

A Good Teacher Adapts Their Knowledge for Distillation

Chengyao Qian Trung Le Mehrtash Harandi
Monash University, Australia

{Chengyao.Qian, trunglm, mehrtash.harandi}@monash.edu

Abstract

Knowledge distillation (KD) is an effective method for enhancing a small model, named student, by training it under the supervision of larger teacher models. However, existing studies indicate that a substantial capacity gap between the student and teacher can lead to poor learning for the student model. This capacity gap problem limits the applicability of KD and necessitates careful selection of the teacher’s size. Despite its importance, the underlying cause of the capacity gap problem remains underexplored. In this paper, we reveal that a substantial disparity in the output distributions of teacher and student models is a key factor behind this issue. To demonstrate this, we decompose the KD loss into two components: class-wise similarity and intra-class distribution, and analyze the contribution of each term. Our analysis shows that a large distributional mismatch can lead to poor student learning. Inspired by this observation, we propose the Adapted Intra-class Distribution (AID) method, wherein the teacher model is finetuned to optimize its intra-class distribution to better align with the student’s capacity prior to knowledge distillation. This approach effectively bridges the capacity gap between teacher and student models and consistently achieves state-of-the-art performance across a diverse range of architectures.

1. Introduction

Deep learning methods have achieved remarkable success in computer vision tasks [9, 20, 27]. However, their high performance comes at the cost of significant computational resource demands, which limits their applicability on resource-constrained devices, *e.g.* mobile devices and embedded systems. To overcome the limitation of high computational costs, knowledge distillation (KD) has been proposed. This technique aims to enhance the performance of small student models by learning the implicit knowledge from large teacher models [11].

KD methods can be divided into two main categories: logit-based and feature-based distillation. Logit-based ap-

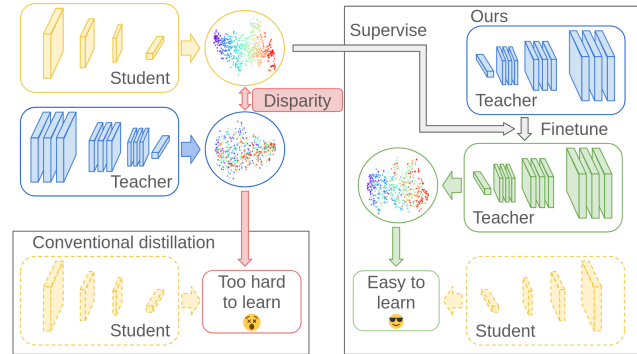


Figure 1. Capacity gap problem is caused by the disparity of intra-class distribution between teacher and student. Our method finetunes the pre-trained teacher models before KD to let the intra-class distribution of teacher match the student’s capacity.

proaches train student models to replicate the prediction probabilities of the teacher [11, 12, 32], whereas feature-based approaches guide the student to mimic the teacher’s internal representations [24, 28]. Although feature-based techniques provide richer information, they often require extra parameters to align the teacher’s features with the student’s. Recent findings have revealed a notable limitation of KD: student models trained under large teacher models frequently perform worse than those guided by mid-sized teachers known as the capacity gap problem [22].

To alleviate this capacity gap, some approaches have employed mid-sized teacher models as the bridge to assist smaller student models [22, 31]. However, this strategy introduces significant additional training overhead due to the extra mid-sized teachers. Moreover, studies [5, 34] indicate that teacher models that are not fully trained (*i.e.*, early-stopped) are more effective for training small students than fully trained teachers. While this approach improves student performance, it requires retraining teacher models, which increases computational resource demands and training time.

In this study, we study the capacity gap problem in KD and demonstrate that teacher must adapt their intra-class

distributions to better match the capacity of the small student. To uncover the root cause of the capacity gap problem, we analyze the Kullback–Leibler (KL) divergence to reveal the “dark knowledge” underlying KD. Our analysis shows that the KD loss consists of two components: class-wise similarity and intra-class distribution. The class-wise similarity component provides information on how an image relates to other classes and effectively serves as a label smoothing regularization, remaining robust regardless of the teacher’s size. In contrast, the intra-class distribution captures the relationships among samples within the same class and is highly sensitive to the teacher’s capacity. As a result, a large teacher can learn intra-class distributions that are challenging for a small student to mimic, ultimately degrading the student’s performance. We propose a novel Adapted Intra-class Distribution (AID) Knowledge Distillation method. This approach fine-tunes the teacher model to produce a more suitable intra-class distribution before training the student model. By only using the KL loss, our method surpasses existing state-of-the-art (SOTA) methods across various datasets.

Our contributions are summarized as follows:

- We decompose the KD loss (Kullback–Leibler divergence) to uncover the capacity gap issue in KD. Our analysis identifies two essential components within the KD loss: class-wise similarity and intra-class distribution. The class-wise similarity acts as a label smoothing regularizer for each class and remains consistent across teacher models of different sizes.
- We show that the misaligned intra-class distribution of the large teacher model contributes to the capacity gap problem. In particular, the intra-class distributions learned by the large teacher are challenging for a small student model which is overlooked by existing work.
- We propose an adaptive teacher approach that fine-tunes the intra-class distribution of a large teacher to better align with the capacity of a small student. This adjustment effectively bridges the capacity gap and enhances the performance of student models.
- Our method sets new baseline performance across various datasets. For instance, Resnet20 trained using our approach with teacher ResNet110×4 on CIFAR-100 achieves a top-1 accuracy of 72.70%, which is approximately 1.6% higher than the best existing method.

2. Related Work

Knowledge distillation.

Knowledge distillation was initially proposed by [11], where KL divergence is used to let the student mimic the teacher’s output logits. Building on this, Decouple KD [37] separates the KD loss into contributions from target and non-target classes. In addition to logits-based methods, feature-based approaches proposed by [28] focus on align-

ing the student’s features with those of the teacher. More recent methods incorporate contrastive learning and feature relational information [24, 33], and research has also focused on selecting optimal features and layers for alignment [3, 14, 18]. A significant limitation of feature-based methods is their reliance on additional parameters to match feature dimensions.

Teacher Assistant Knowledge Distillation (TAKD) [22] first brought attention to the capacity gap in KD, revealing that KD’s performance deteriorates when there is a large disparity between the sizes of teacher and student models. To overcome this, TAKD employs mid-sized teacher models as intermediaries to bridge the gap and enhance student performance. Similarly, Densely Guided Knowledge Distillation (DGKD) [31] combines insights from multiple mid-sized teachers to further improve the student’s outcomes. However, a key drawback of both approaches is their reliance on training several mid-sized teacher models, which incurs significant computational costs. Research has also found that teacher models stopped early during training often yield better distillation results than their fully-trained counterparts [5, 34]. Despite this, these methods still require retraining teacher models, which is computationally expensive given their large size.

Furthermore, the influence of the temperature in the KD loss has been thoroughly explored [7, 13, 17, 19], leading to dynamic temperature strategies that aim to narrow the prediction gap between teacher and student models. Recent investigations have also looked into normalizing logits to boost KD performance [4, 32]. Additionally, batch-wise logits alignment techniques [12, 15] have been introduced, which focus on aligning not just individual sample logits but also the logits across channels at the batch level.

Despite these advancements in bridging the capacity gap between the large teacher and the small student, the performance of student models supervised by large teachers still lags behind those supervised by mid-size teachers. Moreover, the underlying causes of the capacity gap problem remain insufficiently explored. This motivated us to investigate the reasons behind the capacity gap and propose a novel method that fine-tunes the teacher to produce a more suitable intra-class distribution without requiring retraining of the teacher models. It has been noted that the limitations of pre-trained teachers have been explored by [5, 8, 26, 34]; however, all these methods involve retraining the teacher models.

3. Methodology

3.1. Preliminary

Consider a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, where $\mathbf{x} \in \mathbb{R}^n$ and $y \in \mathcal{Y} = \{1, 2, \dots, K\}$. Let $\mathbf{z}_i^s = f^s(\mathbf{x}_i; \boldsymbol{\theta}^s) \in \mathbb{R}^K$ represent the output logits of the student model, where f :

$\mathbb{R}^n \rightarrow \mathbb{R}^K$ is the student network, and θ^s are its trainable parameters. We use the superscript t for the corresponding teacher outputs.

For a given sample $\mathbf{x} \sim \mathcal{D}$, the vanilla KD loss is defined by

$$\mathcal{L}_{\text{KD}}(\mathbf{x}; \theta^s) = \text{KL}(\phi(\mathbf{z}^t/\tau) \parallel \phi(\mathbf{z}^s/\tau)), \quad (1)$$

where $\phi : \mathbb{R}^K \rightarrow \Delta^{K-1}$ is the softmax function, KL denotes the Kullback–Leibler divergence, and $\tau > 0$ is the temperature.

The student’s overall training objective, $\mathcal{L}_{\text{total}}(\theta^s)$, sums the classification loss and the weighted KD loss over all samples:

$$\mathcal{L}_{\text{total}}(\theta^s) = \sum_i^m \left\{ \mathcal{L}_{\text{cls}}(\mathbf{x}_i, \mathbf{y}_i; \theta^s) + \beta \mathcal{L}_{\text{KD}}(\mathbf{x}_i; \theta^s) \right\}, \quad (2)$$

where β is a hyperparameter that balances the two losses. $\mathcal{L}_{\text{cls}}(\mathbf{x}, \mathbf{y}; \theta^s) = -\mathbf{y}^\top \log(\phi(\mathbf{z}^s))$ is the standard cross-entropy loss, with \mathbf{y} representing the one-hot label.

3.2. Know the KD Better

To uncover the underlying cause of the capacity gap problem, we revisit the KD loss and decompose it into two components: class-wise similarity and intra-class distribution. Our empirical findings indicate that the disparity in distributions between the teacher and student is a key contributor to the capacity gap. Motivated by this insight, we introduce a novel approach—*AID knowledge distillation*—which fine-tunes the teacher model to generate an intra-class distribution that is more aligned with the capacity of the student model.

Existing methods [23, 35] have demonstrated that KD can be treated as label smoothing. In this paper, we show that the KD loss not only acts as label smoothing but also implicitly encodes information about the intra-class distribution. Specifically, the vanilla KD loss defined in Equation (1) can be reformulated as:

$$\begin{aligned} \mathcal{L}_{\text{KD}}(\mathbf{x}; \theta^s) &= \phi(\mathbf{z}^t/\tau)^\top \log \left(\frac{\phi(\mathbf{z}^t/\tau)}{\phi(\mathbf{z}^s/\tau)} \right) \\ &= \underbrace{\phi(\mathbf{z}^t/\tau)^\top \log \phi(\mathbf{z}^t/\tau)}_{\text{Constant}} - \phi(\mathbf{z}^t/\tau)^\top \log \phi(\mathbf{z}^s/\tau) \end{aligned} \quad (3)$$

The teacher model is kept frozen during training; consequently, the term $\phi(\mathbf{z}^t/\tau)^\top \log \phi(\mathbf{z}^t/\tau)$ remains constant throughout the training process and does not contribute to the update of the student model. Ignoring the constant term $\phi(\mathbf{z}^t/\tau)^\top \log \phi(\mathbf{z}^t/\tau)$, then the optimized student model

is:

$$\theta^{s*} \in \arg \min_{\theta^s} - \sum_i^m \left[\mathbf{y}_i^\top \log(\phi(\mathbf{z}_i^s)) + \underbrace{\beta \phi(\mathbf{z}_i^t/\tau)^\top \log \phi(\mathbf{z}_i^s/\tau)}_{\text{KD term}} \right]. \quad (4)$$

Let $\mathbf{a}[j]$ denote the j -th element of \mathbf{a} . Then, the KD terms in Equation (4) can be written as

$$\phi(\mathbf{z}_i^t/\tau)^\top \log(\phi(\mathbf{z}_i^s/\tau)) = \sum_j \phi(\mathbf{z}_i^t/\tau)[j] \log(\phi(\mathbf{z}_i^s/\tau)[j]).$$

As such, for a class K , we can decompose the term $\frac{1}{m} \sum_i \phi(\mathbf{z}_i^t/\tau)[K] \log(\phi(\mathbf{z}_i^s/\tau)[K])$ as:

$$\begin{aligned} &\frac{1}{m} \sum_i \phi(\mathbf{z}_i^t/\tau)[K] \log(\phi(\mathbf{z}_i^s/\tau)[K]) \\ &= \underbrace{\frac{1}{m} \sum_i \mu^t[K] \log(\phi(\mathbf{z}_i^s/\tau)[K])}_{\text{class-wise similarity}} \\ &\quad + \underbrace{\text{cov}(\phi(\mathbf{z}^t/\tau)[K], \log(\phi(\mathbf{z}^s/\tau)[K]))}_{\text{intra-class distribution}}, \end{aligned} \quad (5)$$

where $\mu^t[K] = \frac{\sum_i \phi(\mathbf{z}_i^t/\tau)[K]}{m}$ calculates the teacher’s average prediction for class K with temperature τ and cov is covariance.

Class-wise similarity. The average prediction $\mu^t[K]$ remains constant during training. Consequently, the class-wise similarity can be treated as label smoothing for each class. Unlike the label smoothing approach proposed in [35], which assigns a uniform probability across all non-target classes, the class-wise similarity term in the KD loss provides specific class relationships from the teacher to the student. For example, in the context of vehicle images, the probability assigned to a motorcycle should be higher than that assigned to unrelated classes like dogs. This class-wise similarity is robust across models of varying sizes.

Intra-class distribution. The intra-class distribution sheds light on how a teacher model assigns predicted probabilities to samples within the same class, effectively indicating the relative difficulty of each sample. A lower predicted probability implies that a sample is more challenging, while a higher probability denotes an easier instance. For example, consider a clear image of a cat compared to a cat image with distracting noise such as a dog’s face. Even though both images are correctly classified as cats, the teacher exhibits higher confidence in the clear image, suggesting that

its features are more representative of the cat class. As a result, the student model should prioritize learning from the clear image for a more reliable representation of cat features.

Furthermore, a large teacher model, with its extensive capacity, can capture a broader range of features, including those that help distinguish noisy images. In contrast, a mid-sized model may struggle with such images, assigning lower probabilities due to the confusing noise. Given that the student is a small model, receiving high probability predictions from a large teacher for noisy images can mislead its training, as the student may attempt to mimic these overly confident yet less reliable predictions.

The contribution of each term. In order to investigate the contribution of each term in the KD loss to the student model, we decouple the KD loss and train the student models with each term individually. The class-wise similarity term, as described in Equation (5), can be considered as label smoothing regularization. The loss function combines this class-wise similarity with cross-entropy loss.

$$\mathcal{L}_{\text{cls-wise}}(\theta^s) = \sum_{i=1}^m \left\{ \mathcal{L}_{\text{cls}}(\mathbf{x}_i, \mathbf{y}_i; \theta^s) - \beta \mu^t \log(\phi(z^s/\tau)) \right\}, \quad (6)$$

where μ^t is the average prediction of the teacher with temperature τ .

The intra-class distribution term is crucial because it captures the relative difficulty of each sample. By reducing the covariance between the teacher’s and the student’s predictions, we enable the student to emulate the teacher’s assessment of each image’s difficulty. The loss function for the intra-class distribution term is given by:

$$\mathcal{L}_{\text{intra-cls}}(\theta^s) = \sum_{i=1}^m \left\{ \mathcal{L}_{\text{cls}}(\mathbf{x}_i, \mathbf{y}_i; \theta^s) - \beta \text{cov}_{\mathcal{B}}(\phi(\mathbf{z}_i^t/\tau), \log(\phi(\mathbf{z}_i^s/\tau))) \right\}, \quad (7)$$

The experiments, as shown by the blue line in Figure 2, demonstrate an interesting finding: when student models are trained solely using the class-wise similarity term, their performance remains stable regardless of the teacher model’s size. This stability indicates that the capacity gap issue is not present in this setting.

Conversely, when student models are trained with the intra-class distribution term, performance deteriorates under supervision from a large teacher, as illustrated by the orange line in Figure 2. This outcome reinforces our hypothesis from the previous section that an unsuitable teacher distribution contributes to the capacity gap problem. In a brief

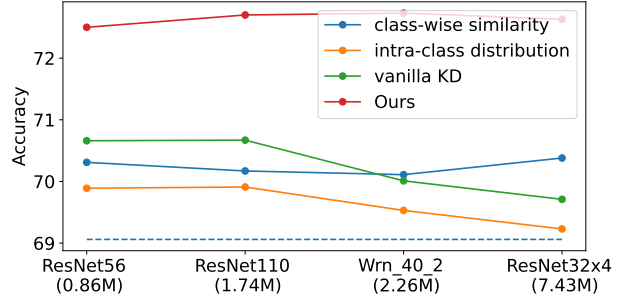


Figure 2. Accuracy of ResNet20 student model under different knowledge distillation strategies. The x-axis represents various teacher models sorted by increasing model size. The blue dashed line indicates the accuracy of the ResNet20 student trained from scratch without distillation. The green line shows the performance using vanilla KD. The solid blue line corresponds to distillation of only class-wise similarity from the teacher, while the orange line represents distillation of only intra-class distribution terms. The red line denotes the performance achieved by our proposed method.

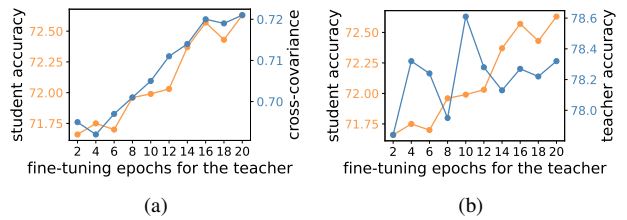


Figure 3. This figure presents the accuracy of the student model (ResNet20) relative to (a) the cross-covariance between teacher and student predictions, and (b) the accuracy of the teacher model.

summary, a larger teacher model, due to its greater capacity to extract diverse features, tends to exhibit higher confidence on challenging samples that often mislead smaller models. When these a high-capacity teacher guides the small student model, they inadvertently encourage the students to mimic an intra-class distribution that is not well-suited for their capacity, leading to suboptimal learning. Since the teacher model remains fixed during standard KD, traditional methods are unable to correct this misalignment. While some approaches [5, 8, 34] have resorted to retraining teacher models for a better match, this comes with significant computational costs. To address this, we introduce the ATI KD method, which refines the teacher’s intra-class distribution without the need for extensive retraining, thereby establishing a new state-of-the-art baseline for knowledge distillation.

3.3. Adjust Teacher Intra-class Distribution

In the previous section, we demonstrated that the capacity gap problem is due to a misalignment in the intra-class distributions between the large teacher and small student.

Since conventional KD fixes the pre-trained teacher during distillation, it cannot resolve this disparity. While some retraining methods [8, 26] attempt to adjust the teacher, they incur high computational costs and can even degrade teacher performance. In contrast, our approach leverages an adaptive teacher strategy for small student model, which proves more effective than traditional retraining techniques.

We fine-tune the pre-trained teacher before applying KD to adjust its intra-class distributions, making it more compatible with the small student. Specifically, we use a trained student model to fine-tune the teacher over several epochs, enabling the teacher to produce a more appropriate intra-class distribution. The loss function employed for fine-tuning is the KD loss defined in Equation (2). During this process, we observe an increase in the cross-covariance between the teacher’s and the student’s predictions, indicating that the teacher is adapting its knowledge to better match the student’s capacity.

As illustrated in Figure 3a, the right y-axis shows the cross-covariance between teacher and student predictions across training epochs, while the left y-axis depicts the student accuracy when supervised by the corresponding teacher. It is evident that student accuracy improves as the cross-covariance increases. Notably, without fine-tuning, the cross-covariance is below 0.4, and the student (ResNet20) achieves an accuracy of 69.83% when supervised by a ResNet32×4 teacher.

We demonstrate the effectiveness of our approach with the red line shown in Figure 2. To ensure a fair comparison between fine-tuned and un-fine-tuned teachers, we plot student accuracy against teacher accuracy in Figure 3b. Interestingly, a high teacher accuracy does not necessarily translate to better student performance. For example, while the un-fine-tuned ResNet32×4 achieves an accuracy of 79.42%, which is notably higher than that of its fine-tuned one, the student model supervised by the fine-tuned ResNet32×4 ultimately performs better than the one supervised by the un-fine-tuned version.

The pseudo code for our approach is presented in Algorithm 1. We assume that a pre-trained student model is available, and that training the student is relatively cheap due to its smaller size compared to the teacher model. Our method involves using this pre-trained student to fine-tune the teacher model, and then employing the fine-tuned teacher to supervise the training of a student model from scratch with vanilla KD loss.

4. Experiments

Models and Methods for comparison. We evaluate our method using the CIFAR-100 [16], Oxford-IIIT Pet [25] and ImageNet [6] datasets. For CIFAR-100, the teacher models with number of parameters include ResNet110×4: 27.2M [10], ResNet56×4: 13.6M [10], ResNet32×4:

Algorithm 1 AID Knowledge Distillation

Input: Pretrained teacher $f^t(\mathbf{x}; \theta_t)$ and student $f^s(\mathbf{x}; \theta^s)$. Dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$. Temperature τ . Fine-tune epochs and KD epochs. Loss $L_{\text{cls}}(\mathbf{z}, \mathbf{y}) = -\mathbf{y}^\top \log(\phi(\mathbf{z}))$ and $L_{\text{KD}}(\mathbf{z}^s; \mathbf{z}^t) = \text{KL}(\phi(\mathbf{z}^t/\tau) \parallel \phi(\mathbf{z}^s/\tau))$

```

1: for each fine-tune epoch do
2:   for each  $(\mathbf{x}_i, y_i) \in \mathcal{D}$  do
3:      $\mathbf{z}_i^s, \mathbf{z}_i^t \leftarrow f^s(\mathbf{x}_i; \theta^s), f^t(\mathbf{x}_i; \theta^t)$ 
4:     update  $\theta^t$  towards minimizing  $L_{\text{cls}}(\mathbf{z}_i^t, \mathbf{y}_i) + \beta L_{\text{KD}}(\mathbf{z}_i^s; \mathbf{z}_i^t)$   $\triangleright$  The student model is frozen during this stage.
5:   end for
6: end for
7: for each KD epoch do
8:   for each  $(\mathbf{x}_i, y_i) \in \mathcal{D}$  do
9:      $\mathbf{z}_i^s, \mathbf{z}_i^t \leftarrow f^s(\mathbf{x}_i; \theta^s), f^t(\mathbf{x}_i; \theta^t)$ 
10:    update  $\theta^s$  towards minimizing  $L_{\text{cls}}(\mathbf{z}_i^s, \mathbf{y}_i) + \beta L_{\text{KD}}(\mathbf{z}_i^s; \mathbf{z}_i^t)$   $\triangleright$  The teacher model is frozen during this stage.
11:   end for
12: end for

```

7.43M [10], WRN-40-2: 2.25M [36], VGG13: 9.46M [30], ResNet50: 25.6M [10], ResNet110:1.73M [10], and ResNet56: 0.86M [10]. The student models comprise SHN-V2: 1.35M [21], ResNet8×4: 1.23M [10], MN-V2: 0.81M [29], WRN-16-2: 0.7M [36], WRN-40-1: 0.57M [36], ResNet20: 0.28M [10], ResNet14: 0.18M [10] and ResNet8: 0.08M [10].

We compare our method with SOTA methods including KD [11], DKD [37], TAKD [22], DGKD [31], CTKD [17], MLKD [15], STD [32], MSE [8], SemCKD [1], ReviewKD [3], SimKD [2]. For SimKD, which introduces an additional layer in the student model, we use a 1×1 convolutional layer to reduce the influence of extra parameters for a fair comparison.

Training Setting. The pre-trained student models are obtained using the same settings as in the KD experiments without KD loss. The hyperparameter settings are as follows:

CIFAR-100: All models are trained for 240 epochs with an initial learning rate of 0.05, which is reduced by a factor of 0.1 at epochs 150, 180, and 210. The decay rate is set to 0.1. The batch size is set to 64, and data augmentation techniques include random cropping and horizontal flipping. The temperature τ is set to 4. β is set to 1. For teacher fine-tuning, the learning rate is set at 0.005 with a decay rate of 0.1. The loss for fine-tuning is Equation (2). The number of epochs for fine-tuning is between 10 and 30.

ImageNet: For the ImageNet experiments, all models are trained for 100 epochs with a batch size of 512. The learning rate starts at 0.1 and is decreased by a factor of 0.1 at

Teacher Student	ResNet56×4 ResNet8	ResNet110×4 ResNet8	ResNet56×4 ResNet14	ResNet110×4 ResNet14	ResNet56×4 WRN-16-2	ResNet110×4 WRN-16-2
Teacher	78.92	79.26	78.92	79.26	78.92	79.26
Student	62.28	62.28	68.13	68.13	73.26	73.26
Ratio (T/S)	170	340	76	152	19	39
KD [11]	61.13 \pm 0.17	61.47 \pm 0.11	66.96 \pm 0.10	66.96 \pm 0.04	75.42 \pm 0.36	75.57 \pm 0.21
FitNet [28]	59.80 \pm 0.16	59.70 \pm 0.07	68.02 \pm 0.22	67.53 \pm 0.41	75.18 \pm 0.31	74.90 \pm 0.33
TAKD [22]	62.64 \pm 0.09	62.25 \pm 0.17	67.59 \pm 0.04	67.43 \pm 0.12	73.42 \pm 0.29	72.65 \pm 0.51
DGKD [31]	61.45 \pm 0.16	60.96 \pm 0.27	67.61 \pm 0.10	67.70 \pm 0.18	75.08 \pm 0.134	74.92 \pm 0.59
SemCKD [1]	49.64 \pm 0.53	27.25 \pm 1.19	60.69 \pm 1.06	32.80 \pm 0.83	70.86 \pm 0.18	66.18 \pm 0.44
SimKD [2]	49.82 \pm 0.47	47.29 \pm 0.20	64.42 \pm 0.19	62.87 \pm 0.69	76.07 \pm 0.40	75.34 \pm 0.46
DKD [37]	58.61 \pm 0.14	57.88 \pm 0.48	67.30 \pm 0.68	66.97 \pm 0.26	76.65 \pm 0.57	76.09 \pm 0.23
CTKD [17]	56.98 \pm 0.39	45.16 \pm 0.25	66.46 \pm 0.40	62.06 \pm 0.78	73.91 \pm 0.13	74.50 \pm 0.22
MSE [8]	56.66 \pm 0.58	56.52 \pm 1.04	66.76 \pm 0.55	65.64 \pm 0.92	74.15 \pm 0.68	74.42 \pm 0.37
STD [32]	59.79 \pm 0.35	59.43 \pm 0.25	67.72 \pm 0.19	67.93 \pm 0.27	76.14 \pm 0.34	75.93 \pm 0.17
Ours	63.28\pm0.05	63.44\pm0.23	69.99\pm0.16	69.73\pm0.35	76.66\pm0.23	76.68\pm0.37

Table 1. Top-1 accuracy(%) on CIFAR-100 for large capacity gap between teacher and student. We use the reimplementation from the STD repository. The best results are bolded. ResNet56×4 and ResNet110×4 are trained on CIFAR-100 with 240 epochs. “Ratio”: the size ratio between teacher and student.

Teacher Student	ResNet56 ResNet20	ResNet110 ResNet20	WRN-40-2 ResNet20	ResNet32×4 ResNet20	ResNet56×4 ResNet20	ResNet110×4 ResNet20
Teacher	72.34	74.31	75.61	79.42	78.92	79.26
Student	69.06	69.06	69.06	69.06	69.06	69.06
Ratio (T/S)	3.1	6.2	8.0	26.5	47.6	97.1
KD [11]	70.66 \pm 0.26	70.67 \pm 0.15	70.10 \pm 0.11	69.83 \pm 0.14	70.03 \pm 0.29	69.76 \pm 0.11
DKD [37]	71.97 \pm 0.36	72.01 \pm 0.47	71.38 \pm 0.34	70.52 \pm 0.13	71.09 \pm 0.20	71.13 \pm 0.18
MLKD [15]	72.19 \pm 0.27	71.89 \pm 0.78	67.92 \pm 0.56	63.71 \pm 0.34	-	51.73 \pm 0.45
STD [32]	71.43 \pm 0.17	71.48 \pm 0.31	70.71 \pm 0.39	69.58 \pm 0.26	71.21 \pm 0.19	71.05 \pm 0.48
Ours	72.50\pm0.16	72.68\pm0.24	72.73\pm0.15	72.63\pm0.09	72.71\pm0.25	72.70\pm0.39

Table 2. Top-1 accuracy(%) on CIFAR-100 for different size of teacher. We use the reimplementation from the STD repository. The best results are bolded. ResNet56×4 and ResNet110×4 are trained on CIFAR-100 with 240 epochs. For the remaining teacher models, we use the pretrained models provided by the STD repository. “Ratio”: the size ratio between teacher and student.

epochs 30, 60, and 90, with a weight decay of 0.0001.

Oxford pets: For the Oxford pets experiments, all models are trained for 100 epochs with a batch size of 64. The learning rate starts at 0.1 and is decreased by a factor of 0.1 at epochs 30, 60, and 90, with a weight decay of 0.0001.

More experimental details are in Appendix.

4.1. Results

CIFAR-100. To emphasize the capacity gap problem, we introduce ResNet110×4 and ResNet56×4 as large teacher models and use ResNet14 and ResNet8 as tiny student models in our KD experiments. ResNet110×4 and ResNet56×4 denote models whose width is four times as ResNet110 and ResNet56, respectively. These models were selected due to their significant differences in size and performance, which create a pronounced disparity in model capacities. By em-

ploying such an extreme teacher-student size ratio, we aim to evaluate the effectiveness of our proposed method compared to existing methods.

The results for cases with a large capacity gap between the teacher and student are shown in Table 1. For TAKD [22], the teacher progression follows the pathway: ResNet110×4 → ResNet56×4 → ResNet32×4 → ResNet110 → ResNet56 → ResNet32 → ResNet20 → ResNet14, while DGKD [31] utilizes all available mid-size teachers. In the scenario with the largest teacher-student size ratio, where the teacher is 340 times larger than the student (using ResNet110×4 as the teacher and ResNet8 as the student), our proposed method outperforms the state-of-the-art approaches by **1.2%**. Notably, when the teacher-student size ratio is extremely high, standard KD even degrades student performance relative to training the student from

Teacher Student	ResNet32×4 SHN-V2	WRN-40-2 MN-V2	VGG13 MN-V2	ResNet50 MN-V2	ResNet32×4 WRN-16-2	ResNet32×4 WRN-40-2	WRN-40-2 ResNet8×4
Teacher	79.42	75.61	74.64	79.34	79.42	79.42	75.61
Student	71.82	64.60	64.60	64.60	73.26	75.61	72.50
Ratio (T/S)	5.5	2.8	11.7	31.6	10.6	3.3	1.8
KD [11]	74.45	68.36	67.37	67.35	74.90	77.70	73.97
FitNet [28]	73.54	68.64	64.16	63.16	74.70	77.69	74.61
RKD [24]	73.21	69.27	64.52	64.43	74.86	77.82	75.26
CRD [33]	75.65	70.28	69.73	69.11	75.65	78.15	75.24
DKD [37]	77.07	69.28	69.71	70.35	75.70	78.46	75.56
CTKD [17]	75.37	68.34	68.50	68.67	74.57	77.66	74.61
STD [32]	75.56	69.23	68.61	69.02	75.26	77.92	77.11
Ours	77.81	70.53	69.98	70.39	76.87	78.74	77.45

Table 3. Top-1 accuracy(%) on CIFAR-100 with conventional teacher-student pairs. The best results are in **bold**. “Ratio”: the size ratio between teacher and student.

Student	Teacher	CRD	ReviewKD	TAKD	DGKD	CKD	STD	DKD	Ours
ResNet18	ResNet34	71.38	71.61	71.37	71.73	70.98	71.42	71.70	72.01
ResNet18	ResNet50	70.90	70.96	-	-	70.74	71.78	72.04	72.26

Table 4. Top-1 accuracy(%) on ImageNet.

Teacher Student	ResNet110×4		
	WRN-40-1	SHN-V2	MN-V2
Teacher	79.26	79.26	79.26
Student	71.98	71.82	64.60
Ratio (T/S)	47.7	20.1	33.6
KD [11]	73.99	76.36	66.20
TAKD [22]	71.37	74.91	65.07
DGKD [31]	73.42	76.99	67.59
DKD [37]	74.76	76.62	61.12
MLKD [15]	70.75	-	66.34
CTKD [17]	68.27	77.60	66.28
STD [32]	74.35	77.07	68.37
Ours	75.46	77.96	68.48

Table 5. Top-1 accuracy(%) on CIFAR-100. We use the reimplementation from the STD repository. The best results are bolded. “Ratio”: the size ratio between teacher and student.

scratch, whereas our method consistently improves student performance.

Table 2 shows the performance outcomes when using teacher models of different sizes. For all existing methods, we observe that increasing the teacher’s size leads to a decrease in student performance. In contrast, our proposed method enhances the student’s performance as larger teacher models are used. Moreover, when extremely large teachers are employed, the student’s performance stabilizes, which we interpret as the student reaching its optimal per-

Teacher Student	ResNet50 ResNet18	ResNet101 ResNet18
Teacher	87.52	89.02
Student	85.78	85.78
Ratio (T/S)	2.2	3.8
KD [11]	86.01	86.14
DKD [37]	86.54	86.47
MLKD [15]	85.98	86.35
CTKD [17]	86.28	86.17
STD [32]	86.68	86.77
Ours	87.11	87.23

Table 6. Top-1 accuracy(%) on Oxford-IIIT Pet. We use the reimplementation from the STD repository. The best results are bolded. “Ratio”: the size ratio between teacher and student.

	TAKD	DGKD	MLKD	MSE	Ours
RTE (mins)	699.8	325.9*	535.2	458.2	179.4

Table 7. RTE on CIFAR-100. DGKD* excludes the training for assistant teachers.

formance level.

Table 3 presents the results for a conventional teacher-student pair. Notably, even when the teacher and student models have different architectures and the size ratio between them is not very large, our method still effectively enhances student performance and outperforms existing meth-

Teacher	ResNet32×4	ResNet32×4	ResNet32×4
MFT	ResNet20	ResNet32	WRN-40-1
ResNet20	72.63	72.66	72.43
ResNet32	74.47	74.80	74.96
WRN-40-1	75.08	75.54	75.77

Table 8. Fine-tune teacher with different students. “MFT”: the model used to fine-tune the teacher.

Teacher	ResNet110	WRN-40-2	ResNet32×4
MFT	WRN-40-2	WRN-40-2	WRN-40-2
ResNet20	70.07	70.35	70.04
ResNet32	74.00	73.92	74.04
WRN-40-1	74.61	75.00	74.66

Table 9. Fine-tune teacher with a large model. “MFT”: the model used to fine-tune the teacher.

ods. Additionally, Table 5 shows the results for different architectures of teachers and students with a large size ratio, where our method consistently achieves the best results.

In Appendix we compare our method to other SOTA methods on the CIFAR-100 dataset using the more conventional teacher-student pairs.

ImageNet. Table 4 presents the results on the ImageNet dataset. Our proposed method consistently outperforms existing approaches on this large-scale dataset. These improvements highlight the effectiveness of fine-tuning the teacher’s intra-class distribution, which enables the student to better mimic the teacher’s decision-making process.

Oxford Pets. In our work, we demonstrate that the intra-class distribution represents a key component of the “dark knowledge” in KD, and that disparities in this distribution between teacher and student hinder the effectiveness of knowledge distillation. To evaluate our method in scenarios with more complex intra-class distributions, we conducted experiments on the Oxford-IIIT Pet dataset. As shown in Table 6, our method consistently enhances student performance even in the dataset with complex intra-class distributions.

Run-time efficiency. Since our proposed method requires a trained student model and finetuning teachers, we compare the time needed for ATI KD with existing methods addressing the capacity gap problem, noted as run-time efficiency (RTE). For TAKD and MLKD, we follow the same training settings as mentioned in their original papers. For DGKD, we only account for the time of the final KD process, excluding the training time for the teacher models. In our experiments, the teacher model is ResNet110×4, and the student model is ResNet8. The RTE for vanilla KD is 126 minutes. In our proposed method, we include the time required to train the student model (39 minutes). Retraining

the teacher requires 334 minutes. As a teacher retraining method, MSE [8] also harms the performance of teachers. All RTE in Table 7 are obtained using a single NVIDIA A5500 GPU.

4.2. Ablation Study

Does the teacher model need to be fine-tuned with the target students to achieve better performance?

In this section, we show that teacher models fine-tuned using different students of similar capacity can still provide a suitable intra-class distribution for the students. This observation further validates our hypothesis that the intra-class distribution is closely linked to model capacity, and that a misalignment in this distribution leads to the capacity gap problem. We conduct experiments using ResNet20, ResNet32, and WRN-40-1 as student models, with ResNet32×4 acting as the teacher. The teacher model is fine-tuned using each of the two smaller student models separately, and the resulting fine-tuned teacher is then used to supervise the students. As illustrated in Table 8, even when the teacher is fine-tuned with a different, smaller student, performance improvements are still observed. Notably, ResNet32 and WRN-40-1 have a similar number of parameters, whereas ResNet20 is smaller. Consequently, if the teacher is fine-tuned with a model that is smaller than the target student, it may lose some of the knowledge that the larger student can learn, ultimately reducing the student’s performance.

Additionally, we investigate the fine-tuning of teachers with larger models, specifically WRN-40-2. Table 9 presents the performance of students supervised by these teachers. The limited improvement observed in this scenario supports our hypothesis that large teacher models must adapt their knowledge to the capacity of small students for effective knowledge transfer.

5. Conclusion

In this paper, we comprehensively analyze the capacity gap problem in KD. When small student models are supervised by large teacher models, their performance lags behind that of students supervised by mid-sized teachers. We reveal that the dark knowledge in KD comprises two key components: class-wise similarity and intra-class distribution. The intra-class distribution, which reflects the relative difficulty of samples within each class, is the primary cause of the capacity gap problem. Due to their greater capacity, large teacher models can learn more features and confidently predict samples that are challenging for smaller models. When small student models are pushed to predict high probabilities for these difficult samples, it results in poor learning outcomes. To address this issue, we propose a novel method that fine-tunes pre-trained teacher models to adjust the intra-class distribution to be more suitable for small students.

Acknowledgements

This work was supported by Australian Research Council (ARC) Discovery Program DP250100262, DP230101176, and by the Air Force Office of Scientific Research under award number FA2386-23-1-4044. The authors gratefully acknowledge the anonymous reviewers for their insightful feedback and valuable suggestions, which have significantly improved the quality of this work.

References

- [1] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 7028–7036, 2021.
- [2] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [3] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5008–5017, 2021.
- [4] Zhihao Chi, Tu Zheng, Hengjia Li, Zheng Yang, Boxi Wu, Binbin Lin, and Deng Cai. Normkd: Normalized logits for knowledge distillation, 2023.
- [5] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 4794–4802, 2019.
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009.
- [7] Jia Guo. Reducing the teacher-student gap via adaptive temperatures, 2022.
- [8] Shayan Mohajer Hamidi, Xizhen Deng, Renhao Tan, Linfeng Ye, and Ahmed Hussein Salamah. How to train the teacher model for effective knowledge distillation. In *Proc. European Conf. on Computer Vision (ECCV)*, 2024.
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [11] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2014.
- [12] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [13] Aref Jafari, Mehdi Rezagholizadeh, Pranav Sharma, and Ali Ghodsi. Annealing knowledge distillation. *CoRR*, abs/2104.07163, 2021.
- [14] Mingi Ji, Byeongho Heo, and Sungrae Park. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 7945–7952, 2021.
- [15] Y. Jin, J. Wang, and D. Lin. Multi-level logit distillation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 24276–24285, Los Alamitos, CA, USA, 2023. IEEE Computer Society.
- [16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [17] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 1504–1512, 2023.
- [18] Sihao Lin, Hongwei Xie, Bing Wang, Kaicheng Yu, Xiaojun Chang, Xiaodan Liang, and Gang Wang. Knowledge distillation via the target-aware transformer. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10915–10924, 2022.
- [19] Jihao Liu, Boxiao Liu, Hongsheng Li, and Yu Liu. Meta knowledge distillation, 2022.
- [20] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.
- [21] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proc. European Conf. on Computer Vision (ECCV)*, 2018.
- [22] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 5191–5198, 2020.
- [23] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.
- [24] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3967–3976, 2019.
- [25] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3498–3505. IEEE, 2012.
- [26] Chengyao Qian, Munawar Hayat, and Mehrtash Harandi. Can we distill knowledge from powerful teachers directly? In *Proc. IEEE International Conference on Image Processing (ICIP)*, pages 595–599, 2023.
- [27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 28, 2015.

- [28] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *Proc. Int. Conf. on Learning Representation (ICLR)*, 2015.
- [29] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [31] Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. Densely guided knowledge distillation using multiple teacher assistants. In *Proc. Int. Conf. on Computer Vision (ICCV)*, pages 9395–9404, 2021.
- [32] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [33] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *Proc. Int. Conf. on Learning Representation (ICLR)*, 2020.
- [34] Chaofei Wang, Qisen Yang, Rui Huang, Shiji Song, and Gao Huang. Efficient knowledge distillation from model checkpoints. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- [35] Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [36] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- [37] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 11953–11962, 2022.