^s - z_j^s))}{\frac{1}{2}C(C-1)} \tag{14}$$

For Eq.6 and Eq.14, we can similarly transform them into differential forms of ranking loss by replacing sgn with $tanh$, and gradients are not computed for the output part corresponding to the teacher, which is:

$$L_\tau^{form1} = \frac{\sum_i \sum_{j<i} tanh(z_i^t - z_j^t)_{detach} \cdot tanh(z_i^s - z_j^s)}{\frac{1}{2}C(C-1)} \tag{15}$$

$$L_\tau^{form2} = \frac{\sum_i \sum_{j<i} tanh[(z_i^t - z_j^t)_{detach} \cdot (z_i^s - z_j^s)]}{\frac{1}{2}C(C-1)} \tag{16}$$

Where $()_{detach}$ means not participating in training. Further, since the gradient of the teacher's output is not required, we can directly use the sign function instead of the $tanh$ function to obtain the ranking loss:

$$L_\tau^{form3} = \frac{\sum_i \sum_{j<i} sgn(z_i^t - z_j^t)_{detach} \cdot tanh(z_i^s - z_j^s)}{\frac{1}{2}C(C-1)} \tag{17}$$

Different Forms of Ranking Loss. In our investigation, we examined the performance of the three forms of distillation presented in Sec. A.4, as depicted in Tab. 10. Form1 exhibited the most superior performance, followed by Form2, with Form3 trailing yet still enhancing the efficacy of the original knowledge distillation. This suggests that within the ranking loss, it is imperative to optimize considering the magnitude of the teacher's logits differences as a coefficient for the loss, rather than merely optimizing as a sign function.

Table 10: Different Forms of Ranking Loss. The experiments are conducted on the CIFAR-100, with 9 heterogeneous and 7 homogeneous architectures. The average Top-1 accuracy (%) is reported.

Loss Form	KD	KD+Form1	KD+Form2	KD+Form3
Similar Structure	72.01	**73.47(+1.46)**	73.42(+1.41)	73.38(+1.37)
Different Structure	72.74	**73.50(+0.76)**	73.36(+0.62)	73.05(+0.31)

A.5 ALGORITHM

Algorithm 1: Plug-and-Play Ranking Loss for Logit Distillation

```
Input: Transfer set $\mathcal{D}$ with samples of image-label pair $\{x_n, y_n\}_{n=1}^N$, base temperature $T$, teacher $f_t$,
  student $f_s$, knowledge distillation Loss $\mathcal{L}_{\text{KD}}$, ranking Loss $\mathcal{L}_{\text{RK}}$, the weight of ranking Loss $\gamma$.
Output: Trained student model $f_s$
for $(x_n, y_n)$ in $\mathcal{D}$ do
    Get the logits of Teacher and student:$z^t = f_t(x_n)$, $z^s = f_s(x_n)$
    Calculate the probability with temperature: $q^t = softmax(\frac{z^t}{T})$ , $q^s = softmax(\frac{z^s}{T})$
    Get the normalized logits of Teacher and student: $\hat{z}^t = \frac{z^t - \bar{z^t}}{std(z^t)}$, $\hat{z}^s = \frac{z^s - \bar{z^s}}{std(z^s)}$
    Update $f_s$ towards minimizing: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{KD}}(q^t, q^s) + \gamma \cdot \mathcal{L}_{\text{RK}}(\hat{z}^t, \hat{z}^s)$
end
```

A.6 DERIVATION OF THE KL LOSS AND RANKING LOSS

Derivation of the KL Divergence with Respect to Student Logits in Knowledge Distillation.

Denote the teacher's logits as $\mathbf{z}^t = [z_1^t, z_2^t, \ldots, z_C^t]$.
Denote the student's logits as $\mathbf{z}^s = [z_1^s, z_2^s, \ldots, z_C^s]$.
Let T be the temperature scaling factor used in the softmax function.

Teacher probabilities can be calculated as:

$$q_i^t = \frac{\exp\left(\frac{z_i^t}{T}\right)}{\sum_{j=1}^{C} \exp\left(\frac{z_j^t}{T}\right)}, \quad \text{for } i = 1, 2, \ldots, C. \tag{18}$$

Student probabilities can be calculated as:

$$q_i^s = \frac{\exp\left(\frac{z_i^s}{T}\right)}{\sum_{j=1}^{C} \exp\left(\frac{z_j^s}{T}\right)}, \quad \text{for } i = 1, 2, \ldots, C. \tag{19}$$

The loss function used in knowledge distillation is the scaled Kullback-Leibler (KL) divergence between the teacher and student probability distributions:

$$L_{\text{KD}} = T^2 \cdot \text{KL}\left(q^t \parallel q^s\right) = T^2 \sum_{i=1}^{C} q_i^t \log\left(\frac{q_i^t}{q_i^s}\right). \tag{20}$$

The derivative of the loss with respect to the student logits z_i^s is:

$$\frac{\partial L_{\text{KD}}}{\partial z_i^s} = -T^2 \sum_{k=1}^{C} q_k^t \frac{\partial \log q_k^s}{\partial z_i^s}. \tag{21}$$

Taking the natural logarithm:

$$\log q_k^s = \frac{z_k^s}{T} - \log\left(\sum_{j=1}^{C} \exp\left(\frac{z_j^s}{T}\right)\right). \tag{22}$$

Compute the partial derivative:

$$\frac{\partial \log q_k^s}{\partial z_i^s} = \frac{\partial}{\partial z_i^s}\left(\frac{z_k^s}{T}\right) - \frac{\partial}{\partial z_i^s}\left(\log\left(\sum_{j=1}^{C} \exp\left(\frac{z_j^s}{T}\right)\right)\right). \tag{23}$$

Compute each term separately:

$$\frac{\partial}{\partial z_i^s}\left(\frac{z_k^s}{T}\right) = \frac{1}{T}\delta_{ik}, \quad \delta_{ik} = \begin{cases} 1, & \text{if } i = k, \\ 0, & \text{if } i \neq k. \end{cases} \tag{24}$$

$$\frac{\partial}{\partial z_i^s}\left(\log\left(\sum_{j=1}^{C} \exp\left(\frac{z_j^s}{T}\right)\right)\right) = \frac{1}{\sum_{j=1}^{C} \exp\left(\frac{z_j^s}{T}\right)} \cdot \frac{\partial}{\partial z_i^s}\left(\sum_{j=1}^{C} \exp\left(\frac{z_j^s}{T}\right)\right). \tag{25}$$

The derivative inside the sum is:

$$\frac{\partial}{\partial z_i^s}\left(\sum_{j=1}^{C} \exp\left(\frac{z_j^s}{T}\right)\right) = \exp\left(\frac{z_i^s}{T}\right) \cdot \frac{1}{T} = \frac{\exp\left(\frac{z_i^s}{T}\right)}{T}. \tag{26}$$

Therefore, the second term becomes:

$$\frac{1}{\sum_{j=1}^{C} \exp\left(\frac{z_j^s}{T}\right)} \cdot \frac{\exp\left(\frac{z_i^s}{T}\right)}{T} = \frac{1}{T} q_i^s. \tag{27}$$

Thus, the total derivative is:

$$\frac{\partial \log q_k^s}{\partial z_i^s} = \frac{1}{T}\delta_{ik} - \frac{1}{T}q_i^s = \frac{1}{T}(\delta_{ik} - q_i^s). \tag{28}$$

We now substitute $\frac{\partial \log q_k^s}{\partial z_i^s}$ back into the expression for the derivative of the loss:

$$\begin{aligned} \frac{\partial L_{\text{KD}}}{\partial z_i^s} &= -T^2 \sum_{k=1}^{C} q_k^t \left(\frac{1}{T}(\delta_{ik} - q_i^s) \right) \\ &= -T \sum_{k=1}^{C} q_k^t (\delta_{ik} - q_i^s). \end{aligned} \tag{29}$$

Considering:

$$\sum_{k=1}^{C} q_k^t \delta_{ik} = q_i^t, \tag{30}$$

$$\sum_{k=1}^{C} q_k^t q_i^s = q_i^s \sum_{k=1}^{C} q_k^t = q_i^s \cdot 1 = q_i^s, \quad \sum_{k=1}^{C} q_k^t = 1. \tag{31}$$

Therefore, the loss derivative simplifies to:

$$\frac{\partial L_{\text{KD}}}{\partial z_i^s} = -T\left(q_i^t - q_i^s\right). \tag{32}$$

Derivation of the Ranking Loss with Respect to Student Logits in Knowledge Distillation. The Rank loss is calculated as:

$$L_{RK} = -\frac{\sum_i \sum_{j<i} tanh(k(z_i^t - z_j^t)) \cdot tanh(k(z_i^s - z_j^s))}{\frac{C(C-1)}{2}} \tag{33}$$

The derivation is calculated as:

$$\frac{\partial L_{RK}}{\partial z_i^s} = \frac{1}{\frac{C(C-1)}{2}} \sum_{j \neq i} \frac{\partial}{\partial z_i^s} \frac{1}{2} \cdot \left[\tanh\left(k(z_i^s - z_j^s)\right) \tanh\left(k(z_i^t - z_j^t)\right)\right] \tag{34}$$

$$= \frac{1}{C(C-1)} \sum_{j \neq i} \frac{\partial}{\partial z_i^s} \left[\tanh\left(k(z_i^s - z_j^s)\right) \tanh\left(k(z_i^t - z_j^t)\right)\right] \tag{35}$$

Denote ϕ_{ij} as:

$$\phi_{ij} = \tanh\left(k(z_i^s - z_j^s)\right) \tanh\left(k(z_i^t - z_j^t)\right) \tag{36}$$

Its derivative w.r.t. z_i^s is:

$$\frac{\partial \phi_{ij}}{\partial z_i^s} = \left[1 - \tanh^2\left(k(z_i^s - z_j^s)\right)\right] \cdot k \cdot \tanh\left(k(z_i^t - z_j^t)\right) \tag{37}$$

Finally, the gradient of L_{RK} with respect to z_i^s is:

$$\frac{\partial L_{RK}}{\partial z_i^s} = -\frac{k}{C(C-1)} \sum_{j \neq i} \left[1 - \tanh^2\left(k(z_i^s - z_j^s)\right)\right] \tanh\left(k(z_i^t - z_j^t)\right) \tag{38}$$

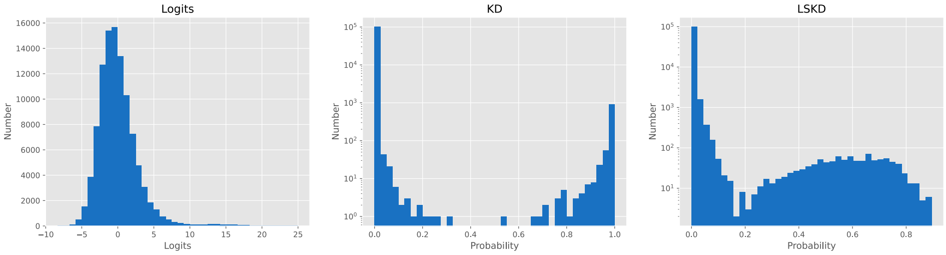


Figure 6: probability produced by KD and LSKD

A.7 Derivation of the Differential Ranking Loss

In this section, we explain how we get the final form of Eq.7. Noticed that:

$$tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1} = 1 - \frac{2}{e^{2x} + 1} \tag{39}$$

Then, we can get the final form of Eq.7 by expanding the $tanh$ function.

$$\tau_d = \frac{\sum_i \sum_{j<i} tanh(k \cdot (z_i^t - z_j^t)) \cdot tanh(k \cdot (z_i^s - z_j^s))}{\frac{1}{2}C(C-1)} \tag{40}$$

$$= \frac{2}{C(C-1)} \cdot \sum_i \sum_{j<i} \left(1 - \frac{2}{1 + e^{2 \cdot (z_i^t - z_j^t) \cdot k}}\right) \cdot \left(1 - \frac{2}{1 + e^{2 \cdot (z_i^s - z_j^s) \cdot k}}\right) \tag{41}$$

A.8 Does Temperature Solve the Problem of Ignoring Smaller Channels

Some methods address the issue of neglecting smaller channels by employing temperature scaling. We examined the probability distribution of LSKD (Sun et al., 2024), an approach using adaptive temperature that has shown promising results. As shown in Fig.6, while LSKD somewhat alleviates the problem of smaller channels, the transformed probabilities still predominantly occupy these smaller channels. Thus, using temperature scaling does not effectively resolve the issue of neglecting smaller channels.

A.9 Further visualization of the Gradient and Ranking

In our study, we conducted a visualization of the gradients associated with both the ranking loss and the KL divergence loss in Fig.7. The logits of the student model were randomly generated but

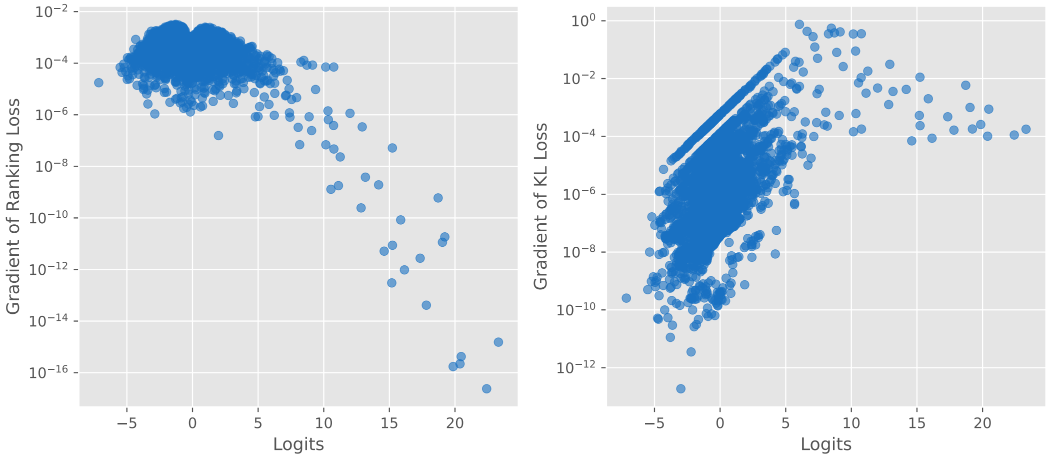


Figure 7: The gradient of ranking loss and KL loss for different values of logits

maintained the same scale as those of the teacher model across each channel. Our findings indicate that the ranking loss consistently provides gradients of similar scale across varying logit values, whereas the gradients from the KL loss are heavily dependent on the specific logit values. This suggests that the ranking loss offers a more uniform attention distribution across different logits compared to the KL loss. **Notably, the ranking loss assigns smaller gradients to logits with larger values, as they are the classification targets and reach the correct rank.** Consequently, these logits receive less attention from the ranking loss.

In addition, we present a visualization of the ranking results generated by KD and KD+Ours in Fig.8. The results indicate that our approach consistently enhances channel alignment, underscoring the robustness and general applicability of our method, particularly in its focused attention on smaller channels.

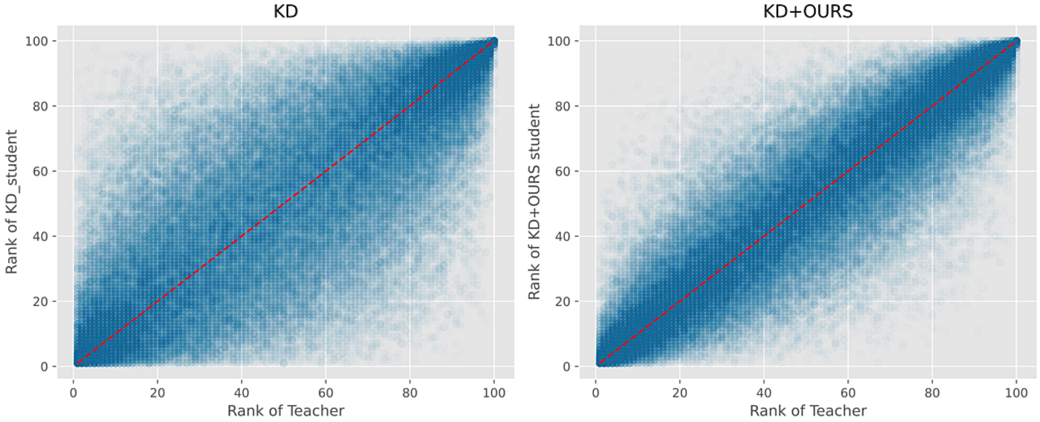


Figure 8: The rank of teacher and student trained by KD and KD + Ours

A.10 Further Ablation Study of the Hyperparameters

To further substantiate the generalization capability and robustness of our method, we conducted comprehensive ablation studies using different combinations of coefficients. As illustrated in Figure 9, our method consistently maintains high performance across various settings, underscoring the strong generalization ability and universal applicability of our method.

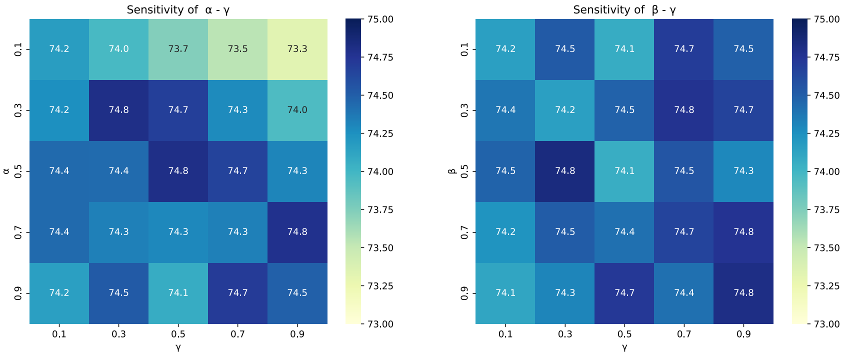


Figure 9: Sensitivity Analysis. **Left:** Sensitivity of $\alpha - \gamma$. **Right:** Sensitivity of $\beta - \gamma$