Revisit the Essence of Distilling Knowledge through Calibration

Wen-Shu Fan [1 2] Su Lu [1 2] Xin-Chun Li [1 2] De-Chuan Zhan [1 2] Le Gan [1 2]

Abstract

Knowledge Distillation (KD) has evolved into a practical technology for transferring knowledge from a well-performing model (teacher) to a weak model (student). A counter-intuitive phenomenon known as capacity mismatch has been identified, wherein KD performance may not be good when a better teacher instructs the student. Various preliminary methods have been proposed to alleviate capacity mismatch, but a unifying explanation for its cause still lacks. In this paper, we propose *a unifying analytical framework to pinpoint the core of capacity mismatch based on calibration*. Through extensive analytical experiments, we observe a positive correlation between the calibration of the teacher model and the KD performance with original KD methods. As this correlation arises due to the sensitivity of metrics (e.g., KL divergence) to calibration, we recommend employing measurements insensitive to calibration such as ranking-based loss. Our experiments demonstrate that ranking-based loss can effectively replace KL divergence, aiding large models with poor calibration to teach better.

1. Introduction

Knowledge Distillation (KD) is a model reuse technique that involves having a large model (teacher) teach a smaller model (student). The function of KD is to enhance the performance of smaller models on resource-constrained platforms, such as mobile devices (Sandler et al., 2018; Zhang et al., 2018; Ma et al., 2018; Tan & Le, 2019).

A prevailing intuition is that a more prominent teacher with stronger expressive capabilities is more likely to impart superior knowledge. However, recent studies have revealed that increasing the capacity of the teacher adversely impacts the performance of KD (Cho & Hariharan, 2019; Li et al., 2021). Researchers attribute this phenomenon to a capacity mismatch between the teacher's and student's models. Some efforts have focused on narrowing the size gap between both architectures (Mirzadeh et al., 2020; Zhu & Wang, 2021), other approaches have aimed to regularize teacher networks to narrow capacity gap (Cho & Hariharan, 2019; Wang et al., 2022). Recently, several studies (Li et al., 2022; Zhao et al., 2022) have aimed to provide a theoretical explanation for why larger models may not teach better in KD. Although these studies offer diverse perspectives to explain why a larger teacher network may not inherently enhance teaching performance, the explanations remain fragmented and lack a more comprehensive framework (Li et al., 2023a;b).

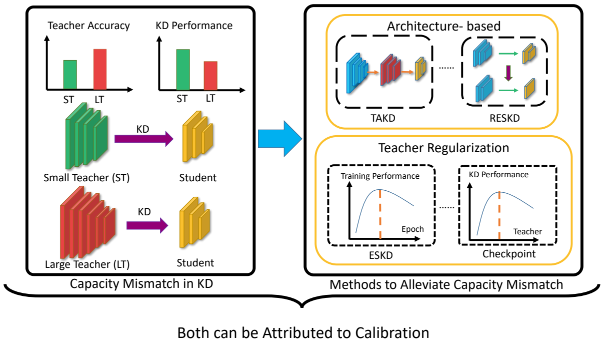

Figure 1. The significance of our work: On the left, a large teacher achieves high performance but struggles in low KD performance, known as capacity mismatch. The right part depicts existing solutions, including architecture adjustments and teacher regularization. Our calibration-based framework provides an explanation for both capacity mismatch and these solutions.

The primary objective of this paper is to explore the reasons behind the occurrence of capacity mismatch. We hold the view that teacher's accuracy does not directly reflect the performance of the distilled student. Instead, we focus on calibration of the teacher model. If the probability associated with the predicted class label can reflect its ground truth correctness likelihood, we say this network is of calibrated confidence (Guo et al., 2017). The research indicates that larger teacher models, while achieving better performance,

[*] Equal contribution [1] School of Artificial Intelligence, Nanjing University, China [2] National Key Laboratory for Novel Software Technology, Nanjing University, China. Correspondence to: De-Chuan Zhan <zhandc@nju.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

tend to have poorer calibration. Therefore, a hypothesis arises: *the decline in the calibration of the larger teacher precisely leads to the deterioration of the student's performance*. Through comparative experiments, we observe a close alignment between changes in KD performance and variations in teacher model's calibration. We also find that a poorly calibrated teacher is unable to guide the student to learn good embeddings, thereby affecting KD performance. These findings provide affirmation for our hypothesis.

We further demonstrate that this hypothesis aligns with the foundational principles of various existing works (Cho & Hariharan, 2019; Mirzadeh et al., 2020; Wang et al., 2022; Liang et al., 2024). In (Mirzadeh et al., 2020), assistant models are employed to narrow the gap between the capacities of the teacher and the student. Through our experiments, we observe that the calibration of assistant models is superior to that of teacher models. We also find an early-stopped teacher (Cho & Hariharan, 2019; Wang et al., 2022) is also well-calibrated. (Li et al., 2022) emphasizes on preserving the diversity of wrong-class logits. These can also be interpreted as enhancing the calibration of teacher models. Therefore, we propose a unified framework for analyzing capacity mismatch from the perspective of calibration.

To enable models with poor calibration to teach effectively, we explore the connection between distillation loss and calibration. Distillation based on KL divergence, where the student network directly mimics the teacher network, fails to alleviate the negative impact of poor calibration in large teacher models. Therefore, reducing the sensitivity of distillation loss to teacher calibration is beneficial. We adopt a representative ranking-based distillation method DIST (Huang et al., 2022) and compare it with traditional distillation methods. The experiments show that ranking-based distillation methods are insensitive to calibration and yield better distillation results than traditional methods.

Our contributions can be summarized as follows:

- We compare the KD performance achieved by teachers with different levels of calibration and observe that a better-calibrated teacher can teach better. These empirical findings substantiate the consistency between KD performance and calibration.

- We propose a unifying framework to explain previous works alleviating capacity mismatch, demonstrating that the commonality among them is the optimization of the teacher's calibration.

- We validate the superior performance of existing ranking-based KD methods and delve into their insensitivity to calibration. This observation underscores the role of ranking methods in mitigating capacity mismatch by reducing sensitivity to calibration.

2. Related Work

Similar to the concept of leveraging pre-trained models to aid training (Zhou, 2016), knowledge distillation (Hinton et al., 2015) allows a smaller model (student) to learn from a larger model (teacher) to enhance its performance. The idea of distilling knowledge can also date back to (Zhou & Jiang, 2004; Buciluǎ et al., 2006), and the smaller model can operate effectively on resource-constrained platforms as a substitute for the larger model.

The "dark knowledge" in the teacher model encompasses the information that the student needs to learn, and it can be categorized into three formats (Gou et al., 2020). Response-based knowledge (Hinton et al., 2015; Chen et al., 2017; Meng et al., 2019; Li et al., 2023c) typically pertains to the neural response of the last output layer of the teacher model. Feature-based knowledge (Romero et al., 2015; Huang & Wang, 2017; Zagoruyko & Komodakis, 2017) resides in intermediate layers, such as feature maps. Relation-based knowledge (Yim et al., 2017; Tian et al., 2020; Kweon et al., 2021) involves the relationships between different layers or data samples, such as the ranking of logits for different classes (Huang et al., 2022). These previous methods emphasize the representation of knowledge and the process of knowledge transfer, overlooking the nature of the teacher itself. We reevaluate these methods from the perspective of calibration and find that, conventional knowledge distillation methods exhibit a high dependence on the calibration level of the teacher. This dependency contributes to the issue of capacity mismatch.

Capacity mismatch refers to a counter-intuitive phenomenon when teacher model is much larger than student. According to conventional wisdom, larger models generally exhibit stronger generalization capabilities and are expected to teach better students. However, (Cho & Hariharan, 2019) found that as the teacher model continuously increases in size, the performance of the student network initially improves but eventually declines. Methods alleviating capacity mismatch can be categorized into 2 major classes. One aspect is from model architecture standpoint. (Mirzadeh et al., 2020) introduced an intermediate-size model as an assistant to mitigate the capacity gap between the large teacher and the student. (Zhu & Wang, 2021) divided the distillation task into multiple steps, distilling only the parts of the model where the gradient direction aligns with the cross-entropy loss. (Li et al., 2021) generated a new residual network after each distillation step and integrated the student network with the residual networks from each step. The other aspect is to regularize teacher. (Cho & Hariharan, 2019) observed that a fully trained teacher may not be an optimal instructor and proposed stopping the teacher's training from a specific checkpoint. Similarly, the work by (Wang et al., 2022) involved selecting an appropriate checkpoint of the teacher model through mutual information. (Kweon et al., 2021) showed that adjusting teacher through bidirectional distillation with the help of student can effectively cope with a large performance gap between teacher and student. The study by (Menon et al., 2021) closely aligns with our work. It identified that when the teacher provides true (Bayes) class-probabilities, it reduces the variance of the student objective, thereby reducing the upper bound of generalization error. The paper argues that accurate teachers do not necessarily provide good probabilities and elucidates why more accurate teachers can be detrimental to distillation (Müller et al., 2019). While this article provides inspiration for our work, it does not delve into the core reason of calibration. The study by (Zhu et al., 2023) adoped a calibration perspective, yet its emphasis is on recalibrating deep neural networks trained on distilled data, rather than addressing capacity mismatch in KD. For the first time, we approach the understanding of capacity mismatch from the perspective of calibration. We establish a unifying framework of the aforementioned works and show they implicitly correct the teacher's calibration.

3. Calibration Matters in KD

In this section, we first introduce the preliminaries of KD. Then, we construct an optimally calibrated teacher to demonstrate its superior teaching performance compared to a poorly calibrated teacher. Following that, we conduct quantitative experiments to illustrate the correlation between teacher's calibration and KD performance. Finally, we present t-SNE visualizations of students to depict how teacher's calibration influences student learning.

3.1. Preliminary

In a C-class classification task with labels $\mathcal{Y} = [C] = \{1, 2, \cdots, C\}$, we denote the output of a sample pair $\{\mathbf{x}, y\}$ by the teacher neural network as $\mathbf{f}_t(\mathbf{x})$, and $\mathbf{f}_s(\mathbf{x})$ for a student neural network. Upon applying these logits as inputs to the softmax layer, the resulting predictions are represented by $\mathbf{p}_t$ and $\mathbf{p}_s$. We use a temperature parameter τ in the softmax function and $* \in \{t, s\}$ to calculate $\mathbf{p}_t(\tau)$ and $\mathbf{p}_s(\tau)$:

$$\mathbf{p}_*(\tau) = \exp\left(\mathbf{f}_*(\mathbf{x})/\tau\right)/Z_*(\tau), \tag{1}$$

where $Z_*(\tau) = \sum_{j=1}^{C} \exp(f_{*,j}(\mathbf{x})/\tau)$. $f_{*,j}(\mathbf{x})$ is a scalar which refers to the logit of the j-th class in $\mathbf{f}_*(\mathbf{x})$.

In Vanilla Knowledge Distillation (KD) (Hinton et al., 2015), the objective for the student is to assimilate knowledge from the teacher by minimizing the Kullback-Leibler (KL) divergence between the predictions of the student ($\mathbf{p}_s$) and that of the teacher ($\mathbf{p}_t$). To account for potential inaccuracies in the teacher's predictions, cross-entropy loss (CE loss) between the target label and the student's output is also incorporated. Consequently, the total loss of Vanilla KD consists of two components, as expressed below:

$$\ell = (1-\alpha)\ell_{\text{CE}}(y, \mathbf{p}_s(1)) + \alpha\ell_{\text{KD}}(\mathbf{p}_t(\tau), \mathbf{p}_s(\tau)). \tag{2}$$

Here, the CE loss term utilizes the target label to directly refine the student model's output. To facilitate a closer match between the prediction of the student and the one-hot target label, the temperature in this part is set to 1.

Our focus is solely on exploring the correlation between teacher calibration and KD performance in this paper. Therefore, any component in the total loss that is unrelated to the property of teacher should be disregarded. In this context, the CE loss in Equation (2) is excluded by setting the value of α to 1 in our experiments.

Aside from accuracy, the reliability of a machine learning model's confidence in its predictions is also critical (Nixon et al., 2019). For example, a 90% posterior credible interval generally should contain the true outcome 90% of the time (Kuleshov et al., 2018). One mathematical formulation of the reliability of confidence is calibration (Nixon et al., 2019). A well-calibrated network should provide a calibrated confidence measure in addition to its prediction, that is, the probability associated with the predicted class label should reflect its ground truth correctness likelihood. In a multi-class classification task, the label space is $\mathcal{Y}$. Label of a sample is defined as $Y \in \mathcal{Y} = \{1, \ldots, K\}$. $\hat{Y}$ is a class prediction of this sample and $\hat{P}$ is its associated confidence, i.e. probability correctness. Perfect calibration is defined as (Guo et al., 2017):

$$\mathbb{P}(\hat{Y} = Y \mid \hat{P} = p) = p, \quad \forall p \in [0, 1] \tag{3}$$

However, $\hat{P}$ in (3) is continuous random variable. So $\mathbb{P}(\hat{Y} = Y \mid \hat{P} = p)$ cannot be computed directly. There are several measurements to approximate $\mathbb{P}(\hat{Y} = Y \mid \hat{P} = p)$, such as Reliability Diagrams, Expected Calibration Error (ECE) and Maximum Calibration Error (MCE). ECE is the most popular used measurements. To calculate ECE, predictions can be divided into M interval bins (each of size $1/M$) and calculate the accuracy of each bin. Let B_m be the set of indices of samples whose prediction confidence falls into the interval $I_m = (\frac{m-1}{M}, \frac{m}{M}]$. ECE computes a weighted average of this error across bins (Nixon et al., 2019):

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \Big| \text{acc}(B_m) - \text{conf}(B_m) \Big| \tag{4}$$

where n is the number of samples. $\text{acc}(B_m)$ and $\text{conf}(B_m)$ are the accuracy and confidence of bin B_m. Lower value of ECE leads to better calibration.

3.2. Optimal-calibrated Teacher Teaches Better

First, we compare the KD performance between a conventional teacher and an optimally calibrated teacher. Accord-

Table 1. Comparison of the generalization performance on the 10-class classification and 100-class classification task between models trained with one-hot labels and posterior distribution.

	10-class classification		100-class classification	
Test run	One-hot label	Posterior	One-hot label	Posterior
Test 1	87.06%	89.50%	91.50%	94.70%
Test 2	88.67%	89.63%	94.00%	94.73%
Test 3	88.77%	89.73%	94.37%	94.87%
Test 4	90.27%	89.93%	94.66%	94.90%
Test 5	90.87%	90.30%	95.47%	95.17%
Average	89.13%	**89.82%**	94.00%	**94.87%**
Variance	179%	**8%**	180%	**3%**

ing to the definition of calibration, a model is considered optimally calibrated if its output precisely matches the posterior distribution.

However, the posterior distribution is always inaccessible and intractable, with its functional formulation remaining unknown. Hence, we utilize a pre-configured GMM to generate an artificial known posterior distribution, as GMM can simulate any form of distribution by adjusting the parameters of individual Gaussian distributions. *The posterior distribution has the same form as the output of the teacher model, so the posterior distribution can be essentially regarded as a teacher model with optimal calibration.* A model trained by the posterior distribution can be viewed as a student distilled by an optimally calibrated teacher, assuming the training loss is also based on KL divergence. As a contrast, we train another identical model using one-hot labels. Being trained by one-hot labels can be considered as being distilled by a conventionally poorly calibrated teacher.

We simulate a 10-class classification task with two input dimensions. We construct a GMM with 10 categories, specifying 10 two-dimensional vectors as the means for each category, and defining covariance matrices to represent the relationships between these 10 categories. The GMM generates a dataset of 10,000 training samples along with their corresponding one-hot labels. We compute the posterior distribution for each training sample based on the specified Gaussian distribution. Following this, the GMM generates 3,000 test samples. We train the model using the calculated posterior distribution and compare its generalization performance with being trained by one-hot labels. Each training session consists of 200 epochs, and the learning rate is 0.03.

To mitigate the impact of random factors, we conducted five training runs for each training approach. Analogous experiments are conducted on a 100-class classification task. The experimental results are presented in Table 1.

From the table above, it is evident that training with the posterior distribution, indicating optimal model calibration, yields increased accuracy and stability on the test dataset. So teacher's output closely resembling one-hot distribution negatively impacts KD. This phenomenon can be elucidated by considering the influence of "dark knowledge" as proposed in (Hinton et al., 2015). For instance, in a 3-class classification task, there is an image labeled as "cat". Its corresponding posterior distribution might be (0.8, 0.15, 0.05) for the labels "cat", "dog" and "truck". The disparity between 0.15 and 0.05 signifies that a dog bears more resemblance to a cat than a truck. Model trained by one-hot labels tend to output (1, 0, 0) for that image. Although this model can achieve high accuracy, it might struggle to predict a new picture when distinguishing between a dog and a cat becomes challenging even for human observers. This difficulty arises because, when confronted with an image resembling both a cat and a dog, being trained by one-hot labels impedes the model from truly understanding why this specific image is a cat/dog rather than the other one.

3.3. Relationship between Calibration and KD Performance

After demonstrating that teachers with optimal calibration are more conducive to improving KD performance, we further delve into the relationship between calibration and KD performance in more specific experiments.

We employ CIFAR-10/CIFAR-100 (Krizhevsky et al., 2009) and Tiny-Imagenet (Tavanaei, 2020) datasets for our experiments. We use different versions of ResNet (He et al., 2016) and WideResNet (Zagoruyko & Komodakis, 2016) as teachers. Students are represented by models with lower capacity, specifically ResNet14, MobileNetV2 (Sandler et al., 2018), and WRN-22-1. The KD method is vanilla KD (Hinton et al., 2015).

Among various calibration metrics (Guo et al., 2017), we opt for Expected Calibration Error (ECE) (Naeini et al., 2015), the most widely employed measurement. It has been demonstrated that temperature scaling (TS) is a straightforward and effective calibration method, as highlighted in (Guo et al., 2017). Consequently, we employ temperature scaling to tune the calibration of teacher network.

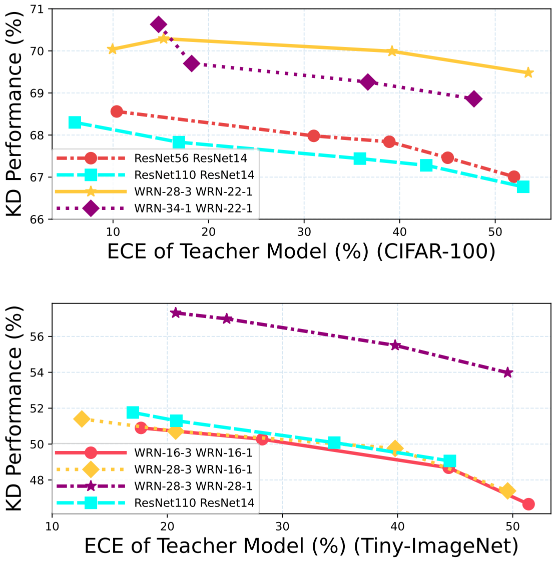

Figure 2. The distillation results for homogeneous teacher and student architectures on the CIFAR-100 and Tiny-ImageNet dataset. In each legend row in the figures, the first refers to the teacher model, and the second to the student model.

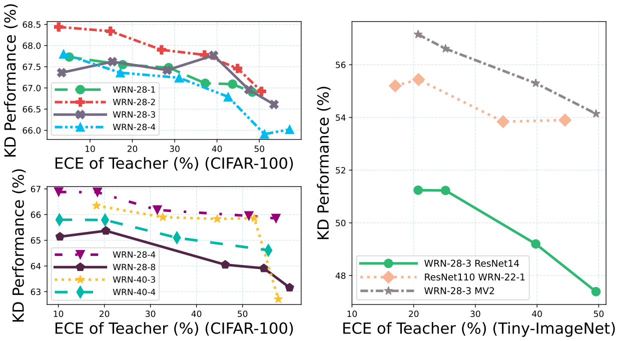

Figure 3. The distillation results for heterogeneous teacher and student architectures on the CIFAR-100 and Tiny-ImageNet dataset. In the left part, student in the upper is ResNet14. Student in the lower is MoblileNetV2. In each legend row in the right figure, the first refers to the teacher model, and the second to the student model.

Our exploration of the correlation between calibration and KD involves two main facets:

Given teacher-student pair The sole modification involves adjusting the temperature of KD to alter the calibration of the teacher model. We conduct experiments in both homogeneous scenarios (Figure 2) and heterogeneous scenarios (Figure 3) on CIFAR-100 and Tiny-ImageNet datasets. In homogeneous scenarios, both teacher and student models adopt either ResNet or WideResNet architectures. In heterogeneous scenarios, the teacher and student models differ in their architectures. We can see that calibration and KD performance exhibit the same changing trend.

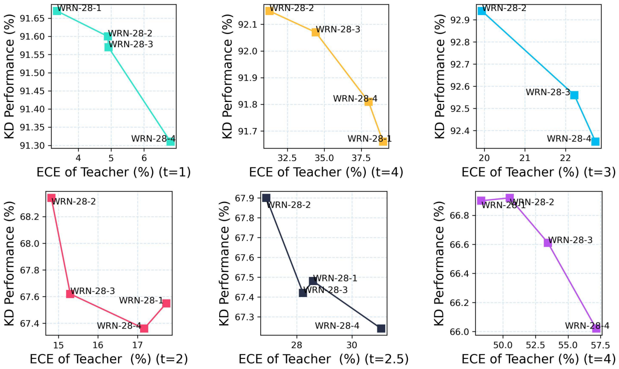

Figure 4. The comparison of KD performance by different-calibrated teachers with fixed temperature. In the upper figure, student is ResNet14 on CIFAR-100 dataset. The lower figure shows the results on CIFAR-10 dataset. Student in 3 subplots are ResNet14, ResNet14 and WRN-22-1. Names of teachers are shown in pictures.

Given temperature We fix the temperature and see KD performance of different teachers. Student model is also determined. This time calibration is changed by the change of teachers. We compare the KD performance on dataset CIFAR-100 and CIFAR-10. Figure 4 demonstrates that as the calibration deteriorates with different teacher, KD performance declines.

From the results of the above experiments, where calibration is adjusted through two methods, we observe a positive correlation between KD performance and teacher calibration.

In addition to response-based KD methods like vanilla KD, we also investigate the relationship between feature-based KD methods and calibration. We select the FitNet method (Romero et al., 2015) and conduct comparative experiments on CIFAR-100. The results, which are shown in the appendix, indicate that our findings also apply to feature-based knowledge.

3.4. t-SNE Visualization of Student

We demonstrate how different levels of calibration in teacher models affect KD performance. We employ t-SNE visualization to showcase the classification results of the trained student model for each category. Here, we choose WRN-28-3 as the teacher model and ResNet14 as student. The student's classification results on the CIFAR-10 dataset is depicted in Figure 5. From the visualization results, we can observe that when the calibration of the teacher model deteriorates, although the inter-class distances are slightly larger, some confusing samples that are similar to several classes are not easily distinguishable. In the right panel of Figure 5, there are a certain number of undistinguished samples among the green, purple, yellow, and pink categories. These samples are so many and even can be considered as an additional 11th category. This result indicates that when the calibration of the teacher model is poor, the distillation process fails to teach the student to distinguish confusing categories effectively. The student model is unable to construct a robust embedding, leading to suboptimal distillation performance.

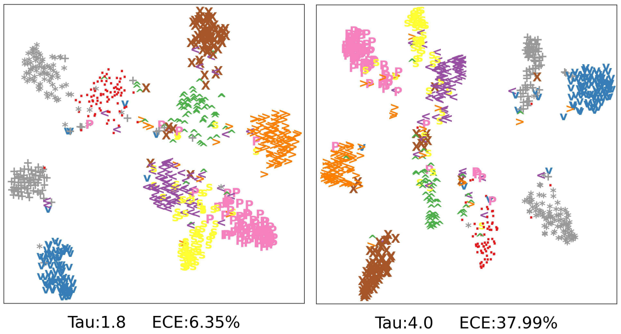

Figure 5. The t-SNE visualization results of the features of student models distilled by teachers with different levels of calibration. Different colors and markers represent different categories.

4. Unifying Explanation to Previous Works

The forthcoming section aims to illustrate that previous works (Mirzadeh et al., 2020; Cho & Hariharan, 2019; Zhu & Wang, 2021; Li et al., 2021) alleviating capacity mismatch can be cohesively explicated by examining the interplay between calibration and KD.

As discussed in Section 2, (Mirzadeh et al., 2020) and (Cho & Hariharan, 2019) represent two methods addressing capacity mismatch from two major perspectives (architecture adjustment and teacher regularization). The justification of explaining these methods through calibration can be applied to explaining similar approaches.

4.1. Architecture-based Method: TAKD

Introduction of TAKD Mirzadeh et al. (Mirzadeh et al., 2020) raised Teacher Assistant Knowledge Distillation (TAKD), employing assistant models to tackle the problem of capacity mismatch. The size of the assistant model lies between that of the teacher model and the student model. TAKD involves distilling knowledge from the teacher network to the assistant network first, followed by distilling knowledge from the assistant network to the student network. They contend that the incorporation of assistant models can reduce the upper bound of error in KD.

We propose an alternative explanation for the effectiveness of the TAKD method. In TAKD, the assistant model distilled by teacher is utilized to distill student. Calibration of the assistant model may be better than teacher model trained solely by one-hot labels. Consequently, the assistant model instructing the student model can be interpreted as a smaller model with better calibration taking on the role of the teacher model. A question naturally arises: *Is the success of TAKD due to the smaller size of the assistant model or the better calibration of the assistant model?*

We conduct experiments to investigate the success factor of TAKD. We fix the distillation temperature to 1.8 on the CIFAR-100 dataset, so calibration of assitant model is changed only by its size. To align with scenarios involving capacity mismatch, we employ larger teacher networks, such as WRN-28-3 and WRN-28-4, as the teacher model. Conversely, the student network utilizes a smaller model, such as ResNet14 and WRN-22-1. Initially, the teacher model instructs the assistant model, and we employ the Expected Calibration Error (ECE) as an evaluation metric to gauge the calibration of the assistant model on the test dataset. After being trained by teacher, the ECE of assistant model is recorded. The results are presented in Table 2. To mitigate experimental variability, we employ fixed sets of diverse random seeds and conducted multiple comparative experiments using the same set of teacher-assistant-student models in the appendix.

Firstly, we observe a correlation between the TAKD performance and the calibration of the assistant model. Subsequently, we note that as the assistant model increases in size, its calibration improves, resulting in enhanced TAKD performance. This strongly underscores that the success of TAKD is attributed to the improved calibration of the assistant model, rather than its size.

We also perform experiments on the Tiny-ImageNet dataset (results shown in Table 3). In this case, we keep the teacher-assistant-student pair fixed. We manipulate the distillation temperature to induce significant changes in the calibration of the assistant model and observe whether there are substantial variations in the TAKD performance. The results confirm our hypothesis.

Based on the aforementioned experimental results, it is evident that the performance of TAKD improves with the enhancement of the assistant's calibration. At this point, we can explain why TAKD can alleviate capacity mismatch based on calibration: *The TAKD process comprises multiple stages, involving several rounds of assistant distillation followed by the ultimate stage of student network distillation. The former stages aim to boost the performance of the better-calibrated assistant model in the final stage.*

Table 2. Results of TAKD performance by assistant models varied on size on CIFAR-100 dataset.

Teacher	Assistant	Size of Assistant	Student	ECE of assistant	KD performance
WRN-28-3	ResNet110	1.7M	ResNet14	**9.57%**	**68.44%**
WRN-28-3	WRN-28-2	1.5M	ResNet14	9.94%	67.96%
WRN-28-3	WRN-28-1	0.38M	ResNet14	9.92%	67.97%
ResNet110	ResNet56	0.86M	ResNet14	**9.24%**	**69.00%**
ResNet110	ResNet32	0.47M	ResNet14	9.62%	68.13%
ResNet110	ResNet20	0.28M	ResNet14	9.92%	68.12%
WRN-28-3	WRN-28-2	1.5M	WRN-22-1	**10.34%**	**71.01%**
WRN-28-3	WRN-28-1	0.38M	WRN-22-1	11.08%	70.36%

Table 3. Results of TAKD performance by different calibration of fixed assistant models on dataset Tiny-ImageNet.

Teacher	Assistant	Student	Tau	ECE of assistant	KD performance
WRN-16-3	ResNet32	WRN-16-1	1.8	23.5%	49.66%
WRN-16-3	ResNet32	WRN-16-1	4.0	48.03%	46.04%
WRN-28-3	WRN-28-2	WRN-28-1	1.8	15.55%	56.45%
WRN-28-3	WRN-28-2	WRN-28-1	4.0	50.13%	52.87%

4.2. Teacher Regularization: ESKD

Introduction of ESKD (Cho & Hariharan, 2019) initially highlights the detrimental impact of a large teacher network on instructing the student. They posit that a large teacher network, when trained for only a few epochs, exhibits behavior akin to a smaller network. In response, they propose the early-stopped method, where a large teacher network is trained for only a few epochs (referred to as an early-stopped teacher) and subsequently employed for distillation. This approach is abbreviated as ESKD (Early-Stopped Knowledge Distillation). It needs to be supplemented that, the authors proposed two approaches to enhance KD performance in (Cho & Hariharan, 2019). One involves discontinuing the KD loss after a certain number of distillation epochs, utilizing only the target label for guidance in subsequent teaching. The other approach advocates early termination of teacher training. Since we focus on correlation between ESKD and calibration of teacher, we omit the discussion of the former method, as it leaves the calibration of the teacher model unchanged.

The efficacy of ESKD can also be attributed to the effectiveness of calibration. The calibration of the early-stopped teacher is not inferior and may even be superior to that of the fully trained teacher. To conduct the relative experiments, we save training checkpoints of the teacher model every 10 epochs on CIFAR-100 dataset. Subsequently, we select the trained teacher models at several checkpoints and compare the performance of students taught by them. Results are in Table 4.

Table 4. The ECE value and the KD performance of a teacher network in training checkpoints.

Teacher	Student	Tau	Checkpoint	ECE	KD
WRN-16-3	WRN-16-1	1.0	140	7.26%	67.12%
			200	7.2%	67.31%
WRN-16-3	WRN-16-1	1.6	140	7.91%	68.43%
			200	7.89%	68.01%
WRN-28-3	WRN-28-1	1.0	130	8.47%	71.09%
			200	9.65%	69.67%
WRN-28-3	WRN-28-1	1.6	140	4.96%	71.59%
			200	5.16%	71.71%

It is evident that teachers trained with fewer epochs can achieve comparable performance and calibration to the fully trained teacher. This observation suggests that an extensive training duration for the teacher model might be unnecessary, as numerous epochs contribute minimally to both accuracy enhancement and calibration refinement. Consequently, we opt to train the teacher model for only 100 epochs, initiating the learning rate decay after the 50th epoch. We use TS to tune calibration of these teachers and compare their KD performance. From the results in Table 5, we can see the calibration and KD performance of such a short-trained teacher is a little superior to a full-trained teacher. More training results are in appendix.

Here, we can explain the effectiveness of regularization methods for teacher models, such as ESKD, from the perspective of model calibration: *These methods optimize the calibration of large teacher models through techniques such as early stopping and selecting checkpoints, imparting more accurate knowledge to the students.*

Table 5. The comparison between teacher trained with fewer epochs and full-trained teacher. T refers to temperature.

	FEW-TRAINED	FULL-TRAINED
NAME	WRN-16-3	WRN-16-3
CLASSIFICATION	74.28%	**75.17%**
ECE(T=1.6)	**7.48%**	7.89%
KD(T=1.6)	**68.62%**	68.01%
NAME	WRN-28-3	WRN-28-3
CLASSIFICATION	75.58%	**76.83%**
ECE(T=1.6)	**3.48%**	5.16%
KD(T=1.6)	**71.95%**	71.71%

4.3. Explanation to Other Methods

The first two sections elucidated two representative methods for mitigating capacity mismatch. Other relative methods can be interpreted through the lens of calibration. For instance, in Residual KD (Li et al., 2021), the authors employed various residual-student models, employing a training process akin to TAKD. (Liang et al., 2024) applied TAKD to self distillation, which shares the same principle of TAKD in alleviating capacity mismatch. We have previously explored the correlation between TAKD and calibration, so a similar connection with calibration can be inferred for the above-mentioned work. (Wang et al., 2022) used information bottleneck to find the suitable checkpoint of the teacher model. The effectiveness of searching for checkpoints has been discussed in Section 4.2. Many other methods share a similar attribution to calibration, but due to space constraints, we will not delve into them here.

5. Fundamental Approach to Alleviate Capacity Mismatch

The correlation between calibration of teacher model and KD performance arises from the fact that most existing KD methods require the student to closely mimic the teacher. Distance measurements, such as KL divergence, are commonly used to precisely align the outputs of the student and teacher. So when capacity mismatch happens, the teacher model is too large that its calibration may not be so good (Guo et al., 2017). The calibration of the student network will also be adversely affected. Hence the student has poor generalization ability and KD performance is bad.

It is practical to raise KD methods which are not sensitive to calibration to solve capacity mismatch. We attempt to relax the training requirements for the student, not insisting on an exact match with the teacher output. Instead, we only focus on some key indicators to align with the teacher, such as maintaining the order of likelihood probabilities across all classes, which is achieved through ranking-based KD.

We have designed ablation experiments to show ranking-based KD method is not sensitive to calibration as KL-based KD (distance measurement is KL divergence). We select DIST (Huang et al., 2022) as a representative of ranking-based methods. The CE loss in DIST is also dropped and we only use inter-class and intra-class relation loss. DIST focuses on ensuring consistency in the class relationships between teacher and student model outputs. We take the inter-relation loss in DIST as an example to demonstrate how DIST calculates loss:

Here we define Y_i as the i-th sample in dataset. $Y_i^{(s)}$ is the student output of Y_i, while $Y_i^{(t)}$ is the teacher output of Y_i. When the batch size is B, the inter-relation loss in DIST is formulated as: (Huang et al., 2022)

$$L_{\text{inter}} := \frac{1}{B} \sum_{i=1}^{B} d_{\mathrm{p}} \left(Y_i^{(\mathrm{s})}, Y_i^{(\mathrm{t})} \right) \tag{5}$$

where $d_{\mathrm{p}}(\boldsymbol{u}, \boldsymbol{v}) := 1 - \rho_{\mathrm{p}}(\boldsymbol{u}, \boldsymbol{v})$, here $\rho_{\mathrm{p}}(\boldsymbol{u}, \boldsymbol{v})$ is the Pearson correlation coefficient between $\boldsymbol{u}$ and $\boldsymbol{v}$:

$$\begin{aligned} \rho_{\mathrm{p}}(\boldsymbol{u}, \boldsymbol{v}) &:= \frac{\operatorname{Cov}(\boldsymbol{u}, \boldsymbol{v})}{\operatorname{Std}(\boldsymbol{u}) \operatorname{Std}(\boldsymbol{v})} \\ &= \frac{\sum_{i=1}^{C} (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{C} (u_i - \bar{u})^2 \sum_{i=1}^{C} (v_i - \bar{v})^2}} \end{aligned} \tag{6}$$

Combining (5) and (6), it is evident that the DIST method prioritizes the comparison of the consistency in output probability rankings for each category between the teacher network and the student network.

We manipulate the two types of KD methods on CIFAR-100 dataset with a wide range of temperature. By calculating Spearman correlation (Dodge, 2008) between ECE and KD performance, it is easy to compare the sensitivity of each KD method to calibration. The detailed results are shown in Table 6 and more results are displayed in appendix.

From the results we can see that when applying DIST, the absolute value of the Spearman correlation coefficient is smaller than that of the KL-based KD. Thus, KD performance of ranking-based KD is not sensitive to calibration as KL-based KD does. We also visualize the relevant experimental data and fit a line on the graph to show the relationship between KD performance and calibration, as depicted in Figure 6. As we can see, the curve's slope for results of vanilla KD approaches ±1, indicating that the performace of the KL-based method varies more significantly with increasing or decreasing calibration. In contrast, the curve for the ranking-based method is relatively flat, suggesting that the KD performance of the ranking-based method is insensitive to calibration. The above experimental results and

Table 6. Some experimental results to show the sensitivity of KD with different measurements with calibration

TEACHER	STUDENT	RESULTS (TOP-1 ACC (%))												SPEARMAN (%)
RESNET-56	RESNET-14	TAU	1.0	1.5	2.0	2.5	3.0	3.5	4.0	5.0	6.0	7.0	8.0	-
		ECE	12.04	1.79	10.39	21.33	31.01	38.92	45.03	51.97	57.04	59.92	61.65	-
		KD	68.09	68.28	68.56	68.57	67.98	67.84	67.46	67.01	67.20	66.75	66.80	-91.82
		DIST	67.35	68.19	68.77	68.63	68.92	68.41	68.41	68.9	68.93	68.39	66.97	0.46
RESNET-110	RESNET-14	TAU	1.0	1.5	2.0	2.5	3.0	3.5	4.0	5.0	6.0	7.0	8.0	-
		ECE	14.08	4.70	6.00	16.89	27.12	35.84	42.78	52.93	57.22	59.65	61.12	-
		KD	67.18	67.72	68.30	67.83	68.42	67.44	67.28	66.77	66.87	67.15	65.98	-76.36
		DIST	67.23	68.09	68.20	68.74	68.15	68.88	68.65	68.74	69.35	67.88	67.18	1.82
WRN-28-1	RESNET-14	TAU	1.0	1.5	2.0	2.5	3.0	3.5	4.0	5.0	6.0	7.0	8.0	-
		ECE	8.70	5.09	17.73	28.59	37.22	43.67	48.34	54.06	57.07	58.78	59.83	-
		KD	67.76	67.73	67.55	67.48	67.11	67.09	66.9	66.82	66.64	66.28	65.68	-99.10
		DIST	67.22	68.26	68.19	68.48	68.84	68.35	68.12	67.99	67.79	67.36	66.33	-50.00

the success of DIST demonstrate that *selecting distillation methods insensitive to calibration is crucial in alleviating capacity mismatch.* This choice can assist in achieving good distillation performance with a given large teacher network.

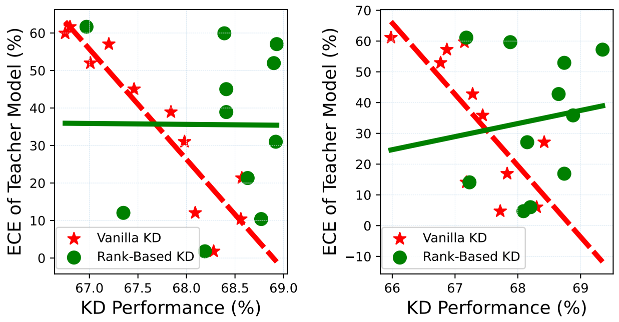

Figure 6. Visualization of KD with different measurements. Green scatters represent ranking-based KD while red scatters represent KL-based KD. Curve slope can show if measurements is related to calibration closely.

6. Conclusion

In this paper, we provide a novel perspective on capacity mismatch through calibration. We observe a significant impact of the calibration of the teacher model on student performance in our experiments. To alleviate the sensitivity to calibration, we propose employing calibration-insensitive metrics, such as ranking-based loss, as substitutes for KL divergence in the KD loss. Our findings suggest that these calibration-insensitive metrics contribute to enhancing the teaching effectiveness of large teachers.

We contend that our interpretation, rooted in model calibration, is both comprehensive and robust. We introduce a unifying framework to systematically summary previous research on capacity mismatch. We hope that our work can cast light on future relevant studies.

Acknowledgements

This work is supported by National Science and Technology Major Project (Grant No. 2022ZD0114805). Professor De-Chuan Zhan is the corresponding author.

Impact Statement

The present paper is a pure technical study, analyzing the problem from a new and unified perspective. We believe it does not raise issues related to social morality or ethics.

References

Buciluǎ, C., Caruana, R., and Niculescu-Mizil, A. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006.

Chen, G., Choi, W., Yu, X., Han, T. X., and Chandraker, M. Learning efficient object detection models with knowledge distillation. In *Advances in Neural Information Processing Systems 30*, pp. 742–751, 2017.

Cho, J. H. and Hariharan, B. On the efficacy of knowledge distillation. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pp. 4793–4801, 2019.

Dodge, Y. *The concise encyclopedia of statistics*. Springer Science & Business Media, 2008.

Gou, J., Yu, B., Maybank, S. J., and Tao, D. Knowledge distillation: A survey. *CoRR*, abs/2006.05525, 2020.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, pp. 1321–1330, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Hinton, G. E., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

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

Huang, Z. and Wang, N. Like what you like: Knowledge distill via neuron selectivity transfer. *CoRR*, abs/1707.01219, 2017.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 2801–2809, 2018.

Kweon, W., Kang, S., and Yu, H. Bidirectional distillation for top-k recommender system. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pp. 3861–3871, 2021.

Li, C., Cheng, G., and Han, J. Boosting knowledge distillation via intra-class logit distribution smoothing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023a.

Li, J., Guo, Z., Li, H., Han, S., Baek, J.-w., Yang, M., Yang, R., and Suh, S. Rethinking feature-based knowledge distillation for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20156–20165, 2023b.

Li, X., Li, S., Omar, B., Wu, F., and Li, X. ResKD: Residual-guided knowledge distillation. *IEEE Trans. Image Process.*, 30:4735–4746, 2021.

Li, X., Fan, W., Song, S., Li, Y., Li, B., Shao, Y., and Zhan, D. Asymmetric temperature scaling makes larger networks teach well again. In *Advances in Neural Information Processing Systems 35*, 2022.

Li, X., Yang, Y., and Zhan, D. MrTF: model refinery for transductive federated learning. *Data Min. Knowl. Discov.*, 37:2046–2069, 2023c.

Liang, P., Zhang, W., Wang, J., and Guo, Y. Neighbor self-knowledge distillation. *Information Sciences*, 654: 119859, 2024.

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Meng, Z., Li, J., Zhao, Y., and Gong, Y. Conditional teacher-student learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 6445–6449, 2019.

Menon, A. K., Rawat, A. S., Reddi, S. J., Kim, S., and Kumar, S. A statistical perspective on distillation. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, pp. 7632–7642, 2021.

Mirzadeh, S., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., and Ghasemzadeh, H. Improved knowledge distillation via teacher assistant. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pp. 5191–5198, 2020.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In *Advances in Neural Information Processing Systems 32*, pp. 4696–4705, 2019.

Naeini, M. P., Cooper, G. F., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 38–41, 2019.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., and Bengio, Y. Fitnets: Hints for thin deep nets. In *3rd International Conference on Learning Representations, ICLR*, 2015.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, pp. 6105–6114, 2019.

Tavanaei, A. Embedded encoder-decoder in convolutional networks towards explainable ai. *arXiv preprint arXiv:2007.06712*, 2020.

Tian, Y., Krishnan, D., and Isola, P. Contrastive representation distillation. In *8th International Conference on Learning Representations, ICLR*, 2020.

Wang, C., Yang, Q., Huang, R., Song, S., and Huang, G. Efficient knowledge distillation from model checkpoints. In *Advances in Neural Information Processing Systems 35*, 2022.

Yim, J., Joo, D., Bae, J., and Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *Proceedings of the British Machine Vision Conference, BMVC*, 2016.

Zagoruyko, S. and Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *5th International Conference on Learning Representations, ICLR*, 2017.

Zhang, X., Zhou, X., Lin, M., and Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Zhao, B., Cui, Q., Song, R., Qiu, Y., and Liang, J. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11953–11962, 2022.

Zhou, Z. Learnware: on the future of machine learning. *Frontiers Comput. Sci.*, 10(4):589–590, 2016.

Zhou, Z. and Jiang, Y. Nec4.5: Neural ensemble based C4.5. *IEEE Trans. Knowl. Data Eng.*, 16(6):770–773, 2004.

Zhu, D., Fang, Y., Lei, B., Xie, Y., Xu, D., Zhang, J., and Zhang, R. Rethinking data distillation: Do not overlook calibration. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pp. 4912–4922, 2023.

Zhu, Y. and Wang, Y. Student customized knowledge distillation: Bridging the gap between student and teacher. In *IEEE/CVF International Conference on Computer Vision, ICCV*, pp. 5037–5046, 2021.

A. Appendix

A.1. Results of corrlation between calibration and feature-based KD performance

Figure 7 demonstrates that for feature-based KD methods, teacher's calibration is also correlated to KD performance.

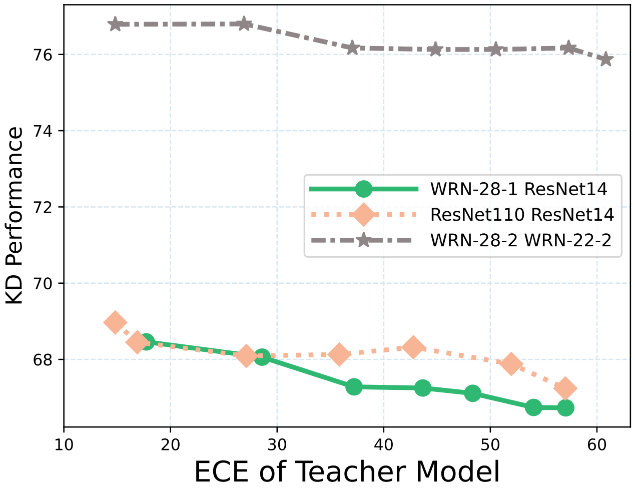

Figure 7. KD performance by feature-based KD methods on the CIFAR-100 dataset. In each legend row, the first refers to the teacher model, while the second pertains to the student model.

A.2. Results of TAKD with different random seeds

Table 7. Results of TAKD performance with different assitant models. The random seed is fixed to several predefined values to reduce experimental randomness.

Teacher	Assistant	Student	ECE of assistant	KD performance	Random seed
WRN-28-4	WRN-28-2	WRN-22-1	**9.99%**	**70.52%**	0
WRN-28-4	WRN-28-1	WRN-22-1	10.20%	70.30%	0
WRN-28-4	WRN-28-2	WRN-22-1	**10.28%**	**70.77%**	1
WRN-28-4	WRN-28-1	WRN-22-1	10.60%	70.14%	1
WRN-28-4	WRN-28-2	WRN-22-1	**9.70%**	**70.14%**	2
WRN-28-4	WRN-28-1	WRN-22-1	10.57%	70.04%	2
WRN-28-4	WRN-28-2	WRN-22-1	**10.03%**	**70.32%**	5
WRN-28-4	WRN-28-1	WRN-22-1	10.32%	70.34%	5
WRN-28-4	WRN-28-2	ResNet14	**10.06%**	**68.48%**	1
WRN-28-4	WRN-28-1	ResNet14	10.45%	67.93%	1
WRN-28-4	WRN-28-2	ResNet14	**9.82%**	**69.16%**	2
WRN-28-4	WRN-28-1	ResNet14	10.38%	68.83%	2

A.3. More results of early-stopped teacher

Table 8. The detailed comparison between teacher trained with fewer epochs and full-trained teacher. T refers to temperature.

	FEW-TRAINED	FULL-TRAINED
NAME	WRN-16-3	WRN-16-3
CLASSIFICATION	74.28%	**75.17%**
ECE(T=1.6)	**7.48%**	7.89%
KD(T=1.6)	**68.62%**	68.01%
ECE(T=3.0)	**37.56%**	39.90%
KD(T=3.0)	**67.77%**	67.54%
ECE(T=4.0)	**50.27%**	51.56%
KD(T=4.0)	**67.35%**	67.01%
NAME	WRN-28-3	WRN-28-3
CLASSIFICATION	75.58%	**76.83%**
ECE(T=1.6)	**3.48%**	5.16%
KD(T=1.6)	**71.95%**	71.71%
ECE(T=3.0)	**37.74%**	41.76%
KD(T=3.0)	**71.58%**	71.39%
ECE(T=4.0)	**52.29%**	53.46%
KD(T=4.0)	**71.23%**	71.11%

A.4. More results of sensitivity comparison

Table 9. The detailed results of sensitivity of KD with different measurements with calibration.

Teacher	Student	Spearman (KD and ECE)	Spearman (DIST and ECE)
WRN-28-2	WRN-22-1	-31.82%	**8.18%**
WRN-28-4	MobileNetV2	-38.18%	**8.18%**