Knowledge Distillation with Refined Logits

Wujie Sun[1,2,3] Defang Chen[4†] Siwei Lyu[4] Genlang Chen[5] Chun Chen[1,3] Can Wang[1,3]
[1]State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2]School of Software Technology, Zhejiang University
[3]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[4]University at Buffalo, State University of New York [5]NingboTech University
sunwujie@zju.edu.cn, {defangch, siweilyu}@buffalo.edu, {cgl, chenc, wcan}@zju.edu.cn

Abstract

Recent research on knowledge distillation has increasingly focused on logit distillation because of its simplicity, effectiveness, and versatility in model compression. In this paper, we introduce Refined Logit Distillation (RLD) to address the limitations of current logit distillation methods. Our approach is motivated by the observation that even high-performing teacher models can make incorrect predictions, creating an exacerbated divergence between the standard distillation loss and the cross-entropy loss, which can undermine the consistency of the student model's learning objectives. Previous attempts to use labels to empirically correct teacher predictions may undermine the class correlations. In contrast, our RLD employs labeling information to dynamically refine teacher logits. In this way, our method can effectively eliminate misleading information from the teacher while preserving crucial class correlations, thus enhancing the value and efficiency of distilled knowledge. Experimental results on CIFAR-100 and ImageNet demonstrate its superiority over existing methods. Our code is available at https://github.com/zju-SWJ/RLD.

1. Introduction

Knowledge distillation utilizes pre-trained high-performing teacher models to facilitate the training of a compact student model [12]. Compared to other model compression methods, such as pruning and quantization [5], knowledge distillation exhibits fewer constraints on the model architecture. This flexibility significantly broadens its applicability, contributing to its increasing prominence in recent research.

Hinton *et al*. [12] were the first to introduce the concept of logit distillation. It is designed to align the logits of teacher and student models following the softmax operations with the Kullback-Leibler (KL) divergence. Most subsequent research has maintained the original concept of logit distillation, instead focusing on exploring feature distillation [3, 4, 18, 28, 37, 38] in more depth by selecting and aligning intermediate-level features between teacher and student models. However, the potential architectural disparity between teacher and student models poses a significant challenge for feature alignment. This is mainly due to the fact that different architectures extract different features [38]. Moreover, the extensive diversity in feature selection further amplifies the complexity of feature distillation and leads to an increase in training time in distillation [3]. Recently, by decoupling the classical logit distillation loss, Zhao *et al*. [48] demonstrate that logit distillation can yield results that are on par with, or even superior to, those of feature distillation. Consequently, logit distillation garnered considerable attention in the research community, thanks to its simplicity, effectiveness, and versatility.

Despite the impressive achievements, most of the recent logit distillation approaches [15, 21, 34] overlook the impact of teacher prediction correctness on the training process. Specifically, incorrect teacher predictions lead to an exacerbated divergence between teacher loss and label loss, which may severely impede the potential enhancements of the student models. Existing correction-based distillation approaches [1, 20, 40] consistently modify the teacher logits (target) using label information. They either exchange the values between the predicted maximum class and the true class [40] (the *swap* operation) or amplify the proportion of the true class within the predicted probabilities [1, 20] (the *augment* operation). We argue that such approaches may alter the correlations among classes, as exemplified in Figure 1. This disruption can obstruct the transmission of "dark knowledge" [12] and hinder performance improvements.

In this paper, we introduce *Refined Logit Distillation* (RLD) to address these challenges. In classification tasks, the true class probability dictates prediction correctness, while class correlations capture high-level semantic relationships that influence classification tendencies. Accord-

†Corresponding Author

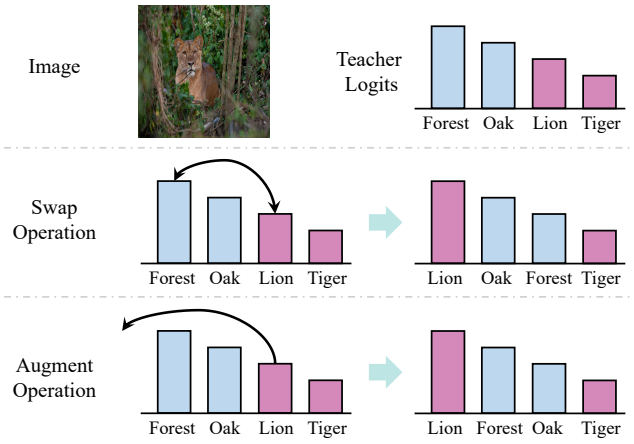


Figure 1. A toy example of existing correction-based distillation approaches. **Classes represented by the same color are highly correlated and should be ranked closely.** The image displayed is a "lion", yet the teacher model incorrectly classifies it as the "forest". Both the swap and augment operations disrupt the close correlation between "lion" and "tiger". A more detailed example of class correlation is provided in the supplementary material.

ingly, RLD consists of two types of knowledge, *sample confidence* (SC) and *masked correlation* (MC). Sample confidence refers to the binary probabilities derived from logits. As for the teacher model, SC comes from the probability associated with the predicted class and the probabilities of the remaining classes. It encapsulates the teacher's prediction confidence for the current sample and is employed to guide the student model. Considering the possible inaccuracies in the teacher's prediction, we align the student's true class probability with teacher's predicted class probability. This alignment not only mitigates the teacher's mistakes, but also guides the student model toward achieving a comparable level of confidence for the current sample. Moreover, it effectively prevents over-fitting. Masked correlation denotes our dynamic approach for selecting a subset of classes for teacher-student alignment. It is designed to mitigate the influence of potentially incorrect teacher predictions on student models while conveying essential class correlations. More specifically, MC involves masking all classes within the teacher logits that have equal or superior rankings compared to the true class. In essence, fewer classes are used for distillation when the teacher makes more mistakes, and more classes are used when it makes fewer mistakes. Using these two complementary types of refined knowledge, the student can achieve better performance.

Our contributions are summarized as follows:

- We reveal that prevalent distillation approaches fail to account for the effects of incorrect teacher predictions, and existing correction-based strategies tend to ruin the valuable class correlations.
- We introduce a novel logit distillation approach termed Refined Logit Distillation (RLD) to prevent over-fitting and mitigate the influence of incorrect teacher knowledge, while preserving the essential class correlations.
- We conduct comprehensive experiments on CIFAR-100 and ImageNet datasets to verify the superior performance of our proposed RLD method.

2. Related Work

The application of knowledge distillation historically concentrated on the image classification task, and progressively extended to a wider range of tasks, including semantic segmentation [23, 32, 36, 42] and image generation [25, 30, 35] within the realm of computer vision. Traditional knowledge distillation typically involves a single teacher and a single student model. As the field evolves, a variety of other paradigms have been proposed, such as online distillation [2, 41], multi-teacher distillation [43, 46], and self-distillation [7, 17]. Since traditional knowledge distillation remains the core foundation of research in this area, we will focus solely on such methods in the discussion below.

In image classification task, existing algorithms can be broadly classified into three categories: logit distillation [12, 15, 21, 34, 48], feature distillation [3, 4, 11, 16, 18, 37, 38], and relation distillation [22, 26, 27]. Logit distillation has become the main focus of current research because of its straightforwardness, effectiveness, and adaptability. The initial logit distillation [12] leverages KL divergence to align the softened output logits of the teacher and student models, thereby significantly enhancing the performance of the student models. DKD [48] revitalizes logit distillation by decoupling this classical loss, enabling it to perform comparably to feature distillation. MLKD [15] leverages multi-level logit knowledge to further enhance model performance. CTKD [21] introduces the curriculum temperature, applying adversarial training and curriculum learning to dynamically determine the distillation temperature for each sample. LSKD [34] processes the logits to adaptively allocate temperatures between teacher and student and across samples, thereby achieving state-of-the-art performance. However, the effect of incorrect teacher predictions on distillation is rarely considered.

Given that logits are intrinsically related to prediction correctness, several methods leverage labels to adjust logits prior to the distillation process. LA [40] swaps the values of the true and predicted classes to correct the teacher model's predictions. RC [1] adds the maximum value in the student's output to the true class, thereby aiding the student model in making accurate and confident predictions. LR [20] combines one-hot labels with the teacher's soft labels to produce a new, precise target for distillation. However, as previously demonstrated in Figure 1, these methods may disrupt class correlations, which can hinder perfor-

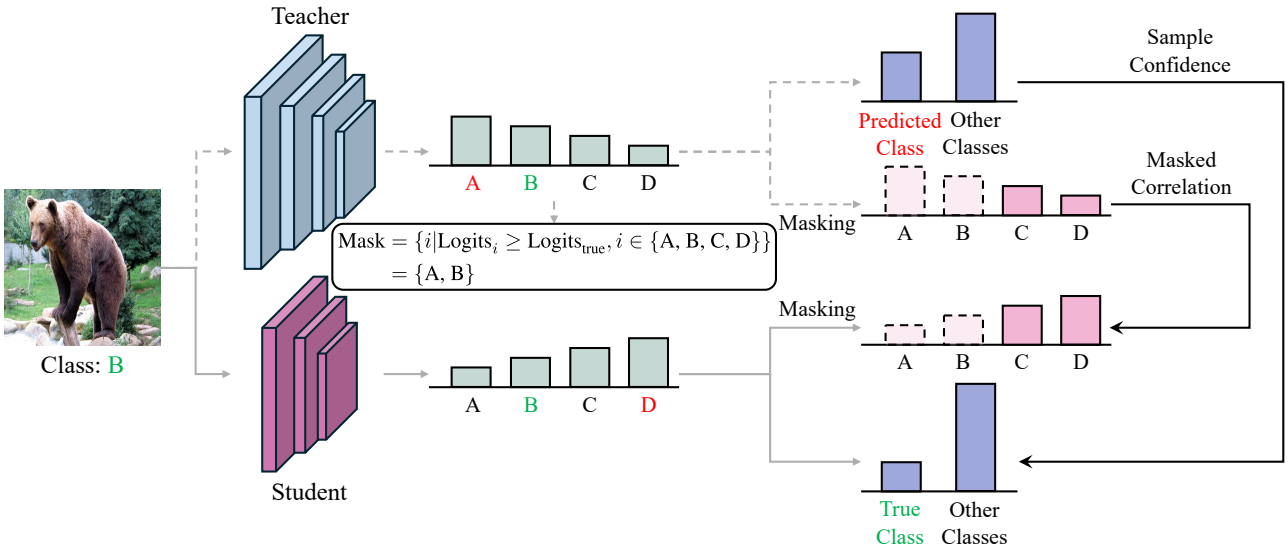


Figure 2. An overview of our proposed Refined Logit Distillation (RLD). In RLD, the teacher model imparts two types of knowledge, denoted as "sample confidence" and "masked correlation", to the student model. The binary sample confidence encapsulates the model confidence for each sample, which helps the student model generate proper-confidence predictions for the true class. The masked correlation denotes the probability distribution acquired after dynamically masking certain classes, which helps to remove misleading information and preserve valuable class correlations during distillation. Both kinds of knowledge are obtained from logits, thus the distillation process does not introduce intermediate layer features.

mance improvement.

3. Preliminaries

We provide an overview of concepts related to knowledge distillation to facilitate readers' understanding.

Consider an image classification task involving C classes. We have a pre-trained teacher model and a student model, denoted as θ^{T} and θ^{S}, respectively. For a single input image x, the output logits z from teacher and student models are denoted as $z^{\mathrm{T}} = \theta^{\mathrm{T}}(x)$ and $z^{\mathrm{S}} = \theta^{\mathrm{S}}(x)$, respectively. By utilizing the softmax function $\sigma(\cdot)$, predicted distributions p^{T} and p^{S} are calculated as follows:

$$p_i = \frac{\exp(z_i)}{\sum_{c=1}^{C} \exp(z_c)}, \tag{1}$$

where p_i represents the predicted value of the i-th class.

To train the student model, the first loss is computed as the cross entropy between the student prediction and the one-hot ground-truth label y:

$$L_{\mathrm{CE}} = -\sum_{c=1}^{C} y_c \log p_c^{\mathrm{S}}. \tag{2}$$

The second loss aligns the softened predictions $\hat{p} = \sigma(z/\tau)$ of the teacher and student models using the KL divergence:

$$L_{\mathrm{KD}} = \tau^2 \mathrm{KL}\left(\hat{p}^{\mathrm{T}}, \hat{p}^{\mathrm{S}}\right) = \tau^2 \sum_{c=1}^{C} \hat{p}_c^{\mathrm{T}} \log \frac{\hat{p}_c^{\mathrm{T}}}{\hat{p}_c^{\mathrm{S}}}, \tag{3}$$

where τ denotes the temperature for the softmax operation.

By combining Equations (2) and (3), we get the classical logit distillation loss for stochastic gradient descent. Such an approach has been experimentally shown to perform better than training solely with labels.

4. Methodology

In this section, we delve into a detailed introduction of our proposed RLD. An overview of RLD is shown in Figure 2.

4.1. Sample Confidence Distillation

Sample confidence (SC) represents the binary distribution b derived from the logits. It encapsulates the model confidence for each sample, thereby aiding the student model in generating proper-confidence predictions for the true class, without unduly restricting the distribution for other classes.

In the context of teacher knowledge, one component of the SC is the maximum predicted probability value $\hat{p}_{\mathrm{max}}^{\mathrm{T}}$, while the other component is the sum of the predicted probabilities for the remaining classes. In contrast, the student SC consists of two components: the predicted probability for the true class $\hat{p}_{\mathrm{true}}^{\mathrm{S}}$, and the sum of the predicted probabilities for the remaining classes. They can be summarized

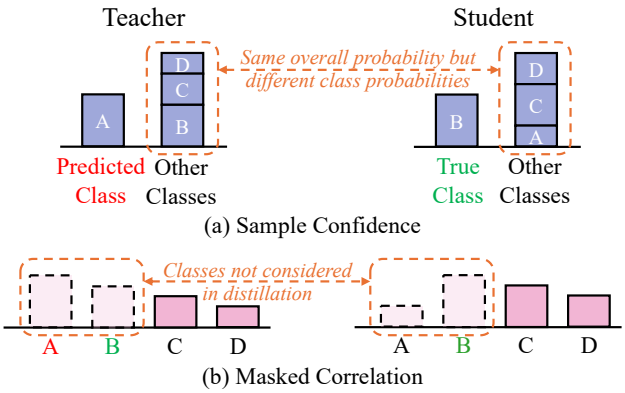


Figure 3. Toy examples elucidating the variances across methods **when the corresponding losses are minimal**. (a) Sample confidence guarantees a proper probability for the true class and eliminates the necessity for the student model to match the intact class probability distribution to that of the teacher model. (b) Masked correlation imposes fewer constraints on the student prediction than traditional knowledge distillation. For the classes that are masked, their probabilities can diverge entirely from those of the teacher model.

in the following formulas:

$$b^{\mathrm{T}} = \{\hat{p}^{\mathrm{T}}_{\max}, 1 - \hat{p}^{\mathrm{T}}_{\max}\}, \tag{4}$$

$$b^{\mathrm{S}} = \{\hat{p}^{\mathrm{S}}_{\mathrm{true}}, 1 - \hat{p}^{\mathrm{S}}_{\mathrm{true}}\}. \tag{5}$$

To transfer this knowledge, we align b^{T} and b^{S} using the KL divergence:

$$L_{\mathrm{SCD}} = \tau^2 \mathrm{KL}\left(b^{\mathrm{T}}, b^{\mathrm{S}}\right). \tag{6}$$

Figure 3(a) more vividly illustrates the aligned knowledge of the teacher and the student when using SCD.

While both L_{CE} and L_{SCD} operate on the true class, gradient analysis (temperature τ omitted) shows that their effects are not entirely identical:

$$\begin{aligned} \frac{\partial L_{\mathrm{CE}}}{\partial z_i} &= p^{\mathrm{S}}_i - y_i, \\ \frac{\partial L_{\mathrm{SCD}}}{\partial z_i} &= \begin{cases} p^{\mathrm{S}}_i - p^{\mathrm{T}}_{\max}, & i = \text{true}, \\ \frac{p^{\mathrm{S}}_i (p^{\mathrm{S}}_{\mathrm{true}} - p^{\mathrm{T}}_{\max})}{p^{\mathrm{S}}_{\mathrm{true}} - 1}, & i \neq \text{true}. \end{cases} \end{aligned} \tag{7}$$

Consequently, RLD incorporates both, thereby endowing the distillation process with greater flexibility.

4.2. Masked Correlation Distillation

Masked correlation (MC) denotes the probability distribution acquired after dynamically masking certain classes. As shown in Figure 3(b), this masking operation relieves the student model from aligning incorrect class rankings, thereby allowing the student model to generate very different output from the teacher without incurring a large loss. Moreover, preserving partial class probabilities empowers the student model to learn valuable class correlations, consequently enhancing the model's performance.

Specifically, the mask M is dynamically derived from teacher logits and labels. We designate all classes whose logit values are **g**reater than or **e**qual to (denoted as "ge") the logit value of the true class as the targets for the masking operation, which can be represented as follows:

$$M_{\mathrm{ge}} = \{i | z^{\mathrm{T}}_i \geq z^{\mathrm{T}}_{\mathrm{true}}, 1 \leq i \leq C\}. \tag{8}$$

After obtaining the mask, we compute the probability distributions for alignment using the following formula:

$$\tilde{p}_i = \frac{\exp\left(z_i/\tau\right)}{\sum_{c=1, c \notin M_{\mathrm{ge}}}^{C} \exp\left(z_c/\tau\right)}, \tag{9}$$

where $1 \leq i \leq C$ and $i \notin M_{\mathrm{ge}}$ are satisfied.

We summarize the distillation loss for the masked correlation knowledge as follows:

$$L_{\mathrm{MCD}} = \tau^2 \mathrm{KL}\left(\tilde{p}^{\mathrm{T}}, \tilde{p}^{\mathrm{S}}\right). \tag{10}$$

When the teacher model makes a more accurate prediction (ranking the true class higher), only a few classes are subjected to the masking operation. It allows the majority of class correlations to be preserved and transferred to the student model. Conversely, if the teacher's prediction is less accurate, the majority of classes are masked. As a result, the student model learns less knowledge, thereby reducing the potential for misinformation to mislead the training process. It also gives the student model more freedom to make predictions for masked classes that differ significantly from those of the teacher model.

4.3. Refined Logit Distillation

By combining Equations (2), (6) and (10), we obtain the final loss for RLD, which is:

$$L_{\mathrm{RLD}} = L_{\mathrm{CE}} + \alpha L_{\mathrm{SCD}} + \beta L_{\mathrm{MCD}}, \tag{11}$$

where hyper-parameters α and β adjust the importance of sample confidence and masked correlation, respectively.

Relevance to DKD. Although RLD and DKD [48] consider logit distillation from distinct perspectives, they become equivalent when the teacher model consistently makes accurate predictions. Besides, DKD does not explicitly explain why transferring non-target class knowledge (i.e., the probability distribution when the true class is masked, referred to as NCKD) can significantly enhance model performance. Beyond the idea that the class relationships embedded in this knowledge facilitate training, RLD offers a new

Type							
	Teacher	ResNet32×4 79.42	VGG13 74.64	WRN-40-2 75.61	ResNet56 72.34	ResNet110 74.31	ResNet110 74.31
	Student	ResNet8×4 72.50	VGG8 70.36	WRN-40-1 71.98	ResNet20 69.06	ResNet32 71.14	ResNet20 69.06
Feature	FitNet	73.50	71.02	72.24	69.21	71.06	68.99
	AT	73.44	71.43	72.77	70.55	72.31	70.65
	RKD	71.90	71.48	72.22	69.61	71.82	69.25
	CRD	75.51	73.94	74.14	71.16	73.48	71.46
	OFD	74.95	73.95	74.33	70.98	73.23	71.29
	ReviewKD	75.63	74.84	*75.09*	71.89	73.89	71.34
	SimKD	*78.08*	74.89	74.53	71.05	73.92	71.06
	CAT-KD	76.91	74.65	74.82	71.62	73.62	71.37
Logit	KD	73.33	72.98	73.54	70.66	73.08	70.67
	CTKD	73.39	73.52	73.93	71.19	73.52	70.99
	DKD	<u>76.32</u>	<u>74.68</u>	<u>74.81</u>	<u>71.97</u>	**74.11**	71.06
	LA	73.46	73.51	73.75	71.24	73.39	70.86
	RC	74.68	73.37	74.07	71.63	73.44	<u>71.41</u>
	LR	76.06	74.66	74.42	70.74	73.52	70.61
	RLD (ours)	**76.64**	**74.93**	**74.88**	**72.00**	<u>74.02</u>	**71.67**

Table 1. Top-1 accuracy (%) on the CIFAR-100 validation set when the teacher and student models are homogeneous. The best and second best results of logit distillation are highlighted in **bold** and <u>underlined</u> text, respectively. For the case where the best result of feature distillation is better than the best result of logit distillation, we highlight it with *italic* text. The reported results are the mean of three trials.

explanation: when the alignment constraint on target class knowledge (TCKD) is weak, masking the true class during distribution alignment offers the student model greater flexibility to adjust the ranking of the true class. This, in turn, mitigates the negative impact of incorrect teacher knowledge, enabling more accurate predictions.

5. Experiments

5.1. Settings

Datasets. We conduct the experiments on two standard image classification datasets: CIFAR-100 [19] and ImageNet [29]. CIFAR-100 comprises 100 distinct classes, with a total of 50,000 images in the training set and 10,000 images in the validation set. Each image in this dataset is of size 32×32 pixels. ImageNet presents a larger and more complex dataset, encompassing 1,000 classes. It includes 1.28 million images in the training set and 50,000 images in the validation set, with each image resolution being 224×224 pixels after pre-processing.

Models. Models used by teachers and students include ResNet [9], WideResNet (WRN) [45], VGG [33], ShuffleNet (SHN) [24, 47], and MobileNet (MN) [14, 31].

Compared Methods. We emphasize that the experimental results in this paper are presented after exhaustive reading of the papers and codes of existing works, and with a focus on fair comparisons between methods. Therefore, MLKD [15] is excluded from the comparison due to differing experimental settings. The compared methods include feature distillation (FitNet [28], AT [18], RKD [26], CRD [37], OFD [10], ReviewKD [4], SimKD [3], and CAT-KD [8]) and logit distillation (KD [12], CTKD [21], DKD [48], LA [40], RC [1], and LR [20]) methods. The performance metrics of all comparison methods, except for the latter three correction-based approaches (LA, RC, and LR), are sourced directly from LSKD [34]. To ensure experimental fairness, we implemented these three correction-based approaches and our proposed RLD method following the experimental setups commonly used in prominent studies (e.g., LSKD [34], DKD [48], and CRD [37]). Additional implementation details are provided in the supplementary material.

5.2. Main Results

CIFAR-100. The top-1 validation accuracy (%) comparison results of RLD and other distillation approaches are reported in Table 1 (homogeneous distillation pairs) and Table 2 (heterogeneous distillation pairs). We can see that RLD is either the optimal or suboptimal logit distillation algorithm in all cases, and is optimal in most cases. This underscores the superiority of RLD and accentuates the significance of making corrections to teacher predictions. While feature distillation can sometimes outperform logit distilla-

Type							
	Teacher	ResNet32×4 79.42	ResNet32×4 79.42	WRN-40-2 75.61	WRN-40-2 75.61	VGG13 74.64	ResNet50 79.34
	Student	SHN-V2 71.82	WRN-40-2 75.61	ResNet8×4 72.50	MN-V2 64.60	MN-V2 64.60	MN-V2 64.60
Feature	FitNet	73.54	77.69	74.61	68.64	64.16	63.16
	AT	72.73	77.43	74.11	60.78	59.40	58.58
	RKD	73.21	77.82	75.26	69.27	64.52	64.43
	CRD	75.65	78.15	75.24	70.28	69.73	69.11
	OFD	76.82	79.25	74.36	69.92	69.48	69.04
	ReviewKD	77.78	78.96	74.34	*71.28*	*70.37*	69.89
	SimKD	78.39	*79.29*	75.29	70.10	69.44	69.97
	CAT-KD	*78.41*	78.59	75.38	70.24	69.13	*71.36*
Logit	KD	74.45	77.70	73.97	68.36	67.37	67.35
	CTKD	75.37	77.66	74.61	68.34	68.50	68.67
	DKD	77.07	78.46	75.56	69.28	69.71	70.35
	LA	75.14	77.39	73.88	68.57	68.09	68.85
	RC	75.61	77.58	75.22	68.72	68.66	68.98
	LR	76.27	78.73	75.26	69.02	69.78	70.38
	RLD (ours)	**77.56**	**78.91**	**76.12**	**69.75**	**69.97**	**70.76**

Table 2. Top-1 accuracy (%) on the CIFAR-100 validation set when the teacher and student models are heterogeneous. The same convention is used as in Table 1.

Teacher/Student	Res34/Res18		Res50/MN-V1	
Accuracy	Top-1	Top-5	Top-1	Top-5
Teacher	73.31	91.42	76.16	92.86
Student	69.75	89.07	68.87	88.76
AT	70.69	90.01	69.56	89.33
OFD	70.81	89.98	71.25	90.34
CRD	71.17	90.13	71.37	90.41
ReviewKD	71.61	90.51	72.56	91.00
SimKD	71.59	90.48	72.25	90.86
CAT-KD	71.26	90.45	72.24	91.13
KD	71.03	90.05	70.50	89.80
CTKD	71.38	90.27	71.16	90.11
DKD	71.70	90.41	72.05	91.05
LA	71.17	90.16	70.98	90.13
RC	71.59	90.21	71.86	90.54
LR	70.29	89.98	71.76	90.93
RLD (ours)	**71.91**	**90.59**	**72.75**	**91.18**

Table 3. Top-1 and top-5 accuracy (%) on the ImageNet validation set. The best and second best results are highlighted in **bold** and underlined text, respectively. The reported results are the mean of three trials.

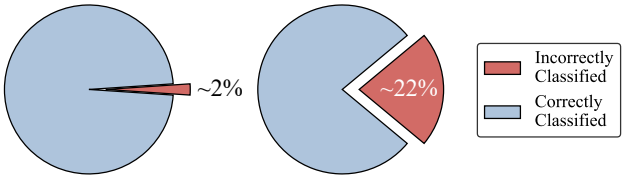
tion, its optimal method varies across teacher-student pairs, and its longer training time and complex algorithm design may hinder practical applicability.

ImageNet. The top-1 and top-5 validation accuracy (%) comparison results of RLD and other distillation approaches are reported in Table 3. On this more challenging dataset, RLD successfully outperforms all existing feature and logit distillation algorithms, consistently achieving optimal performance and demonstrating its superiority.

Analysis. Examining the experimental results detailed in Tables 1 to 3, it is clear that the performance improvement brought about by RLD is more substantial on the ImageNet dataset than on the CIFAR-100 dataset. We presume that this discrepancy stems from the varying accuracy levels of the teacher models on the respective training sets. As shown in Figure 4, the teacher model exhibits high training accuracy on the CIFAR-100 dataset, while it shows comparatively lower accuracy on the ImageNet dataset. Given that RLD aligns with DKD when the teacher predictions are accurate, the high training accuracy on the CIFAR-100 dataset might hinder substantial divergence between these two methods. Conversely, on the ImageNet dataset, where the training accuracy is lower, RLD outperforms DKD by achieving a more substantial improvement.

5.3. Extensions

Reversed Knowledge Distillation. We explore a unique scenario termed reversed knowledge distillation [44], where the teacher performs worse than the student. This study investigates the feasibility of using an inferior teacher model



(a) CIFAR-100 (b) ImageNet

Figure 4. Proportion of predictions from teacher models on the training set. (a) ResNet56. (b) ResNet50.

Teacher	ResNet56 72.34	ResNet110 74.31	VGG13 74.64
Student	WideResNet-40-2 75.61		
KD	76.72	77.37	76.86
DKD	77.34	77.70	77.45
RLD (ours)	**78.03**	**78.28**	**77.88**
Δ	+0.69	+0.58	+0.43

Table 4. Top-1 accuracy (%) on the CIFAR-100 validation set when distilling with inferior teachers. Optimal results are highlighted in **bold**. The reported results are the mean of three trials.

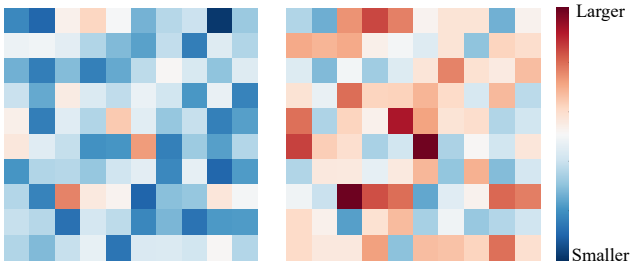
to enhance the performance of a superior student model, particularly in situations where sourcing a more capable teacher model proves challenging. As shown in Table 4, among all distillation pairs, RLD achieves the best performance, and shows a large performance gap compared to DKD. Moreover, the accuracy difference Δ between RLD and DKD shows that poorer teacher model performance leads to greater RLD improvement over DKD. This can be attributed to two main factors: firstly, the use of the inferior teacher allows RLD to be better distinguished from DKD; secondly, the unique setup of reversed knowledge distillation imposes more stringent demands on the quality of knowledge transferred, thereby underscoring the effectiveness of RLD in refining distilled knowledge.

Logit Standardization. We investigate the efficacy of each method when supplemented with logit standardization technique LSKD [34]. The results are shown in Table 5. The optimal results achieved by RLD underscore its superior performance and the vast potential of its integration with other methodologies.

Logit Discrepancy Visualization. We calculate the mean absolute error (MAE) of logits for each class between teacher and student models obtained via DKD and RLD, visualizing these results using the heat map in Figure 5. Despite RLD outperforming DKD, it is observed that the

Teacher	WRN-40-2 75.61	VGG13 74.64	ResNet50 79.34
Student	MobileNet-V2 64.60		
KD	69.23	68.61	69.02
CTKD	69.53	68.98	69.36
DKD	70.01	69.98	70.45
RLD (ours)	**70.35**	**70.63**	**71.06**

Table 5. Top-1 accuracy (%) on the CIFAR-100 validation set when training with logit standardization technique LSKD [34]. The optimal results are highlighted in **bold** text. The reported results are the mean of three trials.



(a) DKD (b) RLD

Figure 5. Visualized teacher-student logit discrepancy learned by DKD and RLD on the CIFAR-100 validation set. For better visualization, 100 classes are reshaped into a 10×10 matrix. The teacher is ResNet32×4, and the student is ResNet8×4.

logit discrepancy yielded by RLD is larger than that of DKD. This observation aligns with our anticipation, given that RLD rectifies certain inaccuracies in teacher knowledge and provides students with greater autonomy in formulating their own predictions. This finding underscores that an unconsidered alignment with teacher knowledge may not be the optimal strategy, and we believe that correction-based approaches deserve more attention and research.

Ablation Study. We perform an ablation study on the components of RLD, and the results are shown in Table 6. The results demonstrate that each component of RLD effectively contributes to enhanced performance. Notably, while masking all classes with values **g**reater than (denoted as "g") those of the true class would similarly eliminate misinformation and preserve class correlations, this masking strategy (denoted as M_{g}) inadvertently integrates true class-related knowledge into both masked correlation and sample confidence. This overlap may create conflicts between the resulting losses, thereby hindering performance improvement. Figure 6 presents a toy example for illustration. After applying M_{g}, the distillation objective of MCD requires that the probabilities of classes B, C, and D be close. This

L_{CE}	L_{SCD}	L_{MCD} M_g	L_{MCD} M_{ge}	Accuracy
✓				72.50
✓	✓			73.55
✓		✓		75.50
✓			✓	75.64
✓	✓	✓		75.53
✓	✓		✓	**76.64**

Table 6. Ablation study on the importance of each component in RLD. Top-1 accuracy (%) on the CIFAR-100 validation set is reported. The teacher is ResNet32×4, and the student is ResNet8×4. The reported results are the mean of three trials.

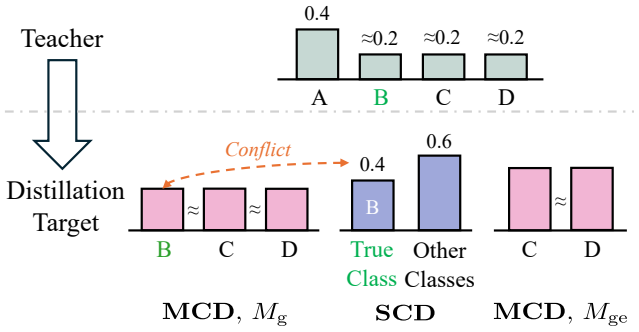


Figure 6. A toy example illustrates how the M_g masking strategy can lead to loss conflict. In this example, the probabilities of classes B, C, and D in the teacher distribution are close, and the value corresponding to B is slightly larger than those of C and D.

conflicts with the objective of SCD, which enforces a probability of 0.4 for class B, since the sum of class probabilities cannot exceed 1. In contrast, this issue does not arise when using M_{ge}. Therefore, we opt for M_{ge} as the masking strategy.

Hyper-parameter Analysis. We investigate the impact of the hyper-parameters α and β, which correspond to the importance of L_{SCD} and L_{MCD}, respectively. As shown in Figure 7, a more detailed hyper-parameter search can significantly enhance the effectiveness of RLD. Notably, the optimal hyper-parameter configurations vary significantly across different distillation pairs, which fundamentally explains why existing studies [34, 48] do not adopt fixed hyper-parameter settings. Additionally, we reveal an important phenomenon: unlike DKD, where optimal performance is achieved at $\alpha = 1$ [48], RLD generally prefers larger α values. This difference may stem from the following mechanism: when the teacher model makes incorrect predictions, smaller α values in the DKD loss can mitigate the negative impact of incorrect teacher knowledge but also

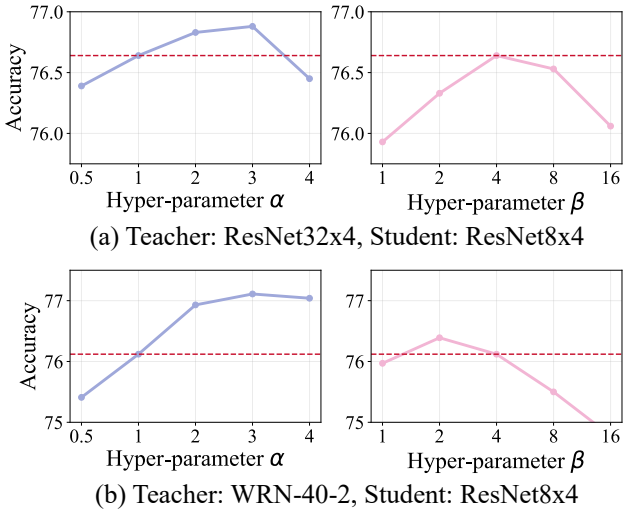


Figure 7. Impact of the hyper-parameters α (L_{SCD}) and β (L_{MCD}) on the CIFAR-100 validation set. By default, for both distillation pairs, $\alpha = 1$ and $\beta = 4$ (corresponding to the accuracies reported in Section 5.2, marked with dashed lines). The reported results are the mean of three trials.

limit the transfer of knowledge to some extent. In contrast, RLD refines the knowledge to effectively eliminate interference from incorrect information, allowing it to adapt to larger α values and further improve model performance.

6. Conclusion

Existing knowledge distillation methods do not consider the impact of incorrect teacher predictions on students. Alternatively, teacher outputs are arbitrarily corrected, disrupting class correlations. In this paper, we introduce Refined Logit Distillation (RLD) to address these issues. RLD enables teacher models to impart two distinct forms of knowledge to the student models: sample confidence and masked correlation. It effectively mitigates over-fitting and eliminates potential misinformation from the teacher models, while maintaining class correlations. Experimental results demonstrate the superiority of RLD.

Future Work. There are a few directions to improve our proposed RLD. For instance, dynamic temperature [21] and meta-learning [13] techniques can be used to tune the hyper-parameters. Additional strategies such as data augmentation [6] and sample selection [20] can be employed to distill high-quality samples. Besides, combining RLD with state-of-the-art feature distillation methods may be a promising avenue of exploration to further improve the distillation performance. We consider extending correction-based knowledge distillation to the feature domain, utilizing techniques such as Class Activation Mapping [39].

Acknowledgment

Wujie Sun and Can Wang are supported by the National Natural Science Foundation of China (No. 62476244), ZJU-China Unicom Digital Security Joint Laboratory and the advanced computing resources provided by the Supercomputing Center of Hangzhou City University.

References

[1] Qizhi Cao, Kaibing Zhang, Xin He, and Junge Shen. Be an excellent student: Review, preview, and correction. *IEEE Signal Processing Letters*, 30:1722–1726, 2023. 1, 2, 5

[2] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen. Online knowledge distillation with diverse peers. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3430–3437, 2020. 2

[3] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11933–11942, 2022. 1, 2, 5

[4] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2021. 1, 2, 5

[5] Tejalal Choudhary, Vipul Mishra, Anurag Goswami, and Jagannathan Sarangapani. A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, 53:5113–5155, 2020. 1

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019. 8

[7] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International conference on machine learning*, pages 1607–1616. PMLR, 2018. 2

[8] Ziyao Guo, Haonan Yan, Hui Li, and Xiaodong Lin. Class attention transfer based knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11868–11877, 2023. 5

[9] 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*, pages 770–778, 2016. 5

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

[11] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3779–3787, 2019. 2

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 1, 2, 5

[13] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021. 8

[14] 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, 2017. 5

[15] Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24276–24285, 2023. 1, 2, 5

[16] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018. 2

[17] Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. Self-knowledge distillation with progressive refinement of targets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6567–6576, 2021. 2

[18] Nikos Komodakis and Sergey Zagoruyko. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017. 1, 2, 5

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

[20] Weichao Lan, Yiu-ming Cheung, Qing Xu, Buhua Liu, Zhikai Hu, Mengke Li, and Zhenghua Chen. Improve knowledge distillation via label revision and data selection. *arXiv preprint arXiv:2404.03693*, 2024. 1, 2, 5, 8

[21] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1504–1512, 2023. 1, 2, 5, 8

[22] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7096–7104, 2019. 2

[23] Yifan Liu, Changyong Shu, Jingdong Wang, and Chunhua Shen. Structured knowledge distillation for dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):7035–7049, 2020. 2

[24] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 5

[25] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023. 2

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

[27] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5007–5016, 2019. 2

[28] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2014. 1, 5

[29] 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. 5

[30] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. 2

[31] 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*, pages 4510–4520, 2018. 5

[32] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5311–5320, 2021. 2

[33] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*. Computational and Biological Learning Society, 2015. 5

[34] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15731–15740, 2024. 1, 2, 5, 7, 8

[35] Wujie Sun, Defang Chen, Can Wang, Deshi Ye, Yan Feng, and Chun Chen. Accelerating diffusion sampling with classifier-based feature distillation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 810–815. IEEE, 2023. 2

[36] Wujie Sun, Defang Chen, Can Wang, Deshi Ye, Yan Feng, and Chun Chen. Holistic weighted distillation for semantic segmentation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 396–401. IEEE, 2023. 2

[37] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020. 1, 2, 5

[38] Can Wang, Defang Chen, Jian-Ping Mei, Yuan Zhang, Yan Feng, and Chun Chen. Semckd: Semantic calibration for cross-layer knowledge distillation. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 1, 2

[39] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 8

[40] Tiancheng Wen, Shenqi Lai, and Xueming Qian. Preparing lessons: Improve knowledge distillation with better supervision. *Neurocomputing*, 454:25–33, 2021. 1, 2, 5

[41] Guile Wu and Shaogang Gong. Peer collaborative learning for online knowledge distillation. In *Proceedings of the AAAI Conference on artificial intelligence*, pages 10302–10310, 2021. 2

[42] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12319–12328, 2022. 2

[43] Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. Reinforced multi-teacher selection for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14284–14291, 2021. 2

[44] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3903–3911, 2020. 6

[45] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*. British Machine Vision Association, 2016. 5

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

[47] 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*, pages 6848–6856, 2018. 5

[48] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022. 1, 2, 4, 5, 8