Debiased Distillation for Consistency Regularization

Lu Wang [1], Liuchi Xu [1], Xiong Yang [2], Zhenhua Huang* [3], Jun Cheng* [4,5]

[1] School of Computer Science and Engineering, Northeastern University, Shenyang, China;
[2] Institute of Applied Artificial Intelligence of the Guangdong-Hong Kong-Macao
Greater Bay Area, Shenzhen Polytechnic University, Shenzhen, China;
[3] School of Computer Science, South China Normal University, Guangzhou, China;
[4] Guangdong Provincial Key Laboratory of Robotics and Intelligent System,
Shenzhen Institute of Advanced Technology, CAS, China;
[5] The Chinese University of Hong Kong, Hong Kong, China
[1] {wanglu@mail,xuliuchi@stumail}.neu.edu.cn, [2] 2018021011@cauc.edu.cn,
[3] huangzhenhua@m.scnu.edu.cn, [4,5] Jun.cheng@siat.ac.cn

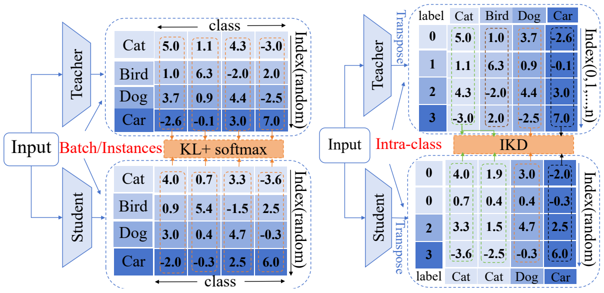
Abstract

Knowledge distillation transfers "dark knowledge" from a large teacher model to a smaller student model, yielding a highly efficient network. To improve the network's generalization ability, existing works use a larger temperature coefficient for knowledge distillation. Nevertheless, these methods may reduce the confidence of the target category and lead to ambiguous recognition of similar samples. To mitigate this issue, some studies introduce intra-batch distillation to reduce prediction discrepancy. However, these methods overlook the inconsistency between background information and the target category, which may increase prediction bias due to noise disturbance. Additionally, label imbalance from random sampling and batch size can undermine network generalization reliability. To tackle these challenges, we propose a simple yet effective Intra-class Knowledge Distillation (IKD) method that facilitates knowledge sharing within the same class to ensure consistent predictions. First, we initialize the matrix and the vector to store logits and class counts provided by the teacher, respectively. Then, in the first epoch, we calculate the sum of logits and sample counts per class and perform KD to prevent knowledge omission. Finally, in subsequent training, we update the matrix to obtain the average logits and compute the KL divergence between the student's output and the updated matrix according to the label index. This process ensures intra-class consistency and improves the student's performance. Furthermore, this method theoretically reduces prediction bias by ensuring intra-class consistency. Extensive experiments on the CIFAR-100, ImageNet-1K, and Tiny-ImageNet datasets validate the superiority of IKD.

Code — https://github.com/yema-web/IKD

Introduction

Deep convolutional neural networks (CNNs) have recently achieved significant success in various computer vision tasks, including image classification (He et al. 2016; Yang et al. 2022a), object detection (Zheng et al. 2023; Zhao et al. 2024), and semantic segmentation (Zou et al. 2024; Liang et al. 2023). While there have been advancements, large-scale CNNs, such as ConvNeXt (Woo et al. 2023), still consume considerable memory and computational overhead, leading to increased training costs and prolonged inference time. In response to these challenges, researchers have explored model compression methods to balance between network accuracy and inference time. Such methods include efficient network design (Zhang et al. 2018a), network pruning (Qin et al. 2024), low-rank factorization (Haeffele and Vidal 2019), quantization (Duan et al. 2023), and knowledge distillation (KD) (Hinton et al. 2014). Among them, KD can effectively deploy lightweight networks on resource-limited devices.



(a) Intra-batch distillation (b) Intra-class distillation

Figure 1: Illustration and comparison of (a) traditional intra-batch distillation methods, including DIST, MLKD, and LCAT, and (b) our method, IKD.

Hinton et al. have formulated KD as an efficient model compression method that transfers the "dark knowledge" from a cumbersome teacher model (teacher) to a compact student model (student). Generally, KD employs the Kullback-Leibler (KL) divergence (Kullback and Leibler 1951) with a temperature coefficient to minimize the discrepancy between the probability distributions of the teacher and student. KD can effectively boost the student's generalization ability on unseen data. Furthermore, some studies have delved deeper into the working mechanism of KD, particularly from the perspective of soft labels (Tang et al. 2020; Helong et al. 2021). Despite these insights, the performance gains from conventional KD methods, such as deep mutual

*Corresponding authors

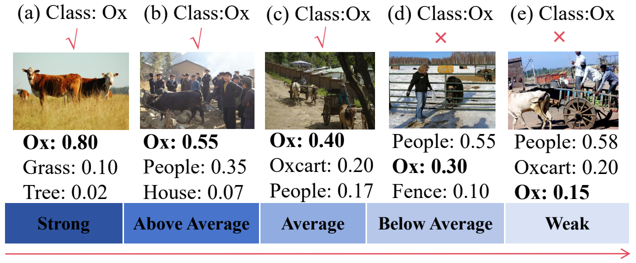


Figure 2: The logit-based KD methods predicts target categories on the ImageNet-1K dataset, ranking the predictions by confidence from strong to weak.

Batch Size	32	64	128	256	Max Gap
DIST	**76.47**	76.04	75.27	74.33	2.14
MLKD	**74.77**	74.63	74.56	74.05	0.72
IKD (Ours)	**76.75**	76.56	76.13	76.04	0.71

Table 1: Illustration of the impact of batch size on the CIFAR-100 test accuracy, with ResNet32×4 and ResNet8×4 being the teacher and student, respectively.

learning (Zhang et al. 2018b), teacher assistants (Mirzadeh et al. 2020; Son et al. 2021), BAN (Furlanello et al. 2018), and early stopping (Cho and Hariharan 2019), may still not meet practical application requirements.

To achieve greater performance gain, recent studies have shifted their focus to extracting knowledge from the logit effectively, which provides essential semantic information for accurate prediction. Compared to feature-based KD methods (Xiaolong et al. 2023; Zong et al. 2023; Romero et al. 2015; Miles and Mikolajczyk 2024; Shen et al. 2024), logit-based KD methods can reduce computational costs and enhance practical feasibility. These advantages have facilitated detailed research investigations into this method (Yang et al. 2024; Wei, Luo, and Luo 2024). For example, Zhao et al. (Zhao et al. 2022) propose decoupling the classical KD (DKD), effectively leveraging logit knowledge to enhance the student's performance. Yang et al. (Yang et al. 2023) present the NKD method, which normalizes non-target logits to strengthen the utilization of soft labels. LSKD (Sun et al. 2024) adaptively allocates temperatures between the teacher and student and across samples for logit distillation. Although logit-based KD methods acquire universal knowledge to improve generalization ability, the introduced noise (e.g., low-confidence categories) may reduce the discrimination between classes, leading to biased predictions.

To mitigate noise-induced classification issues, some researchers have focused on enforcing consistent predictions. For example, CSKD (Yun et al. 2020) utilizes self-distillation (Zhang et al. 2019) to compare prediction distributions between two samples of the same class, and thereby reduce inconsistencies. DIST (Huang et al. 2022) randomly samples instances within batches to reveal their varying semantic similarities. Despite these efforts, intra-batch distillation, depicted in Fig. 1 (a), still faces several challenges. As shown in Figs. 2 (d) and (e), occlusions, such as fences and incomplete head features of the target class, can obscure essential features, leading to predictions that deviate from the target. Furthermore, as illustrated in Table 1, batch size variations, which impact random sampling, may further impair the performance of intra-batch distillation. However, these methods may increase prediction bias, which can not address the inconsistency. In fact, achieving consistency within the same class is crucial for reducing prediction bias. Therefore, this inconsistency underscores the urgent need for innovative solutions.

In response to these critical challenges, we revisit knowledge distillation from the perspectives of intra-class distillation through theoretical analysis, and propose a simple yet effective **I**ntra-class **K**nowledge **D**istillation method, as shown in Fig. 1 (b), named **IKD**, that facilitates knowledge sharing within the same class to ensure consistent predictions. Specifically, we use a matrix to store the average logits and a vector to count samples from the teacher for each class. In the first epoch, we calculate the cumulative sum of logits and sample counts per class, and conduct vanilla KD to prevent knowledge omission. In subsequent training, we update the matrix to obtain the average logits and compute the KL divergence between the student's output and the updated matrix according to the label index. The proposed method can be integrated with the logit-based KD methods to further improve performance by enhancing prediction consistency. As a result, the student can better recognize ambiguous samples by minimizing the noise and discerning distinctions and similarities within the same category. Our main contributions are summarized as follows:

- We propose an Intra-class Knowledge Distillation (IKD) method as a regularization technique to ensure prediction consistency while reducing noise interference.
- We have demonstrated that our method effectively reduces prediction bias through theoretical analysis. Moreover, theoretical derivations suggest that IKD effectively reduces the complexity of generalization.
- Extensive experiments on CIFAR-100, ImageNet-1K, and Tiny-ImageNet consistently demonstrate that our method surpasses existing plug-and-play strategies when integrated with the state-of-the-art logit-based KD method, thereby validating its effectiveness.

Related Work

In this section, we will review and discuss existing works that relate to feature- and logit-based knowledge distillation.

Feature-based Knowledge Distillation. Several studies have shown that intermediate layer features contain abundant information, which may facilitate the learning of downstream tasks (Yang et al. 2022a,b; Zhao et al. 2024). Consequently, these features are typically used as targets for knowledge distillation. For instance, FitNets (Romero et al. 2015) uses intermediate features from the teacher as hints to guide the student's training. AT (Komodakis and Zagoruyko 2017) employs activation maps derived from intermediate features to enhance the knowledge transfer to the student. CRD (Tian, Krishnan, and Isola 2020) introduces contrastive learning to improve knowledge transfer from the teacher to the student. ReviewKD (Chen et al. 2021) guides shallow learning by repeatedly reviewing old knowledge. FCFD (Liu et al. 2023) promotes feature similarity between teacher and student, enabling more faithful imitation and effective learning by the student. CAT-KD (Guo et al. 2023) guides the student by transferring the teacher's class activation maps. However, although feature-based knowledge distillation is information-rich, it demands substantial computational resources and memory capacity. As a result, recent research has shifted towards exploring the logit knowledge as a more efficient target for distillation.

Logit-based Knowledge Distillation. Recent studies have focused on dynamic temperature and decoupling the logit knowledge to improve the effectiveness of this distillation method.

1) Dynamic temperature: Traditional KD uses a fixed temperature, limiting the potential for logit knowledge transfer. To remove this limitation, dynamic temperature has been introduced to improve this transfer (Sun et al. 2024; Jin, Wang, and Lin 2023). For example, CTKD (Li et al. 2023) employs a dynamically learnable temperature to adjust the task difficulty. MLKD (Jin, Wang, and Lin 2023) employs multi-level prediction alignment with multiple temperature coefficients to enhance knowledge transfer. WTTM (Zheng and Yang 2024) uses transformed teacher matching to emphasize temperature scaling in knowledge refinement.

2) Decoupling the logit knowledge: To explore more effective logit knowledge, recent studies have focused on decoupling the logit knowledge (Zhao et al. 2022; Yang et al. 2023): For instance, ATS (Li et al. 2022) decomposes the KL divergence into correction guidance, label smoothing, and class distinguishability to improve distillation. ReKD (Xu et al. 2024) divides the logit knowledge into head and tail categories for targeted transfer. SDD (Wei, Luo, and Luo 2024) breaks down global logit outputs into local outputs, creating distillation pipelines that enhance detailed knowledge acquisition. These methods can improve the student's performance by facilitating knowledge transfer.

However, these methods cannot effectively address the intra-batch noise. Although DIST (Huang et al. 2022), MLKD (Jin, Wang, and Lin 2023), and LCKA (Zhou et al. 2024) incorporate intra-batch distillation, their performance stability is affected by sample randomness and batch size. In this paper, we propose an Intra-class Knowledge Distillation (IKD) that enhances the student's generalization within the same class by emphasizing sample consistency. Our method is a plug-and-play strategy that can be easily integrated with logit-based KD methods and complements the shortcomings of traditional distillation methods.

Methodology

Notation

Given a mini-batch of training samples $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{b} \subseteq \mathcal{D}^{train}$, where $\mathcal{D}^{train}$ denotes the training sample set and b denotes the batch size. For a training sample x_i with label y_i, the feature extractor $\mathcal{F}(\cdot; \omega)$ and the classifier $\mathcal{G}(\cdot; \theta)$ are parameterized by ω and θ, respectively. When a training sample x_i is fed into the network, the feature extractor $\mathcal{F}$ produces the penultimate embedding features $f = \mathcal{F}(x_i; \omega)$. These features are then processed by the classifier $\mathcal{G}$ to yield the output $z_i = \mathcal{G}(f; \theta)$. In addition, we denote the probability distribution of the k-th class and the logit for the teacher as $p_i^t(k)$ and z_i^t, respectively, and for the student as $p_i^s(k)$ and z_i^s, respectively. The probability distributions for the teacher and the student corresponding to the k-th category of the i-th training sample are expressed as follows:

$$p_i^t(k) = \frac{exp(z_k^t/\tau)}{\sum_{j=1}^{C} exp(z_j^t/\tau)}, p_i^s(k) = \frac{exp(z_k^s/\tau)}{\sum_{j=1}^{C} exp(z_j^s/\tau)}, \quad (1)$$

where C denotes the number of categories, and τ denotes the temperature coefficient during distillation.

Conventional Knowledge Distillation

Kullback-Leibler (KL) divergence is typically used as the loss metric in KD. The training aims to make the probability distribution predicted by the student approach that of the teacher, i.e., minimizing the discrepancy between them. The KD loss is defined as follows:

$$\mathcal{L}_{KD} = \frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{C} p_j^t(k) \log(\frac{p_j^t(k)}{p_j^s(k)}), \quad (2)$$

where we omit the temperature coefficient τ for simplicity.

The cross-entropy loss for image classification in the student can be defined as follows:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{j=1}^{N} \sum_{k=1}^{C} y_j^s(k) \log p_i^s(k)), \quad (3)$$

where $y_j^s(k)$ denotes the one-hot label for the k-th category of the j-th training sample for the student.

The total loss $\mathcal{L}$ in knowledge distillation consists of both the KD loss $\mathcal{L}_{KD}$ and the cross-entropy loss $\mathcal{L}_{CE}$. The combined loss $\mathcal{L}$ can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{KD}, \quad (4)$$

where α is a hyperparameter for balancing the two losses.

Motivation

Logit-based KD methods boost student performance by minimizing inter-class variance (Safaryan, Peste, and Alistarh 2024) between the teacher and student. However, these methods may increase boundary uncertainty due to noise interference, such as occlusion and irrelevant categories. Furthermore, samples from the same category may exhibit significant differences in pose, lighting, angle, and scale. Failing to consider intra-class variations may lead to performance degradation adequately. To tackle these challenges, we propose an Intra-class Knowledge Distillation (IKD) method, which can effectively reduce noise introduction and mitigate performance degradation.

Intra-class Knowledge Distillation

Fig. 3 presents an overview of the comprehensive framework, and we propose an IKD that facilitates knowledge sharing within the same class to ensure consistent predictions. IKD preparation is detailed in three steps:

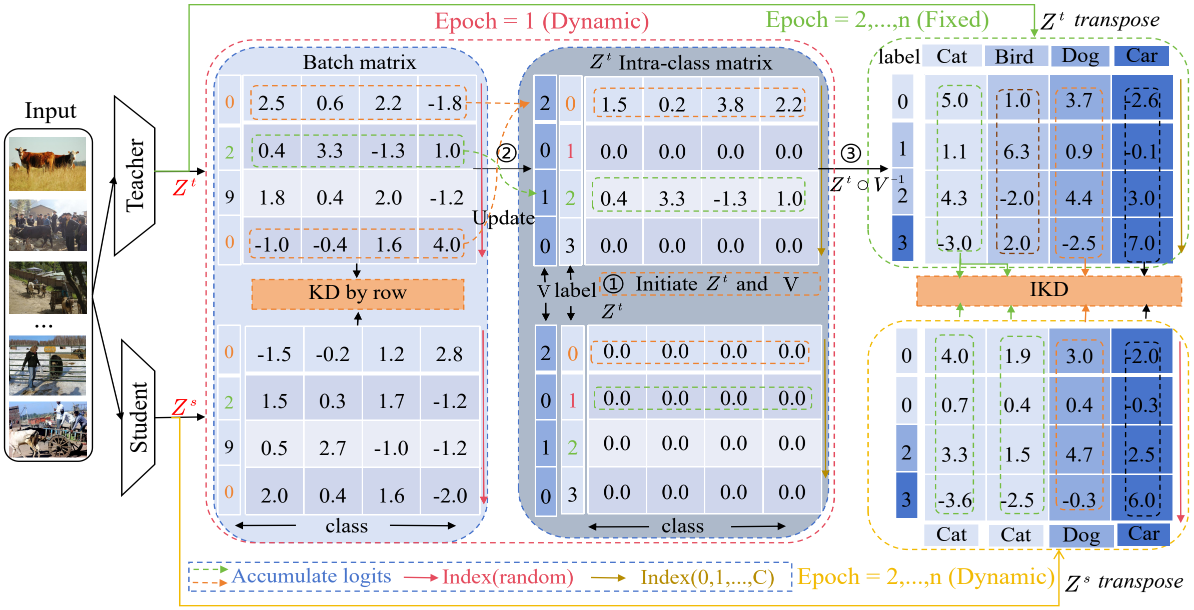


Figure 3: Overview of the proposed IKD framework, including its implementation steps and workflow.

Step 1: We initialize the matrix $\mathbf{Z}^t \in \mathbb{R}^{C\times C}$ and the vector $\mathbf{V} \in \mathbb{N}^C$ as a zero matrix and a zero vector, respectively. They are used to store the average logits of the teacher and count samples according to the label index. They are defined as $\mathbf{Z}^t = (z^t_{i,j})_{1\leq i,j\leq C} \in \mathbb{R}^{C\times C}$ and $\mathbf{V} = [v_1, \quad \cdots \quad, v_C] \in \mathbb{N}^{1\times C}$, respectively, where v_c denotes the count of the samples that belong to category c, with $1 \leq c \leq C$. $z^t_{i,j}$ denotes the average logit of samples from class i for class j.

We use $z^t_{c,:}$ to represent the c-th row of $\mathbf{Z}^t$ corresponding to the average logit of samples from class c.

Step 2: We calculate the cumulative sum of logits and the sample count for each class of the teacher in the first epoch. Since the teacher's parameters are fixed, this calculation is only carried out in the first epoch. Each sample's logits, generated by the teacher, are added to the matrix $\mathbf{Z}^t$ according to the label index. Meanwhile, the sample count is accumulated according to the label index in the vector $\mathbf{V}$. This process repeats until all samples have been iterated through in the first epoch. The update of the elements in $\mathbf{Z^t}$ and $\mathbf{V}$ can be defined as follows:

$$z^t_{c,:} \leftarrow z^t_{c,:} + z_c, \quad v_c \leftarrow v_c + 1, \tag{5}$$

where the term z_c represents the logit corresponding to label index c for a given sample, with $z_c \in \mathbb{R}^{1\times C}$.

Step 3: We update the matrix $\mathbf{Z}^t$ to calculate the average logits for each class as $\mathbf{Z}^t \leftarrow \mathbf{Z}^t \circ \mathbf{V}^{-1}$, where $\circ$ denotes Hadamard product. The term $\mathbf{V}^{-1}$ represents the element-wise inverse of the vector $\mathbf{V}$. To align dimensions, $\mathbf{V}^{-1}$ is broadcast to match the dimensions of each row in the matrix $\mathbf{Z}^t$.

Conducting Knowledge Distillation. During the first epoch, we focus on data statistics (i.e., **Step 1** through **Step 3**) and conduct normal distillation similar to KD (refer to Eq. (4) for details) to reduce knowledge omission.

From the second epoch onward, the IKD is conducted, with the teacher's parameters being fixed. This process utilizes the average logits $\mathbf{Z}^t$ obtained from **Setp 3** in the first epoch, and $\mathbf{Z}^t$ keeps unchanged from one epoch to the next, continuing until the final epoch. Meanwhile, for each batch in every epoch iteration, we compute the KL divergence between the student's output and $\mathbf{Z}^t$ according to the label index. The $\mathcal{L}_{IKD}$ for intra-class distillation is as follows:

$$\mathcal{L}_{IKD} = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{C} p^t(k)\log\left(\frac{p^t(k)}{p^s_j(k)}\right), \tag{6}$$

where the probability that the j-th sample belongs to category $k = c$ is denoted as $p^t(k) = softmax(z^t_{k,:}/T)$ and $p^s_j(k) = softmax(z^s_j/T)$. T denotes the intra-class temperature.

Algorithm 1 presents the pseudo-code of the IKD. Please refer to *Supplementary Materials* for a detailed algorithm description.

Comparison with Intra-batch Knowledge Distillation.

- 1) In intra-batch distillation, batch samples randomly selected may come from different classes, leading to a shift in the student's focus to inter-class similarities, and it is also influenced by the batch size.
- 2) In the IKD method, the teacher's output (i.e., Z^t) is set in the first epoch and fixed thereafter, while in batch distillation, the teacher's output (i.e., z^t_j) dynamically adjusts with each batch. The IKD provides a stable learning target, which can simplify the training process and reduce the risk of overfitting.
- 3) The IKD method distills knowledge within the same category, which can effectively capture subtle intra-class

differences and variations. IKD is ideal for tasks with significant intra-class diversity.

Optimization Objective

The proposed IKD can be integrated seamlessly with logit-based KD methods, such as KD, DKD, NKD, and WTTM, and it has the effect of optimizing student performance and enhancing prediction consistency. Specifically, taking KD as an example, the overall loss expression is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{KD} + \lambda_{avg}\mathcal{L}_{IKD}, \tag{7}$$

where λ_{avg} is the weight coefficient for balancing $\mathcal{L}_{total}$.

Theoretical Analysis

Theoretical Analysis of Distillation Bias. We present the bias-variance decomposition of $\mathcal{L}_{\mathrm{KD}}$ based on the definitions and notations from (Tom 1998; Helong et al. 2021).

Proposition 1. *Based on the theory from (Tom 1998), the bias term can be decomposed and expressed as follows:*

$$\begin{aligned} bias = D_{\mathrm{KL}}\left(y, \overline{y}_{\mathrm{ce}}\right) + \sum_{c=1}^{C} \mathbb{E}_{x|x_c}\left[y \log\left(\frac{\overline{y}_{\mathrm{c,ce}}}{\overline{y}^{t}_{\mathrm{c,kd}}}\right)\right] \\ + \sum_{c=1}^{C} \mathbb{E}_{x|x_c}\left[y \log\left(\frac{\overline{y}^{t}_{\mathrm{c,kd}}}{\overline{y}_{\mathrm{c,kd}}}\right)\right]. \end{aligned} \tag{8}$$

Proof. The detailed proof is provided in *Supplementary Materials*.

The decoupling of the bias term in Proposition 1 shows that considering the third term related to the IKD on the right side of Eq. (8) can increase the consistency within the same class. Notably, $\overline{y}^{t}_{\mathrm{c,kd}}$, representing the teacher's predictions, approximates the one-hot distribution of category c more closely. The IKD method brings $\overline{y}_{\mathrm{c,kd}}$ from the student closer to the $\overline{y}^{t}_{\mathrm{c,kd}}$, reducing the term $\sum_{c=1}^{C} \mathbb{E}_{x|x_c}\left[y \log\left(\frac{\overline{y}^{t}_{\mathrm{c,kd}}}{\overline{y}_{\mathrm{c,kd}}}\right)\right]$, decreasing the bias term. Thus, the total distillation error is reduced.

Theoretical Analysis of Model Complexity. The IKD aims to reduce the student's output variance for each class, thereby minimizing prediction inconsistency. This reduction in variance correlates with lower model complexity, enhancing generalization performance. The details are as follows:

Proposition 2. *Suppose the variance for class c is reduced by a factor of k_c, and we can obtain $h'\left(x_i\right) = \frac{1}{\sqrt{k_c}} h\left(x_i\right)$.*

Proof. The detailed proof is provided in *Supplementary Materials*.

As a result of the reduced variance, we establish a relationship (Sain 1996; Bartlett, Bousquet, and Mendelson 2005) between the output $h'\left(x_i\right)$ in Proposition 2 and the Rademacher complexity $R_n\left(\mathcal{H}\right)$ (Bartlett and Mendelson 2002).

Proposition 3. *When the variance for class c is reduced by a factor of k_c, Rademacher complexity decreases by ΔR_n, yielding $\Delta R_n = (1 - \frac{1}{\sqrt{k_c}})R_n(\mathcal{H})$.*

Proof. Details on ΔR_n are in the *Supplementary Materials*.

From the theoretical derivation above, it is evident that by reducing variance, the IKD method decreases Rademacher complexity, thereby enhancing generalization performance.

Experiments

In this section, we evaluate our proposed Intra-class Knowledge Distillation (IKD) method and show its effectiveness across various classification datasets. In addition, we conduct ablation studies, and visualizations to further validate our method. All experiments are conducted repeatedly three times, and we report the averaged results.

Experimental Setup

In our experiments, we evaluate our method using three classic datasets and provide the implementation details.

CIFAR-100 (Krizhevsky, Hinton et al. 2009) comprises 100 classes, with each image having a resolution of 32×32 pixels. The CIFAR-100 dataset contains 50k training images and 10k validation images.

ImageNet-1K (ILSVRC2012) (Deng et al. 2009) is a large-scale dataset comprising 1k classes. The dataset comprises 1.2 million training images and 50k validation images.

Tiny-ImageNet (Le and Yang 2015) is a subset of the ImageNet-1K dataset, consisting of 200 classes, and the image is 64×64 pixels. The training set contains 100k images, and the validation set contains 10k images.

Implementations details are provided in *Supplementary Materials* due to page constraints.

Main Results and Analysis

Image Classification Results on CIFAR-100. We have conducted extensive experiments across two different architectures to evaluate our method. Specifically, as illustrated in Table 2, we employ an experimental setup in which the teacher and student either share the same architecture or utilize different architectures for a fair comparison.

Our primary focus is on logit-based knowledge distillation as a plug-and-play strategy. In both the same and different architecture setups, we integrate IKD with four baseline methods (KD, DKD, NKD, WTTM) for comparison and compare it with the state-of-the-art (SOTA) plug-and-play methods, including DOT, CTKD, SDD, and LSKD. The experimental results, as presented in Table 2, demonstrate that IKD consistently outperforms these plug-and-play methods, achieving an average performance improvement ranging from 0.63% to 1.44%, regardless of whether the architectures are the same or different. These results highlight the effectiveness of IKD on the CIFAR-100 dataset.

Image Classification Results on Imagenet-1K. We evaluate our method on the ImageNet-1K dataset using Top-1 and Top-5 accuracy as the primary metrics. As shown in Table 3 and Table 4, our method significantly outperforms most SOTA plug-and-play methods across different baselines, strongly supporting its efficacy and potential. However, it is noteworthy that NKD+IKD may result in loss of critical information or inappropriate knowledge transfer, as evidenced by the Top-5 performance in Table 3. We think that NKD emphasizes non-target classes, which may divert attention from target classes. At the same time, IKD focuses on enhancing the performance of target classes, potentially diminishing the learning of non-target classes.

	Method	Homogeneous Architecture							Heterogeneous Architecture		
Distillation Manner	Teacher	ResNet56	ResNet110	ResNet110	ResNet32×4	WRN-40-2	WRN-40-2	VGG13	VGG13	ResNet50	ResNet50
	Accuracy	72.34	74.31	74.31	79.42	75.61	75.61	74.64	74.64	79.34	79.34
	Student	ResNet20	ResNet20	ResNet32	ResNet8×4	WRN-16-2	WRN-40-1	VGG8	MobileNetV2	MobileNetV2	VGG8
	Accuracy	69.06	69.06	71.14	72.50	73.26	71.98	70.36	64.60	64.60	70.36
Features	AT	70.55	70.22	72.31	73.44	74.08	72.77	71.43	59.40	58.58	71.84
	PKD	70.34	70.25	72.61	73.64	74.54	73.45	72.88	67.13	66.52	73.01
	VID	70.38	70.16	72.61	73.09	74.11	73.30	71.23	65.56	67.57	70.30
	FitNet	69.21	68.99	71.06	73.50	73.58	72.24	71.02	64.14	63.16	70.69
	RKD	69.61	69.25	71.82	71.90	73.35	72.22	71.48	64.52	64.43	71.50
	CRD	71.16	71.46	73.48	75.51	75.48	74.14	73.94	69.73	69.11	74.30
	OFD	70.98	70.61	73.23	74.95	75.24	74.33	73.95	69.48	69.04	-
	CC	69.63	69.48	71.48	72.97	73.56	72.21	70.71	64.86	65.43	70.25
	ReviewKD	**71.89**	**71.85**	**73.89**	75.63	76.12	75.09	74.84	70.37	69.89	72.60
	FT	69.84	70.22	72.37	72.86	73.25	71.59	70.58	61.78	60.99	70.29
	CAT-KD	71.62	71.14	73.62	**76.91**	75.60	74.82	74.65	69.13	**71.36**	**75.39**
	FCFD	71.68	-	-	76.80	**76.34**	**75.43**	**74.86**	**70.67**	71.07	-
Logits	KD	70.66	70.67	73.08	73.33	74.92	73.54	72.98	67.37	67.35	73.81
	KD+CTKD	71.08	71.01	73.41	73.71	75.46	74.02	73.15	68.84	68.24	73.63
	KD+DOT	71.11	70.97	73.37	74.98	75.43	73.87	73.88	63.00	64.18	73.95
	KD+LSKD	71.24	71.61	73.76	76.16	**76.22**	74.43	74.23	69.43	69.50	74.42
	KD+SDD	71.52	71.58	73.97	75.09	75.86	74.53	73.91	68.43	69.76	74.69
	KD+IKD	**72.00**	**72.18**	**74.44**	**76.56**	76.01	**74.84**	**74.80**	**69.87**	**70.49**	**75.12**
	Δ	+1.34	+1.51	+1.36	+3.23	+1.09	+1.30	+1.82	+2.50	+3.14	+1.31
	DKD	71.77	70.91	73.93	76.08	75.64	74.87	74.44	69.71	70.35	75.34
	DKD+CTKD	71.81	71.37	73.74	76.32	75.60	74.53	74.61	69.84	70.32	75.29
	DKD+DOT	71.12	71.58	73.57	76.03	75.42	74.49	74.59	62.48	57.89	74.72
	DKD+LSKD	71.63	71.71	73.72	76.56	76.02	74.79	74.68	**70.23**	70.40	75.41
	DKD+SDD	71.49	71.41	74.03	76.41	75.36	74.52	74.27	69.87	**71.46**	75.55
	DKD+IKD	**72.37**	**72.18**	**74.29**	**77.03**	**76.16**	**75.06**	**75.12**	69.87	70.82	**75.66**
	Δ	+0.60	+1.27	+0.36	+0.95	+0.52	+0.19	+0.68	+0.16	+0.47	+0.32
	NKD	71.47	71.23	73.21	76.39	75.07	74.20	74.61	69.78	69.39	74.01
	NKD+CTKD	71.63	71.37	73.76	75.65	75.82	74.38	73.42	69.19	69.67	74.27
	NKD+DOT	71.49	71.36	71.29	70.65	75.22	74.02	70.96	N/A	N/A	67.31
	NKD+LSKD	70.85	70.96	73.09	75.86	75.90	73.97	73.55	68.85	69.27	74.33
	NKD+SDD	**72.05**	72.18	74.04	76.55	75.77	74.03	73.86	69.25	70.23	73.27
	NKD+IKD	72.01	**72.18**	**74.16**	**77.10**	**76.03**	**75.01**	**75.05**	**69.90**	**71.03**	**75.71**
	Δ	+0.54	+0.95	+0.95	+0.71	+0.96	+0.81	+0.44	+0.12	+1.64	+1.70
	WTTM	71.77	71.53	73.82	76.12	76.29	73.99	74.42	68.45	69.18	74.82
	WTTM+CTKD	-	-	-	-	-	-	-	-	-	-
	WTTM+DOT	71.33	71.01	68.34	72.69	72.02	75.22	15.50	N/A	N/A	N/A
	WTTM+LSKD	69.29	69.91	**74.19**	**76.75**	75.80	73.94	73.70	67.01	66.80	74.28
	WTTM+SDD	71.57	71.42	73.78	76.40	75.72	**74.66**	74.38	**69.36**	69.63	**75.09**
	WTTM+IKD	**72.02**	**71.87**	73.95	76.16	**76.44**	74.24	**74.48**	68.52	**69.69**	75.01
	Δ	+0.25	+0.34	+0.13	+0.04	+0.15	+0.21	+0.06	+0.07	+0.51	+0.19
	Ψ	+0.68	+1.02	+0.70	+1.23	+0.68	+0.63	+0.75	+0.73	+1.44	+0.88

Table 2: Test accuracy (%) of students on CIFAR-100 validation set. Δ and Ψ indicate performance and average performance improvement of the IKD, respectively. Here, Ψ = (Δ(KD+IKD) + Δ(DKD+IKD) + Δ(NKD+IKD) + Δ(WTTM+IKD)) / 4.0.

ResNet34 (teacher): 73.31% Top-1, 91.42% Top-5 accuracy. ResNet18 (student): 69.75% Top-1, 89.07% Top-5 accuracy.										
Features	AT	OFD	CRD	ReviewKD	SP	CC	MGD*	RKD	SRRL	CAT-KD
Top-1	70.69	70.81	71.17	71.61	70.62	69.96	71.69	70.40	**71.73**	71.26
Top-5	90.01	89.98	90.13	90.51	89.80	89.17	90.42	89.78	**90.60**	90.45

Logits	KD							DKD							NKD							
	Vanilla	+CTKD	+DOT	+LSKD	+SDD	+IKD	Δ	Vanilla	+CTKD	+DOT	+LSKD	+SDD	+IKD	Δ	Vanilla	+CTKD	+DOT	+LSKD	+SDD	+IKD	Δ	Ψ
Top-1	70.66	71.32	71.72	71.42	71.44	**71.82**	+1.16	71.72	71.51	72.03	71.88	72.02	**72.04**	+0.32	71.97	-	71.47	71.57	**72.33**	72.20	+0.23	+0.57
Top-5	89.88	90.27	90.30	90.29	90.05	**90.57**	+0.69	90.41	90.47	90.50	90.58	**91.21**	90.59	+0.18	91.10	-	90.00	90.11	**91.31**	90.66	-0.44	+0.14

Table 3: Top-1 and Top-5 accuracy (%) results on the ImageNet-1K validation set using homogeneous architecture.

Fine-grained Image Classification Results on Tiny-ImageNet. To further validate the effectiveness of our method, we conduct tests on the more fine-grained Tiny-ImageNet dataset, characterized by high intra-class similarity and compact features. As shown in Table 5 (see *Supplementary Materials*), our method has been proven to be effective on this dataset. The effectiveness of IKD stems from its enhancement of the student's ability to capture and discriminate subtle distinctions within classes. This approach improves the student's detail sensitivity, significantly boost-

ResNet50 (teacher): 76.16% Top-1, 92.87% Top-5 accuracy. MobileNetV1 (student): 68.87% Top-1, 88.76% Top-5 accuracy.										
Features	AT	OFD	CRD	ReviewKD	RKD	AB	MGD*	MGD	SRRL	CAT-KD
Top-1	70.18	71.25	71.32	72.56	68.50	68.89	72.49	71.47	**72.49**	72.24
Top-5	89.68	90.34	90.41	91.00	88.32	88.71	90.94	90.35	90.92	**91.13**

Logits	KD							DKD							NKD							
	Vanilla	+CTKD	+DOT	+LSKD	+SDD	+IKD	Δ	Vanilla	+CTKD	+DOT	+LSKD	+SDD	+IKD	Δ	Vanilla	+CTKD	+DOT	+LSKD	+SDD	+IKD	Δ	Ψ
Top-1	70.49	-	73.09	72.18	72.24	**73.27**	+2.78	72.05	-	**73.33**	72.85	73.08	73.02	+0.97	72.58	-	72.35	72.31	73.12	**73.32**	+0.74	+1.50
Top-5	89.92	-	91.11	90.80	90.71	**91.34**	+1.42	91.05	-	91.22	91.23	91.09	**91.46**	+0.41	90.80	-	90.48	90.46	91.11	**91.41**	+0.61	+0.81

Table 4: Top-1 and Top-5 accuracy (%) results on the ImageNet-1K validation set using heterogeneous architecture.

ing recognition performance.

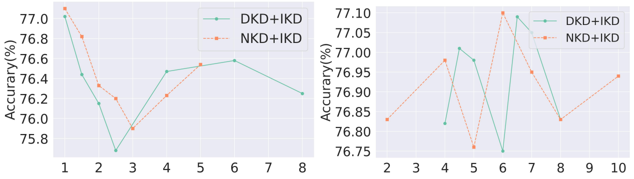
Ablation Studies

In the following experiments, we evaluate the effectiveness of IKD through ablation experiments focusing on intra-class temperature, and weight coefficient.

Hyperparameters Sensitivity Analysis. We conduct extensive ablation studies on the intra-class temperature T and the weight of IKD λ_{avg}.

1) Effect of T: We conduct ablation experiments on the intra-class temperature T. It is observed that lower temperature results in higher classification accuracy. This is probably because if the temperature is set high, noise may be introduced, and the student's generalization performance could also be negatively impacted. The experimental results indicate that appropriately lowering the temperature can reduce noise interference and improve classification accuracy. Therefore, as shown in Fig. 4 (a), choosing a temperature parameter of 1.0 is appropriate for suppressing noise.

2) Effect of λ_{avg}: To optimize the balance of loss weights, Fig. 4 (b) demonstrates that for DKD+IKD, the optimal performance is achieved with a λ_{avg} of 6.5, whereas for NKD+IKD, a λ_{avg} of 6.0 yields the best results. As shown in Fig. 4 (b), λ_{avg} significantly impacts model performance. High settings of λ_{avg} overemphasize intra-class details, while low settings underutilize potential intra-class information, both leading to notable performance fluctuations. For other parameter details, refer to *Supplementary Materials*.



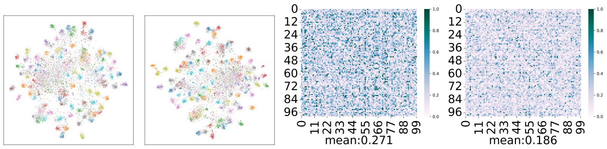
(a) Effect of T

(b) Effect of λ_{avg}

Figure 4: Ablation study evaluates the impact of varying the parameters T and λ_{avg} on the performance of DKD+IKD and NKD+IKD. We employ ResNet32×4 as teacher and ResNet8×4 as student on the CIFAR-100 dataset.

Visualizations

We visualize our experimental results using t-SNE (Van der Maaten and Hinton 2008) and correlation matrices. As shown in Figs. 5(a)-(b), the t-SNE results show that by incorporating KD into the proposed IKD, the extracted features become more separable, demonstrating the effectiveness of our strategy. Furthermore, as illustrated in Figs. 5(c)-(d), integrating IKD with KD yields lower correlation matrix means than KD alone. These results suggest that the student has assimilated more detailed information, and hence its performance is improved.



(a) KD

(b) KD+IKD

(c) KD

(d) KD+IKD

Figure 5: Visualization of the penultimate-layer features trained using KD and KD+IKD, respectively, on the CIFAR-100 dataset using t-SNE and correlation matrices, with ResNet32×4 as the teacher and ResNet8×4 as the student.

Limitation and Discussion

Our research only considers logit layer distillation, which may not capture the full potential of knowledge. Our training has focused only on the classification task, without validating the generalization ability of IKD on other tasks. In future work, We plan to thoroughly test our method across various tasks (e.g., Person Re-ID) to assess its effectiveness in the future. These efforts will deepen the understanding of IKD techniques' versatility.

Conclusion

In this paper, we introduce Intra-class Knowledge Distillation (IKD), a regularization method designed to address the shortcomings of existing distillation techniques, particularly in handling prediction bias. The proposed method not only reduces discrepancies in predictions but also increases the robustness of students. Moreover, IKD theoretically enhances intra-class consistency, which in turn minimizes biases in predictive performance. Our experiments on CIFAR-100, ImageNet-1K, and Tiny-ImageNet demonstrate the effectiveness of the IKD in improving generalization and student performance.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. U21A20487), in part by Guangdong Major Project of Basic and Applied Basic Research (No. 2023B0303000016), in part by Shenzhen Technology Project (No. JCYJ20220818101206014, No. JCYJ20220818101211025), in part by Guangdong Technology Project (No. 2022B1515120067), and in part by Yunnan Science & Technology Project (No. 202305AF150152).

References

Ahn, S.; Hu, S. X.; Damianou, A.; Lawrence, N. D.; and Dai, Z. 2019. Variational information distillation for knowledge transfer. In *CVPR*, pages 9163–9171.

Bartlett, P. L.; Bousquet, O.; and Mendelson, S. 2005. Local rademacher complexities.

Bartlett, P. L.; and Mendelson, S. 2002. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 3(11): pages 463–482.

Chen, P.; Liu, S.; Zhao, H.; and Jia, J. 2021. Distilling knowledge via knowledge review. In *CVPR*, pages 5008–5017.

Cho, J. H.; and Hariharan, B. 2019. On the efficacy of knowledge distillation. In *ICCV*, pages 4794–4802.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255.

Duan, Z.; Lu, M.; Ma, J.; Huang, Y.; Ma, Z.; and Zhu, F. 2023. Qarv: Quantization-aware resnet vae for lossy image compression. *TPAMI*, 46(1): pages 436–450.

Furlanello, T.; Lipton, Z.; Tschannen, M.; Itti, L.; and Anandkumar, A. 2018. Born again neural networks. In *ICML*, pages 1607–1616.

Guo, Z.; Yan, H.; Li, H.; and Lin, X. 2023. Class Attention Transfer Based Knowledge Distillation. In *CVPR*, pages 11868–11877.

Haeffele, B. D.; and Vidal, R. 2019. Structured low-rank matrix factorization: Global optimality, algorithms, and applications. *TPAMI*, 42(6): pages 1468–1482.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

Helong, Z.; Liangchen, S.; Jiajie, C.; Ye, Z.; Guoli, W.; Junsong, Y.; and Zhang, Q. 2021. Rethinking soft labels for knowledge distillation: a bias-variance tradeoff perspective. In *ICLR*.

Heo, B.; Kim, J.; Yun, S.; Park, H.; Kwak, N.; and Choi, J. Y. 2019a. A comprehensive overhaul of feature distillation. In *ICCV*, pages 1921–1930.

Heo, B.; Lee, M.; Yun, S.; and Choi, J. Y. 2019b. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. *AAAI*, 33(01): pages 3779–3787.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2014. Distilling the knowledge in a neural network. In *NIPS Workshop*.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In *CVPR*, pages 484–492.

Huang, T.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2022. Knowledge distillation from a stronger teacher. *NIPS*, 35: pages 33716–33727.

Huang, Z.; and Wang, N. 2017. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*.

Jin, Y.; Wang, J.; and Lin, D. 2023. Multi-level logit distillation. In *CVPR*, pages 24276–24285.

Kim, J.; Park, S.; and Kwak, N. 2018. Paraphrasing complex network: Network compression via factor transfer. *NIPS*, 31.

Komodakis, N.; and Zagoruyko, S. 2017. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Toronto, ON, Canada*.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *AMS*, 22(1): pages 79–86.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): pages 3.

Li, X.-C.; Fan, W.-S.; Song, S.; Li, Y.; Yunfeng, S.; Zhan, D.-C.; et al. 2022. Asymmetric temperature scaling makes larger networks teach well again. *NIPS*, 35: pages 3830–3842.

Li, Z.; Li, X.; Yang, L.; Zhao, B.; Song, R.; Luo, L.; Li, J.; and Yang, J. 2023. Curriculum temperature for knowledge distillation. In *AAAI*, pages 1504–1512.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, pages 7061–7070.

Liu, D.; Kan, M.; Shan, S.; and CHEN, X. 2023. Function-Consistent Feature Distillation. In *ICLR*.

Miles, R.; and Mikolajczyk, K. 2024. Understanding the role of the projector in knowledge distillation. In *AAAI*, pages 4233–4241.

Mirzadeh, S. I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; and Ghasemzadeh, H. 2020. Improved knowledge distillation via teacher assistant. In *AAAI*, pages 5191–5198.

Niu, Y.; Chen, L.; Zhou, C.; and Zhang, H. 2022. Respecting transfer gap in knowledge distillation. *NIPS*, 35: pages 21933–21947.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *CVPR*, pages 3967–3976.

Passalis, N.; and Tefas, A. 2018. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, pages 268–284.

Peng, B.; Jin, X.; Liu, J.; Li, D.; Wu, Y.; Liu, Y.; Zhou, S.; and Zhang, Z. 2019. Correlation congruence for knowledge distillation. In *ICCV*, pages 5007–5016.

Qin, Z.; Wang, K.; Zheng, Z.; Gu, J.; Peng, X.; xu Zhao Pan; Zhou, D.; Shang, L.; Sun, B.; Xie, X.; and You, Y. 2024. InfoBatch: Lossless Training Speed Up by Unbiased Dynamic Data Pruning. In *ICLR*.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. Fitnets: Hints for thin deep nets. In *ICLR*.

Safaryan, M.; Peste, A.; and Alistarh, D. 2024. Knowledge distillation performs partial variance reduction. In *NIPS*, volume 36.

Sain, S. R. 1996. The nature of statistical learning theory.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520.

Shen, C.; Huang, Y.; Zhu, H.; Fan, J.; and Zhang, G. 2024. Student-Oriented Teacher Knowledge Refinement for Knowledge Distillation. In *ACM MM*.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Son, W.; Na, J.; Choi, J.; and Hwang, W. 2021. Densely guided knowledge distillation using multiple teacher assistants. In *ICCV*, pages 9395–9404.

Song, J.; Chen, Y.; Ye, J.; and Song, M. 2022. Spot-adaptive knowledge distillation. *TIP*, 31: pages 3359–3370.

Sun, S.; Ren, W.; Li, J.; Wang, R.; and Cao, X. 2024. Logit Standardization in Knowledge Distillation. In *CVPR*, pages 15731–15740.

Tang, J.; Shivanna, R.; Zhao, Z.; Lin, D.; Singh, A.; Chi, E. H.; and Jain, S. 2020. Understanding and improving knowledge distillation. *arXiv preprint arXiv:2002.03532*.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive representation distillation. In *ICLR*.

Tom, H. 1998. Bias/variance decompositions for likelihood-based estimators. *NC*, 10(6): pages 1425–1433.

Tung, F.; and Mori, G. 2019. Similarity-preserving knowledge distillation. In *ICCV*, pages 1365–1374.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR*, 9(11).

Wei, S.; Luo, C.; and Luo, Y. 2024. Scaled Decoupled Distillation. In *CVPR*, pages 15975–15983.

Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I. S.; and Xie, S. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *CVPR*, pages 16133–16142.

Xiaolong, L.; Lujun, L.; Chao, L.; and Yao, A. 2023. Norm: Knowledge distillation via n-to-one representation matching. In *ICLR*.

Xu, L.; Ren, J.; Huang, Z.; Zheng, W.; and Chen, Y. 2024. Improving Knowledge Distillation via Head and Tail Categories. *TCSVT*, 34(5): pages 3465–3480.

Yang, C.; Zhou, H.; An, Z.; Jiang, X.; Xu, Y.; and Zhang, Q. 2022a. Cross-image relational knowledge distillation for semantic segmentation. In *CVPR*, pages 12319–12328.

Yang, J.; Martinez, B.; Bulat, A.; Tzimiropoulos, G.; et al. 2021. Knowledge distillation via softmax regression representation learning. In *ICLR*.

Yang, S.; Yang, J.; Zhou, M.; Huang, Z.; Zheng, W.-S.; Yang, X.; and Ren, J. 2024. Learning From Human Educational Wisdom: A Student-Centered Knowledge Distillation Method. *TPAMI*, 46(6): pages 4188–4205.

Yang, Z.; Li, Z.; Jiang, X.; Gong, Y.; Yuan, Z.; Zhao, D.; and Yuan, C. 2022b. Focal and global knowledge distillation for detectors. In *CVPR*, pages 4643–4652.

Yang, Z.; Li, Z.; Shao, M.; Shi, D.; Yuan, Z.; and Yuan, C. 2022c. Masked generative distillation. In *ECCV*, pages 53–69.

Yang, Z.; Zeng, A.; Li, Z.; Zhang, T.; Yuan, C.; and Li, Y. 2023. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *ICCV*, pages 17185–17194.

Yue, K.; Deng, J.; and Zhou, F. 2020. Matching guided distillation. In *ECCV*, pages 312–328.

Yun, S.; Park, J.; Lee, K.; and Shin, J. 2020. Regularizing class-wise predictions via self-knowledge distillation. In *CVPR*, pages 13876–13885.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. In *BMVC*, pages 87.1–87.12.

Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; and Ma, K. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *ICCV*, pages 3713–3722.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018a. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, pages 6848–6856.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018b. Deep mutual learning. In *CVPR*, pages 4320–4328.

Zhao, B.; Cui, Q.; Song, R.; and Liang, J. 2023. DOT: A Distillation-Oriented Trainer. In *ICCV*, pages 6189–6198.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled Knowledge Distillation. In *CVPR*, pages 11953–11962.

Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; and Chen, J. 2024. Detrs beat yolos on real-time object detection. In *CVPR*, pages 16965–16974.

Zheng, K.; and Yang, E.-H. 2024. Knowledge Distillation Based on Transformed Teacher Matching. In *ICLR*.

Zheng, Z.; Ye, R.; Hou, Q.; Ren, D.; Wang, P.; Zuo, W.; and Cheng, M.-M. 2023. Localization distillation for object detection. *TPAMI*, 45(8): pages 10070–10083.

Zhou, Z.; Shen, Y.; Shao, S.; Chen, H.; Gong, L.; and Lin, S. 2024. Rethinking Centered Kernel Alignment in Knowledge Distillation. In *IJCAI*, 5680–5688.

Zong, M.; Qiu, Z.; Ma, X.; Yang, K.; Liu, C.; Hou, J.; Yi, S.; and Ouyang, W. 2023. Better Teacher Better Student: Dynamic Prior Knowledge for Knowledge Distillation. In *ICLR*.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2024. Segment everything everywhere all at once. *NIPS*, 36: 19769–19782.