DIST+: Knowledge Distillation from A Stronger Teacher

Tao Huang, Shan You, *Member, IEEE,* Fei Wang, Chen Qian, Chang Xu, *Senior Member, IEEE,*

Abstract—The paper introduces DIST, an innovative knowledge distillation method that excels in learning from a superior teacher model. DIST differentiates itself from conventional techniques by adeptly handling the often significant prediction discrepancies between the student and teacher models. It achieves this by focusing on maintaining the relationships between their predictions, implementing a correlation-based loss to explicitly capture the teacher's intrinsic inter-class relations. Moreover, DIST uniquely considers the semantic similarities between different instances and each class at the intra-class level. The method is further enhanced by two significant improvements: (1) A teacher acclimation strategy, which effectively reduces the discrepancy between teacher and student, thereby optimizing the distillation process. (2) An extension of the DIST loss from the logit level to the feature level, a modification that proves especially beneficial for dense prediction tasks. DIST stands out for its simplicity, practicality, and adaptability to various architectures, model sizes, and training strategies. It consistently delivers state-of-the-art results across a range of applications, including image classification, object detection, and semantic segmentation. The methodology and results are detailed in the paper, and the implementation code is available at https://github.com/hunto/DIST_KD.

Index Terms—Knowledge distillation, Pearson correlation, image classification, object detection, semantic segmentation

1 Introduction

The advent of automatic feature engineering fuels deep neural networks to achieve remarkable success in a plethora of computer vision tasks, such as image classification [1]–[5], object detection [6], [7], and semantic segmentation [8], [9]. In the path of pursuing better performance, current deep learning models generally grow deeper and wider [10], [11]. However, such heavy models are clumsy to deploy in practice due to the limitations of computational and memory resources. For an efficient model with competitive performance to those larger models, knowledge distillation (KD) [12] has been proposed to boost the performance of the efficient model (student) by distilling the knowledge of a larger model (teacher) during training.

The essence of knowledge distillation relies on how to formulate and transfer the knowledge from teacher to student. The most intuitive yet effective approach is to match the probabilistic prediction (response) scores between the teacher and student via Kullback–Leibler (KL) divergence [12]. In this way, the student can be guided with more informative signals during training, and is thus expected to have more promising performance than that being trained stand-alone. Besides this vanilla prediction match, other works [13]–[16] also investigate the knowledge within intermediate representations to further boost the distillation performance, but this usually induces additional training cost as a consequence. For example, OFD [13] proposes to distill the information via multiple intermediate layers, but requires additional convolutions for feature alignments; CRD [15] introduces a contrastive loss to transfer pair-wise relationships, but it needs to hold a memory bank for all 128-d features of ImageNet images, and produces additional 260M FLOPs of computation cost.

Recently, a few studies [17]–[19] have been performed to address the poor learning issue of the student network when the student and teacher model sizes significantly differ. For example, TAKD [18] proposes to reduce the discrepancy of teacher and student by resorting to an additional teaching assistant of moderate model size; DGKD [19] further improves TAKD by densely gathering all the assistant models to guide the student. However, increasing the model size is only one of the popular approaches to have a stronger teacher. There lacks a thorough analysis on the training strategies to derive a stronger teacher and their effect on KD. Most importantly, a generic enough solution is preferred to address the difficulty of KD brought by stronger teachers, rather than struggling to deal with different types of stronger teachers (with larger model size or stronger training strategy) individually.

To understand what makes a stronger teacher and their effect on KD, we systematically study the prevalent strategies for designing and training deep neural networks, and show that:

- Beyond scaling up the model size, a stronger teacher can also be derived through advanced training strategies, *e.g.*, label smoothing and data augmentation [20]. However, given a stronger teacher, the student's performance on the vanilla KD could be dropped, even worse than training from scratch without KD, as shown in Figure 1.

- *Tao Huang and Chang Xu are with School of Computer Science, Faculty of Engineering, The University of Sydney, Australia. E-mail: thua7590@uni.sydney.edu.au, c.xu@sydney.edu.au*
- *Shan You and Chen Qian are with SenseTime Research. E-mail: youshan@senseauto.com, qianchen@sensetime.com*
- *Fei Wang is with University of Science and Technology of China. E-mail: wangfei91@mail.ustc.edu.cn*

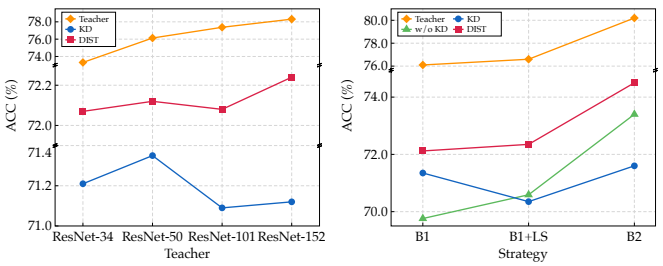

Figure 1. Comparisons of KD and our proposed DIST on ImageNet with different teachers. (a) The ResNet-18 students are trained using baseline strategy with different model sizes of the teacher. (b) The ResNet-18 students are trained using different strategies with ResNet-50 teachers.

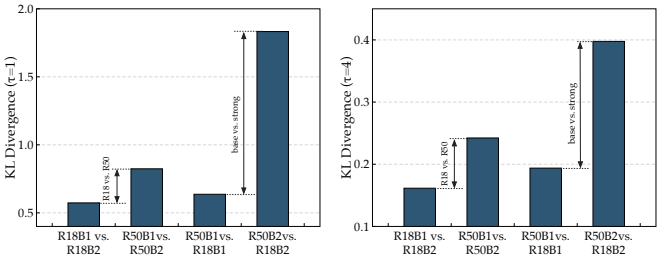

Figure 2. Discrepancy between the predictions of models trained standalone with different strategies on ImageNet validation set. (a) KL divergence with temperature $\tau = 1$. (b) KL divergence ($\tau = 4$). *R18B1* represents ResNet-18 trained with strategy B1. Details of training strategies refer to Table 2.

- The discrepancy between teacher and student tends to get fairly larger when we switch their training strategy to a stronger one (see Figure 2). In this case, an exact recovery of predictions via KL divergence could be challenging and lead to the failure of vanilla KD.
- Preserving the *relation of predictions* between teacher and student is sufficient and effective. When transferring the knowledge from teacher to student, what we really care about is preserving the preference (relative ranks of predictions) by the teacher, instead of recovering the absolute values accurately. Correlation between teacher and student predictions could be favored to relax the exact match of KL divergence and distill the intrinsic relations.

In this paper, we thus leverage the Pearson correlation coefficient [21] as a new match manner to replace the KL divergence. In addition, besides the *inter-class relations* in prediction vector (see Figure 3), with the intuition that different instances have different spectrum of similarities with respect to each class, we also propose to distill the *intra-class relations* for further boosting the performance as Figure 3. Concretely, for each class, we gather its corresponding predicted probabilities of all instances in a batch, then transfer this relation from teacher to student. Our proposed method (dubbed DIST) is super simple, efficient, and practical, which can be implemented with only several lines of code (see Appendix **??**) and has almost the same training cost as the vanilla KD. As a result, the student can be liberated from the burden of matching the exact output of a strong teacher, but only be guided appropriately to distill those truly informative relations.

A preliminary version of this work was presented earlier [22], namely DIST. This journal version extends the initial conference paper in multiple ways. First, we provide a comprehensive analysis of the discrepancy between student and stronger teacher, and find that the devil in KD from a stronger teacher lies in the inconsistency of non-target classes between teacher and student. We then design an efficient way to alleviate this inconsistency by acclimating the teacher's predictions to align with the student's, while maintaining the original accuracy of the teacher for effective distillation. Additionally, besides the logit-level distillation loss in DIST, we extend the loss to the feature level, which provides both pixel-level relational distillation and channel-level relational distillation. Combining the logit-level and feature-level distillation, our final method, dubbed DIST+, achieves further improvement on various tasks, especially on dense prediction tasks such as object detection.

Extensive experiments are conducted on benchmark datasets to verify our effectiveness on various tasks, including image classification, object detection, and semantic segmentation. Experimental results show that our DIST significantly outperforms vanilla KD and those sophisticatedly-designed state-of-the-art KD methods. For example, with the same baseline settings on ImageNet, our DIST achieves the highest 72.07% accuracy on ResNet-18. With the stronger strategy, our method obtains 82.3% accuracy on the recent transformer Swin-T [23], improving KD by 1%. As for DIST+, we obtain further improvement of 0.32% on ResNet-18 and 0.3% on Swin-T.

2 Related Work

2.1 Bridging the Representation Gap in Knowledge Distillation

Knowledge Distillation (KD), a method of transferring knowledge from a larger, complex teacher model to a smaller, more efficient student model, has seen significant advancements recently. One emerging challenge is the representation gap between these teacher and student models. As models grow in complexity and performance, distilling knowledge effectively becomes more intricate. Studies like TAKD [18] and DAKD [19] reveal a counter-intuitive finding: stronger teacher models do not always translate to better KD performance. TAKD [18] suggest a capacity limit for students learning from teachers, introducing the concept of teaching assistants to bridge this gap. Sequential training of assistants, each smaller than the last, culminates in a final assistant that efficiently trains the student. Building on this, Son et al. [19] propose DAKD, enhancing connectivity between all involved models and allowing students to select optimal teachers per sample. Park et al. [24] introduce a different approach with SFTN, where the teacher model is supervised by the student to minimize the representation gap. While innovative, these methods face practical limitations due to their complexity and computational demands.

This paper offers a simpler and efficient solution using correlation-based loss.

2.2 Advancements in KD for Dense Prediction Tasks

Dense prediction tasks like object detection and semantic segmentation present unique challenges for KD, as they require detailed predictions at the pixel level. Several methods have been proposed to enhance KD in these contexts. Chen et al. [25] were pioneers in applying KD to object detection, focusing on classification logits and regressions. Li et al. [26] identified that feature maps in detection models contain richer semantic information than responses, leading to the distillation of FPN features. However, this approach grapples with the imbalance of foreground and background pixels. Recent methods have aimed to select valuable features and develop tailored loss functions to address this imbalance [27]–[32].

In semantic segmentation, KD techniques prioritize maintaining structural semantic connections. He et al. [33] utilize a pretrained autoencoder for optimizing feature similarity in a latent space, alongside transferring non-local pairwise affinity maps. SKD [34] employs pairwise distillation among pixels and adversarial distillation on score maps. IFVD [35] focuses on transferring intra-class feature variation, while CWD [28] and CIRKD [36] introduce channel-wise and relational distillations respectively.

Despite recent advancements, a major challenge persists: state-of-the-art knowledge distillation (KD) methods are often task-specific, hampering their generalizability and incurring substantial experimental costs. To address this, we extend our DIST approach to DIST+, incorporating feature-level relational distillation. This enhancement proves to be simple yet effective, particularly in dense prediction tasks.

3 Revisiting Prediction Match of KD

In vanilla knowledge distillation [12], the knowledge is transferred from a pre-trained teacher model to a student model by minimizing the discrepancy between the prediction scores of the teacher and student models.

Formally, with the logits $\boldsymbol{Z}^{(\mathrm{s})} \in \mathbb{R}^{B\times C}$ and $\boldsymbol{Z}^{(\mathrm{t})} \in \mathbb{R}^{B\times C}$ of student and teacher networks, where B and C denote batch size and the number of classes, respectively, the vanilla KD loss [12] is represented as

$$\begin{aligned}\mathcal{L}_{\mathrm{KD}} &:= \frac{\tau^2}{B}\sum_{i=1}^{B}\mathrm{KL}\left(\boldsymbol{Y}^{(\mathrm{t})}_{i,:}, \boldsymbol{Y}^{(\mathrm{s})}_{i,:}\right) \\ &= \frac{\tau^2}{B}\sum_{i=1}^{B}\sum_{j=1}^{C} Y^{(\mathrm{t})}_{i,j}\log\left(\frac{Y^{(\mathrm{t})}_{i,j}}{Y^{(\mathrm{s})}_{i,j}}\right),\end{aligned} \tag{1}$$

where KL refers to Kullback–Leibler divergence with

$$\boldsymbol{Y}^{(\mathrm{s})}_{i,:} = softmax\left(\boldsymbol{Z}^{(\mathrm{s})}_{i,:}/\tau\right), \quad \boldsymbol{Y}^{(\mathrm{t})}_{i,:} = softmax\left(\boldsymbol{Z}^{(\mathrm{t})}_{i,:}/\tau\right), \tag{2}$$

being the probabilistic prediction vectors, and τ is the temperature factor to control the softness of logits.

In addition to the teacher's soft targets in Equation (1), KD [12] stated that it is beneficial to train the student together with ground-truth labels, and the overall training loss is composed of the original classification loss $\mathcal{L}_{\mathrm{cls}}$ and KD loss $\mathcal{L}_{\mathrm{KD}}$, *i.e.*,

$$\mathcal{L}_{\mathrm{tr}} = \alpha\mathcal{L}_{\mathrm{cls}} + \beta\mathcal{L}_{\mathrm{KD}}, \tag{3}$$

where $\mathcal{L}_{\mathrm{cls}}$ is usually the cross-entropy loss between the predictions of student network and ground-truth labels, α and β are factors for balancing the losses.

3.1 Catastrophic Discrepancy with A Stronger Teacher

As illustrated in Section 1, the effect of a teacher on KD has not been sufficiently investigated, especially when the performance of pre-trained teacher grows stronger, such as with larger model size or being trained with more advanced and competing strategies, *e.g.*, label smoothing, mixup [20], auto augmentations [37], *etc*. With this regard, as Figure 2, we train ResNet-18 and ResNet-50 standalone with strategy B1 and strategy B2[1], and obtain 4 trained models (*R18B1*, *R18B2*, *R50B1*, and *R50B2* with accuracies 69.76%, 73.4%, 76.13%, and 78.5%, respectively), then compare their discrepancy using KL divergence ($\tau = 1$ and $\tau = 4$) on the predicted probabilities $\boldsymbol{Y}$. We have the following observations:

- The outputs of ResNet-18 do not change much with the stronger strategy compared to ResNet-50. This implies that the representational capacity limits the student's performance, and it tends to be fairly challenging for the student to exactly match the teacher's outputs as their discrepancy becomes larger.
- When the teacher and student models are trained with a stronger strategy, the discrepancy between teacher and student would be larger. This indicates that when we adopt KD with a stronger training strategy, the misalignment between KD loss and classification loss would be severer, thus disturbing the student's training.

As a result, the exact match (*i.e.*, the loss reaches the minimal if and only if the teacher and student outputs are exactly identical) with KL divergence seems way too over-ambitious and demanding since the discrepancy between student and teacher can be considerably huge. Since the exact match can be detrimental with a stronger teacher, our intuition is to develop a relaxed manner for matching the predictions between the teacher and student.

4 DIST: Distillation from A Stronger Teacher

4.1 Relaxed Match with Relations

The prediction scores indicate the teacher's confidence (or preference) over all classes. For a relaxed match of predictions between the teacher and student, we are motivated to consider what we really care about for the teacher's output. Instead of the exact probabilistic values, actually, during inference, we are only concerned about their **relations**, *i.e.*, relative ranks of predictions of teacher.

In this way, for some metric $d(\cdot,\cdot)$ with $\mathbb{R}^C \times \mathbb{R}^C \to \mathbb{R}^+$, the exact match can be formulated that $d(\boldsymbol{a},\boldsymbol{b}) = 0$ if $\boldsymbol{a} = \boldsymbol{b}$

1. Training with B2 obtains higher accuracy compared to B1, *e.g.*, 73.4% (B2) vs. 69.8% (B1) on ResNet-18.

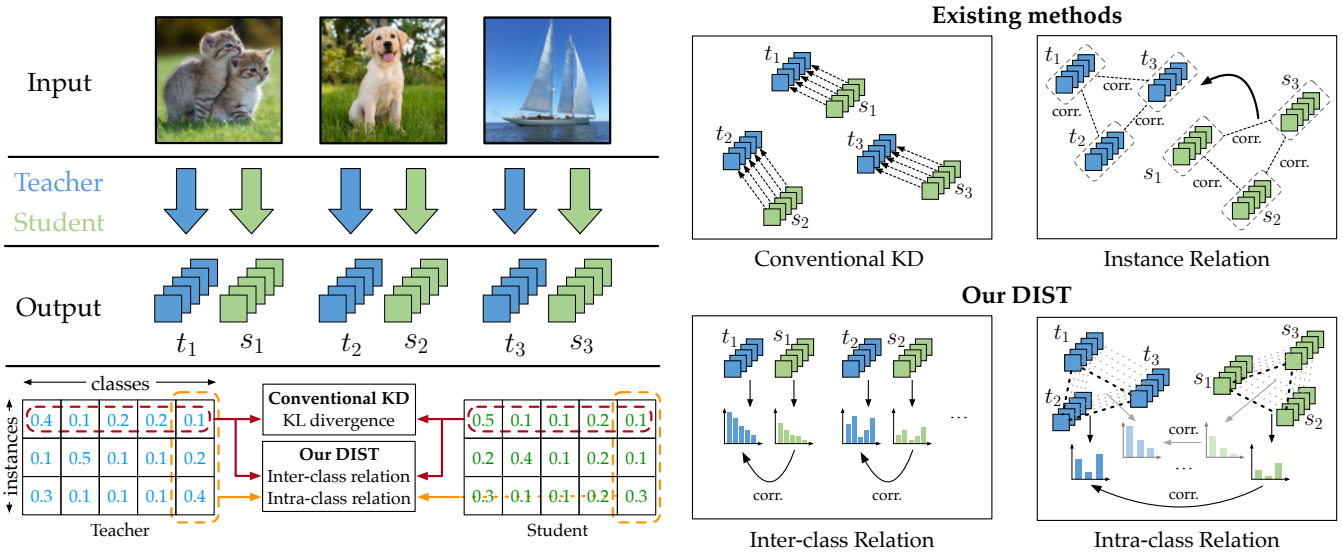

Figure 3. Difference between our DIST and existing KD methods. Conventional KD matches the outputs of student ($s \in \mathbb{R}^5$) to teacher ($t \in \mathbb{R}^5$) point-wisely; instance relation methods operate on the feature level and measure the internal correlations (corr.) between instances in student and teacher separately, then transfer the teacher's correlations to student. Our DIST proposes to maintain the inter-class and intra-class relations between student and teacher. Inter-class relation: correlation between the predicted probabilistic distributions on each instance of teacher and student. Intra-class relation: correlation of the probabilities of all the instances on each class.

for any two prediction vector as $\boldsymbol{Y}^{(\mathrm{s})}_{i,:}$ and $\boldsymbol{Y}^{(\mathrm{t})}_{i,:}$ in the KL divergence of Equation (1). Then as a relaxed match, we can introduce additional mappings $\phi(\cdot)$ and $\psi(\cdot)$ with $\mathbb{R}^C \to \mathbb{R}^C$ such that

$$d(\phi(\boldsymbol{a}), \psi(\boldsymbol{b})) = d(\boldsymbol{a}, \boldsymbol{b}), \forall \boldsymbol{a}, \boldsymbol{b} \tag{4}$$

Therefore, $d(\boldsymbol{a}, \boldsymbol{b}) = 0$ does not necessarily require $\boldsymbol{a}$ and $\boldsymbol{b}$ should be exactly the same. Nevertheless, since we care about the relation within $\boldsymbol{a}$ or $\boldsymbol{b}$, the mappings ϕ and ψ should be isotone and do not affect the semantic information and inference result of the prediction vector.

With this regard, a simple yet effective choice for the isotone mapping is the positive linear transformation, namely,

$$d(m_1\boldsymbol{a} + n_1, m_2\boldsymbol{b} + n_2) = d(\boldsymbol{a}, \boldsymbol{b}), \tag{5}$$

where m_1, m_2, n_1, and n_2 are constants with $m_1 \times m_2 > 0$. As a result, this match could be invariant under separate changes in scale and shift for the predictions. Actually, to satisfy the property Equation (5), we can thus adopt the widely-used Pearson's distance as the metric, *i.e.*,

$$d_p(\boldsymbol{u}, \boldsymbol{v}) := 1 - \rho_p(\boldsymbol{u}, \boldsymbol{v}). \tag{6}$$

$\rho_p(\boldsymbol{u}, \boldsymbol{v})$ is the Pearson correlation coefficient between two random variables $\boldsymbol{u}$ and $\boldsymbol{v}$,

$$\begin{aligned} \rho_p(\boldsymbol{u}, \boldsymbol{v}) &:= \frac{\mathrm{Cov}(\boldsymbol{u}, \boldsymbol{v})}{\mathrm{Std}(\boldsymbol{u})\mathrm{Std}(\boldsymbol{v})} \\ &= \frac{\sum_{i=1}^{C}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{C}(u_i - \bar{u})^2 \sum_{i=1}^{C}(v_i - \bar{v})^2}} \end{aligned} \tag{7}$$

where $\mathrm{Cov}(\boldsymbol{u}, \boldsymbol{v})$ is the covariance of $\boldsymbol{u}$ and $\boldsymbol{v}$, $\bar{u}$ and $\mathrm{Std}(\boldsymbol{u})$ denote the mean and standard derivation of $\boldsymbol{u}$, respectively.

In this way, we can define the **relation as correlation**. More specifically, and the original exact match in vanilla KD [12] can thus be relaxed and replaced by maximizing the linear correlation to preserve the relation of teacher and student on the probabilistic distribution of each instance, which we call *inter-class relation*. Formally, for each pair of prediction vector $\boldsymbol{Y}^{(\mathrm{s})}_{i,:}$ and $\boldsymbol{Y}^{(\mathrm{t})}_{i,:}$, the inter-relation loss can be formulated as

$$\mathcal{L}_{\text{inter}} := \frac{1}{B}\sum_{i=1}^{B} d_p\left(\boldsymbol{Y}^{(\mathrm{s})}_{i,:}, \boldsymbol{Y}^{(\mathrm{t})}_{i,:}\right). \tag{8}$$

Some isotone mappings or metrics can also be used to relax the match as Equation (4), such as cosine similarity investigated empirically in Section 6.4; other more advanced and delicate choices could be left as future work.

4.2 Better Distillation with Intra-relations

Besides the inter-class relation, where we transfer the relation of multiple classes in each instance, the prediction scores of multiple instances in each class are also informative and useful. This scores indicate the similarities of multiple instances to one class. For instance, suppose we have three images containing "cat", "dog", and "plane", respectively, and they have three prediction scores on the 'cat' class, denoted as e, f, and g. Generally, the picture "cat" should have the largest score to the "cat" class, while the "plane" should have the smallest score since it is inanimate. This relation of "$e > f > g$" could also be transferred to the student. Besides, even for the images from the same class, the intrinsic intra-class variance of the semantic similarities is actually also informative. It indicates the prior from the teacher that *which one is more reliable to cast in this class*.

Therefore, we also encourage to distill this *intra-relation* for better performance. Actually, define the prediction matrix $\boldsymbol{Y}^{(\mathrm{s})}$ and $\boldsymbol{Y}^{(\mathrm{t})}$ with each row as $\boldsymbol{Y}^{(\mathrm{s})}_{i,:}$ and $\boldsymbol{Y}^{(\mathrm{t})}_{i,:}$, then the above inter-relation is to maximize the correlation rowwise (see Figure 3). In contrast, for intra-relation, the correspond-

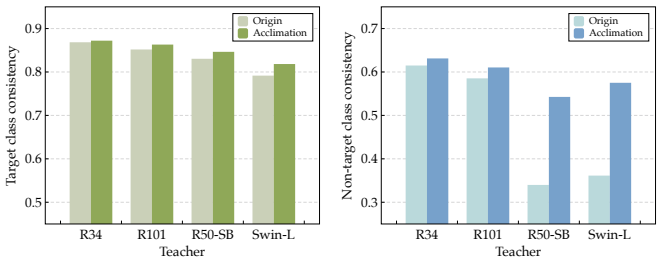

Figure 4. Statistics of the (a) target class consistency and (b) non-target class consistency on ResNet-18 student. We compare the original predictions of teachers with our acclimated predictions on baseline teachers (R34 and R101) and stronger teachers (R50-SB [38] and Swin-L).

ing loss is thus to maximize the correlation column-wisely, *i.e.*,

$$\mathcal{L}_{\text{intra}} := \frac{1}{C}\sum_{j=1}^{C} d_p\left(\boldsymbol{Y}^{(\text{s})}_{:,j}, \boldsymbol{Y}^{(\text{t})}_{:,j}\right). \tag{9}$$

As a result, the overall training loss $\mathcal{L}_{\text{tr}}$ of DIST can be composed of the classification loss, inter-class KD loss, and intra-class KD loss, *i.e.*,

$$\mathcal{L}_{\text{tr}} = \lambda_1\mathcal{L}_{\text{cls}} + \lambda_2\mathcal{L}_{\text{inter}} + \lambda_3\mathcal{L}_{\text{intra}}, \tag{10}$$

where λ_1, λ_2, and λ_3 are factors for balancing the losses. In this way, via the relation loss, we have endowed the student with freedom more or less to match the teacher network's output adaptively, thus boosting the distillation performance to a great extent.

5 DIST+ for Better Relaxed Match

5.1 On the Essential Discrepancy with Stronger Teacher

In the previously discussed DIST method, we redefined the distillation process as a relaxed objective. This approach requires the student model to match the teacher model's predictions through relational correspondence. Building upon this, DIST+ delves deeper into analyzing the internal relationships within the predictions of both the teacher and student models.

For classification tasks, accuracy is a common evaluation metric. This metric typically involves ranking the predicted probabilities, then checking if the class with the highest predicted probability (top-1) matches the ground truth. To assess the disparity between the student's and teacher's predictions, we divide the rank vectors $\boldsymbol{R}$ of the predicted probabilities $\boldsymbol{Y}$ into two categories: the ranks of the target class $\boldsymbol{P}$ with dimensions $[B, 1]$, and the ranks of the non-target class ranks $\boldsymbol{M}$ with dimensions $[B, C-1]$. We then introduce two types of consistencies to quantify these differences.

Target class consistency. This metric quantifies the extent to which the student and teacher models agree on the ranks of the target class. It is expressed as the percentage of samples for which both models assign the same rank to the target class. Formally, the Target Class Consistency (TC) is defined as:

$$TC\left(\boldsymbol{P}^{(s)}, \boldsymbol{P}^{(t)}\right) := \frac{1}{B}\sum_{i=1}^{B}\mathbb{1}\left(P_i^{(s)}, P_i^{(t)}\right), \tag{11}$$

where $\mathbb{1}$ is an indicator function that yields 1 when $P_i^{(s)}$ and $P_i^{(t)}$ are equal, and 0 otherwise.

Table 1
Non-target class consistencies between distilled student and teachers

Method	Baseline teacher		Stronger teacher	
	R34	R101	R50-SB	Swin-L
w/o KD	0.61	0.59	0.34	0.36
KL div.	0.67	0.63	0.40	0.39
DIST	0.73	0.66	0.45	0.48

Non-target class consistency. As we have established, for non-target classes, it is crucial to ascertain if the student and teacher models maintain consistent rank relationships. To evaluate this, we employ the Spearman's rank correlation coefficient as a measure of non-target class consistency. This can be mathematically represented as:

$$NC\left(\boldsymbol{M}^{(s)}, \boldsymbol{M}^{(t)}\right) := \frac{1}{B}\sum_{i=1}^{B}\rho_s\left(\boldsymbol{M}_i^{(s)}, \boldsymbol{M}_i^{(t)}\right), \tag{12}$$

with ρ representing Spearman's rank correlation, calculated as

$$\rho_s(\boldsymbol{r}_a, \boldsymbol{r}_b) := \frac{\text{Cov}(\boldsymbol{r}_a, \boldsymbol{r}_b)}{\text{Std}(\boldsymbol{r}_a), \text{Std}(\boldsymbol{r}_b)}, \tag{13}$$

where $\boldsymbol{r}_a$ and $\boldsymbol{r}_b$ are rank vectors, *Cov* represents the covariance between these rank vectors, and *Std* denotes the standard derivation.

In Figure 4, we assess target and non-target class consistencies using the independently-trained ResNet-18 as the student and various pretrained models as teachers. The predictions on the ImageNet validation set are used to calculate these consistencies, as indicated by the *Origin* legend in the figure. Our analysis yields two key insights:

- **High target class consistency across teachers:** Both sets of teachers — those trained with the baseline strategy (R34 and R101) and those with the stronger strategy (R50-SB [38] and Swin-L) — demonstrate high target class consistency with the student model. This observation implies that the decrease in knowledge distillation performance observed with stronger teachers is likely not due to discrepancies in target class predictions.
- **Significant disparity in non-target class consistency:** A pronounced difference is evident in the non-target class consistency between baseline and stronger teachers. For example, the non-target class consistency is recorded at 0.61 for the R34 teacher, but only 0.34 for the R50-SB teacher. This substantial divergence suggests that the differences between the student and stronger teachers in non-target class predictions may mislead the optimization process in knowledge distillation, leading to diminished performance.

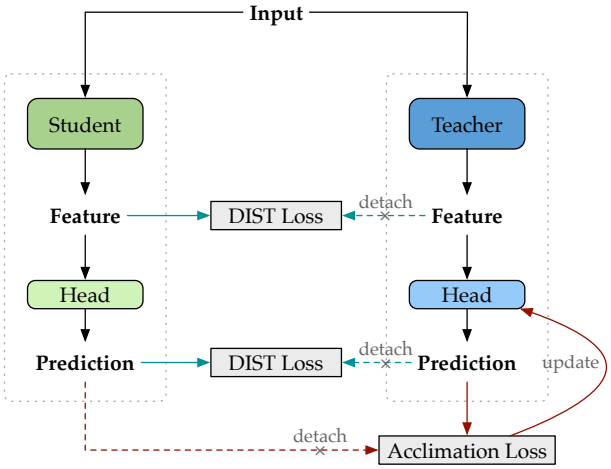

Figure 5. Framework of DIST+. We extend DIST with an additional distillation loss on intermediate features, and an acclimation loss for aligning the teacher predictions to student predictions.

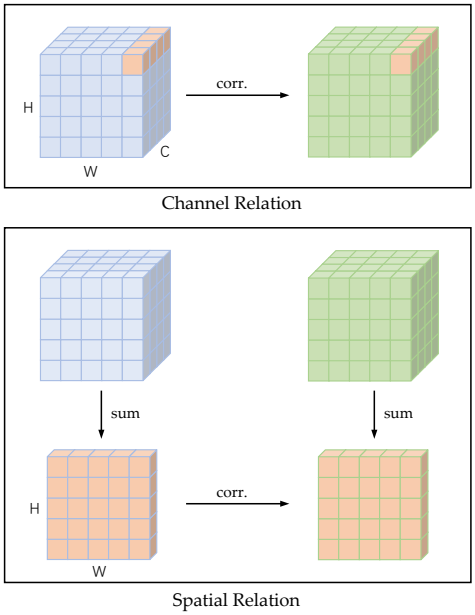

Figure 6. Illustration of channel relation and feature relation in DIST+.

Consequently, to optimize the effectiveness of knowledge distillation, it is crucial to address the discrepancies in the non-target class predictions between the teacher and student models.

5.2 How Much Does Student Learn from the Non-target Classes?

KD methods applied to classification logits frequently involve distillation focusing on non-target classes. A pertinent question arises: *Does this type of supervision aid in reducing the discrepancy between the student and stronger teachers in terms of non-target class predictions?* Contrary to what one might expect, our findings indicate a negative response to this query. As detailed in Table 1, we evaluated the non-target class consistency between the distilled student models and their respective teachers. Both the traditional approach using KL divergence in vanilla KD and our DIST method seem insufficient in narrowing the gap for non-target classes between student and teacher models. Even after distillation, the consistency levels with stronger teachers remain substantially lower compared to those with baseline teachers. This persistent discrepancy could be due to the student model's limited capacity in assimilating the complex information from the more advanced teachers.

Given the notable discrepancy observed in non-target class predictions and the challenges inherent in mitigating it, a question arises: *Is it feasible to omit distillation on non-target classes and focus solely on learning from the teacher's target class predictions?* Recent research, such as DKD [39], however, indicates that relying exclusively on target class knowledge distillation (TCKD) may be counterproductive. It suggests that TCKD alone might not only be unhelpful but could potentially impair performance. The distillation of non-target class information (NCKD) is deemed crucial for the effective transfer of knowledge from a superior teacher model. Therefore, to enhance knowledge distillation from stronger teachers, it becomes imperative to improve both the non-target class consistency and the overall distillation process for non-target classes.

5.3 Acclimating Teacher from Student

In a fortunate turn, our research has led to an intriguing insight. Although students face significant challenges in assimilating the non-target class preferences from stronger teachers, we have discovered a promising avenue. As shown in Figure 4, we fine-tune the last backbone layer and head in teacher models using our proposed acclimation loss (which will be introduced later), and gain significant increments on non-target class-consistency. It appears that these stronger teachers can be effectively adjusted to provide non-target class predictions that are more conducive to student learning. Remarkably, this adaptation can be achieved with minimal impact on the overall accuracy of the teacher model.

This discovery opens up a new dimension in the field of knowledge distillation. Recognizing the possibility of modifying stronger teachers to align better with student models, we explore a methodical approach to facilitate this adaptation. Our strategy focuses on fine-tuning specific components of the teacher model, using the student's predictions as a guiding framework. This approach is grounded in the understanding that slight adjustments in key areas of the teacher model can significantly enhance the student's learning process. By doing so, we aim to harmonize the teacher's expertise with the student's learning capacity, particularly in the context of non-target class predictions.

In our approach, we begin by initializing the teacher model with its pretrained weights. The key step involves fine-tuning the last convolution layer and classification head of the teacher model, guided by the student's predictions during the distillation process. This fine-tuning is critical for aligning the teacher's non-target class probabilities with those of the student. Formally, let $\boldsymbol{E}^{(t)}$ represent the teacher's non-target class probabilities and $\boldsymbol{E}^{(s)}$ denote the corresponding probabilities from the student, both having the shape $[B, C-1]$. The acclimation process is designed to modify the teacher model so that its non-target class relationships mirror those of the student. This alignment is

Table 2
Training Strategies on Image Classification Tasks. *BS*: batch size; *LR*: learning rate; *WD*: weight decay; *LS*: label smoothing; *EMA*: model exponential moving average; *RA*: RandAugment [37]; *RE*: random erasing; *CJ*: color jitter

Strategy	Dataset	Epochs	Total BS	Initial LR	Optimizer	WD	LS	EMA	LR scheduler	Data augmentation
A1	CIFAR-100	240	64	0.05	SGD	5×10^{-4}	-	-	$\times 0.1$ at 150,180,210 epochs	crop + flip
B1	ImageNet	100	256	0.1	SGD	1×10^{-4}	-	-	$\times 0.1$ every 30 epochs	crop + flip
B2	ImageNet	450	768	0.048	RMSProp	1×10^{-5}	0.1	0.9999	$\times 0.97$ every 2.4 epochs	$\{B1\}$ + RA + RE
B3	ImageNet	300	1024	5e-4	AdamW	5×10^{-2}	0.1	-	cosine	$\{B2\}$ + CJ + Mixup + CutMix

Table 3
Evaluation results of baseline settings on ImageNet. We use ResNet-34 and ResNet-50 released by Torchvision [40] as our teacher networks, and follow the standard training strategy (B1).

Student (teacher)		Teacher	Student	KD [12]	OFD [13]	CRD [15]	SRRL [41]	Review [42]	DIST	DIST+
ResNet-18 (ResNet-34)	Top-1	73.31	69.76	70.66	71.08	71.17	71.73	71.61	72.07	**72.39**
	Top-5	91.42	89.08	89.88	90.07	90.13	90.60	90.51	90.42	90.67
MobileNet (ResNet-50)	Top-1	76.16	70.13	70.68	71.25	71.37	72.49	72.56	73.24	**73.47**
	Top-5	92.86	89.49	90.30	90.34	90.41	90.92	91.00	91.12	91.22

quantified using Pearson's distance, as described in Equation (6), specifically:

$$\mathcal{L}_{\mathrm{ta}} := \frac{1}{B} d_p \left(\boldsymbol{E}^{(s)}, \boldsymbol{E}^{(t)} \right). \tag{14}$$

This equation represents the loss function used to acclimate the teacher model, thereby optimizing it to produce non-target class predictions that are more in sync with the student's learning pattern.

We implement the acclimation of the teacher model during the distillation process. Specifically, as shown in Figure 5, in each iteration, we utilize the predictions from both the teacher and student models to compute the teacher acclimation loss function $\mathcal{L}_{\mathrm{ta}}$. This loss is then used to generate gradients that are backpropagated exclusively through the teacher model. It's important to note that the gradients generated by $\mathcal{L}_{\mathrm{ta}}$ are not employed in optimizing the student model. Instead, their sole purpose is to fine-tune the teacher model, ensuring that it becomes more conducive to the student's learning process.

5.4 Feature-level Relaxed Relation Match

In our described method, KD is conducted through an analysis and harmonization of the relationships in teacher and student predictions. However, it is widely recognized that intermediate features, as opposed to merely predicted probabilities, harbor richer information which can lead to a more nuanced and effective distillation process [15], [30], [42], [43]. In DIST+, we expand our approach from logit-level distillation to include feature-level distillation. Our findings indicate that incorporating feature-level relationships effectively enhances the performance of the distillation.

In the realm of feature distillation, we identify and utilize two key axes of relationships: channel relation and spatial relation. As illustrated in Figure 6, channel relation focuses on transferring the inter-channel dynamics from the teacher model to the student model. In contrast, spatial relation deals with aggregating responses across channels to capture the overall spatial response patterns, which are then transferred from the teacher to the student. This dual approach ensures that both the nuanced inter-channel interactions and the broader spatial response patterns are effectively distilled into the student model.

Mathematically, we define $\boldsymbol{F}^{(t)}$ as the feature map used for distillation in the teacher model, and $\boldsymbol{F}^{(s)}$ as the corresponding feature map in the student model. Both feature maps share the same dimensional structure, represented as $[B, D, H, W]$, where D denotes the channel dimension, and H and W represent the height and width of the feature map, respectively. The computation of the two distinct types of feature relations, channel relation and spatial relation, is carried out based on these feature maps.

Channel relation loss. The channel relation loss is computed using the Pearson distance across the channel dimension:

$$\mathcal{L}_{\mathrm{cr}} := \frac{1}{BHW} \sum_{i=1}^{B} \sum_{j=1}^{H} \sum_{k=1}^{W} d_p \left(\boldsymbol{F}^{(s)}_{i,:,j,k}, \boldsymbol{F}^{(t)}_{i,:,j,k} \right). \tag{15}$$

Spatial relation loss. The spatial relation loss, on the other hand, is computed on the aggregated feature across the spatial dimension. First, we aggregate each feature map over the channel dimension and then reshape these aggregated features into $\hat{\boldsymbol{F}}^{(s)}$ and $\hat{\boldsymbol{F}}^{(t)}$, each with the shape $[B, HW]$. The spatial relation loss is then calculated as

$$\mathcal{L}_{\mathrm{sr}} := \frac{1}{B} \sum_{i=1}^{B} d_p \left(\hat{\boldsymbol{F}}^{(s)}_i, \hat{\boldsymbol{F}}^{(t)}_i \right). \tag{16}$$

This approach ensures a comprehensive alignment of the student's feature map with that of the teacher, both in terms of channel-wise relationships and overall spatial patterns.

Table 4
Performance of ResNet-18 and ResNet-34 on ImageNet with different sizes of teachers

Student	Teacher	Top-1 ACC (%)				
		student	teacher	KD	DIST	DIST+
ResNet-18	ResNet-34	69.76	73.31	71.21	72.07	**72.39**
	ResNet-50		76.13	71.35	72.12	**72.46**
	ResNet-101		77.37	71.09	72.08	**72.43**
	ResNet-152		78.31	71.12	72.24	**72.55**
ResNet-34	ResNet-50	73.31	76.13	74.73	75.06	**75.23**
	ResNet-101		77.37	74.89	75.36	**75.48**
	ResNet-152		78.31	74.87	75.42	**75.57**

Table 5
Top-1 accuracies (%) of students trained with strong strategies on ImageNet. The *Swin-T* is trained with strategy B3 in Table 2, others are trained with B2. ResNet-50 and Swin-L teachers have 80.1% and 86.3% accuracies, respectively. †: trained by [38]. ‡: Pretrained on ImageNet-22K

Teacher	Student	w/o KD	KD	RKD	SRRL	DIST	DIST+
ResNet-50†	ResNet-18	73.4	72.6	72.9	71.2	74.5	**74.8**
	ResNet-34	76.8	77.2	76.6	76.7	77.8	**78.0**
	MobileNetV2	73.6	71.7	73.1	69.2	74.4	**74.7**
	EfficientNet-B0	78.0	77.4	77.5	77.3	78.6	**78.7**
Swin-L‡	ResNet-50	78.5	80.0	78.9	78.6	80.2	**80.4**
	Swin-T	81.3	81.5	81.2	81.5	82.3	**82.6**

As a result, the overall loss function for training the student in DIST+ is as follows:

$$\mathcal{L}_{\mathrm{tr}} = \lambda_1 \mathcal{L}_{\mathrm{cls}} + \lambda_2 \mathcal{L}_{\mathrm{inter}} + \lambda_3 \mathcal{L}_{\mathrm{intra}} + \lambda_4 \mathcal{L}_{\mathrm{cr}} + \lambda_5 \mathcal{L}_{\mathrm{sr}}, \quad (17)$$

where λ is the loss weight for balancing the losses.

6 Experiments

6.1 Image Classification

Settings. Training strategies. The training strategies of image classification task are summarized in Table 2. **CIFAR-100.** For fair comparisons, we use the same training strategies (referred to *A1* in Table 2) and pretrained models following CRD [15]. **ImageNet.** B1: for comparisons with previous KD methods, we train our baselines with the same simple training strategy as CRD [15]. B2: to validate the effectiveness of KD methods on modern training strategies, we follow EfficientNet [44] and design a training strategy B2, which can significantly improve the performance compared to B1. B3: the strategy B3 is used for training Swin-Transformers [23], and contains even more stronger data augmentations and regularization.

Loss weights. On CIFAR-100 and ImageNet, we set $\lambda_1 = 1$, $\lambda_2 = 2$, and $\lambda_3 = 2$ in Eq.(10) for DIST, and additional $\lambda_4 = \lambda_5 = 1$ for DIST+. On object detection and semantic segmentation, these factors are all equal to 1. For KD [12], we set $\alpha = 0.9$, $\beta = 1$ in Eq.(3), and use a default temperature $\tau = 4$. Specifically, instead of using $\tau = 1$ on ImageNet, we choose a larger temperature $\tau = 4$ on CIFAR-100, as it is easy to get overfit and the learned probabilistic distribution is sharp on CIFAR-100.

Baseline results on ImageNet. We first compare our method with prior works using the baseline settings. As shown in Table 3, our DIST and DIST+ significantly outperforms prior KD methods. Note that our method is only conducted on the outputs of models, and has a similar computational cost as KD [12]. Nevertheless, it even achieves better performance compared to those sophisticatedly-designed methods. For example, CRD [15] needs to preserve a memory bank for all 128-d features of ImageNet images, and produces additional 260M FLOPs of computation cost; SRRL [41] and Review [42] require additional convolutions for feature alignments.

Distillation from stronger teacher models. As the stronger teachers come from larger model sizes and stronger strategies, we here first conduct experiments to compare our DIST with the vanilla KD on different scales (model sizes) of ResNets with baseline strategy B1. As shown in Table 4, when the teacher goes larger, the ResNet-18 students perform even worse than that with a medium-sized ResNet-50 teacher. Nevertheless, our DIST shows an upward trend with larger teachers, and the improvements compared to KD also become more significant, indicating that our DIST tackles better on the large discrepancy between the student and larger teacher.

Distillation from stronger training strategies. Recently, the performance of models on ImageNet has been significantly improved by the sophisticated training strategies and strong data augmentations (*e.g.*, TIMM [38] achieves 80.4% accuracy on ResNet-50 while the baseline strategy B1 only obtains 76.1%). However, most of the KD methods still conduct experiments with simple training settings. It is seldomly investigated whether the KD methods are suitable to the advanced strategies. In this way, we conduct experiments with advanced training strategies and compare our method with vanilla KD, instance relation-based RKD [45], and SRRL [41].

We first train traditional CNNs with strong strategies, and also use a strong ResNet-50 with 80.1% accuracy trained by [38] as the teacher. As results shown in Table 5, on both similar architectures (ResNet-18, ResNet-34) and dissimilar architectures (MobileNetV2, EfficientNet-B0), our DIST can achieve the best performance. Note that RKD and SRRL can perform worse than training from scratch, especially when the students are small (ResNet-18 and MobileNet) or the architectures of teacher and student are fairly different (ResNet-50 and Swin-L), this might be because they focus on the intermediate features, which can be more challenging for the student to recover teacher's features compared to predictions.

Furthermore, we experiment on the recent state-of-the-art Swin-Transformer [23]. The results show that our DIST gains improvements on even more stronger models and strategies. For example, with Swin-L teacher, our method improves ResNet-50 and Swin-T by 1.7% and 1.0%, respectively. Moreover, the extended version, DIST+, achieves significant improvements compared to DIST. This demonstrates the effects of DIST+ with teacher acclimation and feature-level relaxed distillation loss.

CIFAR-100. The results on CIFAR-100 dataset in Table 6 show that, by distilling on the predicted logits, our DIST outperforms most of the sophisticatedly-designed feature

Table 6
Evaluation results on CIFAR-100 dataset. The upper and lower models denote teacher and student, respectively. Results of the compared methods are reported by CRD [15]

Method	Same architecture style			Different architecture style		
	WRN-40-2 WRN-40-1	ResNet-56 ResNet-20	ResNet-32x4 ResNet-8x4	ResNet-50 MobileNetV2	ResNet-32x4 ShuffleNetV1	ResNet-32x4 ShuffleNetV2
Teacher	75.61	72.34	79.42	79.34	79.42	79.42
Student	71.98	69.06	72.50	64.6	70.5	71.82
FitNet [46]	72.24±0.24	69.21±0.36	73.50±0.28	63.16±0.47	73.59±0.15	73.54±0.22
CC [14]	72.21±0.25	69.63±0.32	72.97±0.17	65.43±0.15	71.14±0.06	71.29±0.38
VID [47]	73.30±0.13	70.38±0.14	73.09±0.21	67.57±0.28	73.38±0.09	73.40±0.17
RKD [45]	72.22±0.20	69.61±0.06	71.90±0.11	64.43±0.42	72.28±0.39	73.21±0.28
PKT [48]	73.45±0.19	70.34±0.04	73.64±0.18	66.52±0.33	74.10±0.25	74.69±0.34
AB [13]	72.38±0.31	69.47±0.09	73.17±0.31	67.20±0.37	73.55±0.31	74.31±0.11
FT [49]	71.59±0.15	69.84±0.12	72.86±0.12	60.99±0.37	71.75±0.20	72.50±0.15
CRD [15]	74.14±0.22	71.16±0.17	75.51±0.18	69.11±0.28	75.11±0.32	75.65±0.10
KD [12]	73.54±0.20	70.66±0.24	73.33±0.25	67.35±0.32	74.07±0.19	74.45±0.27
DIST	74.73±0.24	71.75±0.30	76.31±0.19	68.66±0.23	76.34±0.18	77.35±0.25
DIST+	74.82±0.27	72.01±0.23	76.19±0.36	68.72±0.15	76.55±0.35	77.28±0.33

Table 7
Results on COCO validation set. T: teacher; S: student. *: We implement KD using $\tau = 1$ and other settings are the same as DIST

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage detectors						
T: Cascade Mask RCNN-X101	45.6	64.1	49.7	26.2	49.6	60.0
S: Faster RCNN-R50	38.4	59.0	42.0	21.5	42.1	50.3
KD [12]*	39.7	61.2	43.0	23.2	43.3	51.7
FKD [50]	41.5	62.2	45.1	23.5	45.0	55.3
CWD [28]	41.7	62.0	45.5	23.3	45.5	**55.5**
DIST	40.4	61.7	43.8	**23.9**	44.6	52.6
DIST + mimic	41.8	62.4	45.6	23.4	46.1	55.0
DIST+	**42.1**	**62.6**	**46.2**	23.6	**46.4**	55.2
One-stage detectors						
T: RetinaNet-X101	41.0	60.9	44.0	23.9	45.2	54.0
S: RetinaNet-R50	37.4	56.7	39.6	20.0	40.7	49.7
KD [12]*	37.2	56.5	39.3	20.4	40.4	49.5
FKD [50]	39.6	58.8	42.1	22.7	43.3	52.5
CWD [28]	40.8	**60.4**	43.4	22.7	44.5	**55.3**
DIST	39.8	59.5	42.5	22.0	43.7	53.0
DIST + mimic	40.1	59.4	43.0	23.2	44.0	53.6
DIST+	**41.0**	60.3	**43.7**	**23.3**	**44.6**	55.1

Table 8
Results on Cityscapes val dataset. All models are pretrained on ImageNet

Method	mIoU (%)
T: DeepLabV3-R101	78.07
S: DeepLabV3-R18	74.21
SKD [51]	75.42
IFVD [35]	75.59
CWD [28]	75.55
CIRKD [36]	76.38
DIST	77.10
DIST+	**77.36**
S: PSPNet-R18	72.55
SKD [51]	73.29
IFVD [35]	73.71
CWD [28]	74.36
CIRKD [36]	74.73
DIST	76.31
DIST+	**76.52**

distillation methods. While the extended version with feature distillation, DIST+, achieves an overall better performance than the logits-only DIST.

6.2 Object Detection

Settings. We further investigate the effectiveness of DIST on downstream tasks. We conduct experiments on MS COCO object detection dataset [52], and simply leverage our DIST as an additional supervision on the final predictions of classes.

Training strategies. Following [28], [50], we use the same standard training strategies (the official $2\times$ schedule in MMDetection [53]) and utilize Cascade Mask R-CNN [7] with ResNeXt-101 backbone as the teacher for two-stage student of Faster R-CNN [6] with ResNet-50 backbone; while for one-stage RetinaNet [54] with ResNet-50 backbone, the RetinaNet with ResNeXt-101 backbone is utilized as the teacher. All loss weights of KD are set to 1.

As shown in Table 7, our DIST achieves competitive results on COCO validation set. For comparisons, we train the vanilla KD under the same settings as our DIST, the results show that our DIST significantly outperforms vanilla KD by simply replacing the loss functions. Moreover, by combining DIST with mimic, which minimizes the mean square error between FPN features of teacher and student, we can even outperform the state-of-the-art KD methods designed for object detection. Note that by conducting feature distillation in our DIST+, we achieve further improvements compared to *DIST + mimic*. For instance, we obtain 42.1 AP on Faster-RCNN ResNet-50 student, significantly improving without-KD baseline by 3.7.

6.3 Semantic Segmentation

Settings. We also perform experiments on semantic segmentation, a challenging dense prediction task. Following [28], [35], [36], we train DeepLabV3 [55] and PSPNet [8] with ResNet-18 backbone on Cityscapes dataset, and adopt our DIST on the predictions of classification head using a teacher with ResNet-101 backbone of DeepLabV3.

Table 9
Ablation of inter-class and intra-class relations on ImageNet. The student and teacher models are ResNet-18 and ResNet-34, respectively

Method	Inter	Intra	ACC (%)
KD	-	-	71.21
DIST (KL div.)	✗	✓	70.61
DIST (KL div.)	✓	✓	71.62
DIST	✓	✗	71.63
DIST	✗	✓	71.55
DIST	✓	✓	**72.07**

Table 10
Comparisons of training KD with or without the classification loss on ImageNet. The student and teacher models are ResNet-18 and ResNet-34, respectively. The original accuracy of ResNet-18 without KD is 69.76%

Method	w/ cls. loss	w/o cls. loss
KD	71.21	68.12
DIST	**72.07**	**70.65**

Training strategies. Following CIRKD [36], we adopt a standard data augmentation, which consists of random flipping, random scaling in the range of $[0.5, 2]$, and a crop size of 512×1024. We train the models using an SGD optimizer with a momentum of 0.9, and a polynomial annealing learning rate scheduler is adopted with an initial value of 0.02. We train the mask tokens for 2000 iterations in the mask learning stage, and then train the student for 40000 iterations. All the loss weights in KD are set to 1.

As the results summarized in Table 8, with only the supervision of class predictions, our DIST can significantly outperform existing knowledge distillation methods on semantic segmentation task. For example, our DIST outperforms recent state-of-the-art method CIRKD [36] by 1.58% on PSPNet-R18. This demonstrate our effectiveness on relation modeling. Furthermore, the improved DIST+, with feature distillation, further outperforms DIST by 0.26% on DeepLabV3-R18 student.

6.4 Ablation Studies

Effects of inter-class and intra-class correlations. This paper proposes two types of relations: inter-class and intra-class relations. To validate the effectiveness of each relation, we conduct experiments to train students with these relations separately. The results on Table 9 verify that, both inter-class and intra-class relations can outperform the vanilla KD; also, the performance could be further boosted by combining them together.

Effect of intra-class relation in vanilla KD. To investigate the effectiveness of intra-class relation in vanilla KD, we adopt experiments to train our DIST using KL divergence as the relation metric, denoted as *DIST (KL div.)*[2]. As the results summarized in Table 9, adding intra-class relation in the vanilla KD can also improve the performance (from 71.21% to 71.62%). However, when the student is trained with intra-class relation only, the improvement of using KL divergence is less significant than using Pearson correlation (70.61% vs. 71.55%), since the means and variances of intra-class distributions could be varied.

2. Specifically, the vanilla KD is the same as DIST (KL div.) with inter-class relation only.

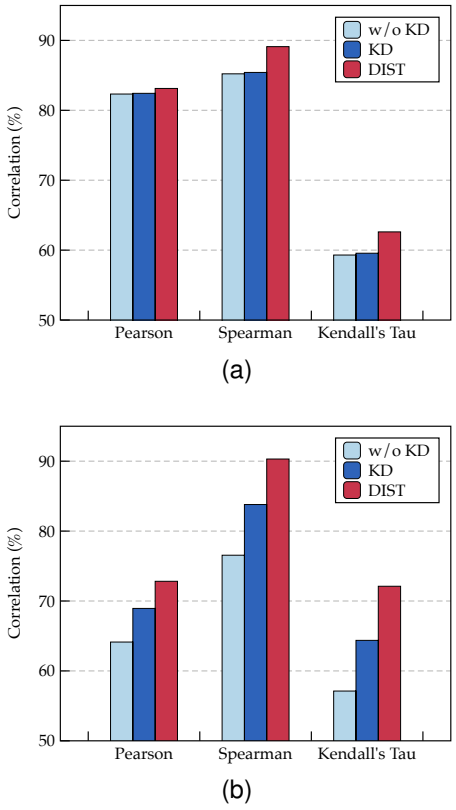

Figure 7. (a) Inter-class correlation and (b) intra-class correlation between ResNet-18 student and ResNet-34 teacher. We train the methods on ImageNet with B1 strategy.

Effect of training students with KD loss only. Training student with only the KD loss can better reflect the distillation ability and the information richness of supervision signals. As results in Table 10 show that, when the student is trained with only the KD loss, our DIST significantly outperforms the vanilla KD. Without using the ground-truth labels, it can even outperform the standalone training accuracy, which indicates the effectiveness of our DIST in distilling those truly-beneficial relations.

Correlations between teacher and student. To validate the effectiveness of our correlation-based loss, we measure the correlations between teacher and student models, where the student models are trained by plain classification loss, KD, and our DIST. We choose commonly used Pearson correlation coefficient, Spearman's [56] and Kendall's Tau [57] rank correlation coefficients as the metrics of correlation. As summarized in Figure 7, our DIST obtains higher inter-class and intra-class correlations compared to baselines.

Using cosine similarity in DIST. In our method, the relation matching can be any function with the same form of Eq.(4). We simply adopt a commonly used Pearson correlation as our relation metric in DIST. Here we conduct experiments to investigate the efficacy of our method with cosine similarity.

Both cosine similarity and Pearson correlation coefficient

Table 11
Ablation of cosine similarity and Pearson correlation in DIST. We train the student ResNet-18 and teacher ResNet-34 on ImageNet with or without label smoothing (LS)

Method	w/o LS	w/ LS
Teacher	73.31	73.78
KD ($\tau = 4$)	71.21	70.71
KD ($\tau = 1$)	71.49	71.37
DIST (cosine)	71.79	71.63
DIST (Pearson)	**72.07**	**72.18**

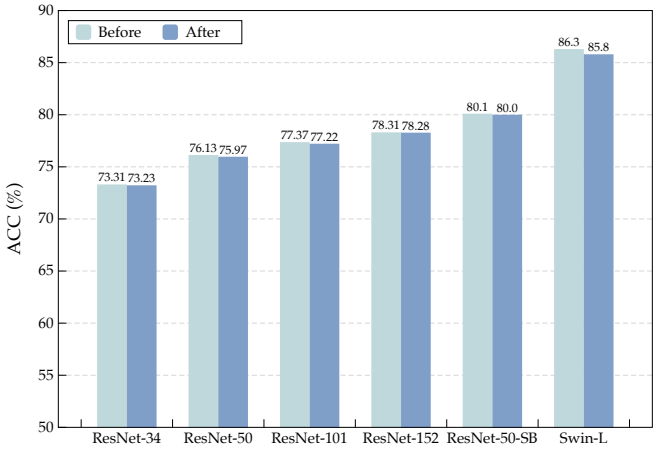

Figure 8. Top-1 accuracies of teachers before and after acclimation.

can evaluate the relations between teacher and student. Compared to the scale invariance in cosine similarity, the Pearson correlation has an additional shift-invariance by centering the vectors first (see Eq.(7)), and it could be more robust to the distribution changes. We conduct experiments to compare these two metrics in our DIST and train the models with or without label smoothing. Since recent studies [58], [59] state that KD with high temperatures is incompatible with label smoothing, we also train the models with KD ($\tau = 1$). As shown in Table 11, adopting DIST with Pearson correlation achieves higher accuracies compared to KD and DIST with cosine similarity, especially when the teacher and student are trained with label smoothing (the predicted probabilistic distributions would be shifted by it). As a result, the scale-and-shift-invariant Pearson correlation may be a better metric for measuring relations in DIST.

6.5 Further Ablation Studies for DIST+

Performance comparison of teachers before and after acclimation. In Figure 8, we summarize the accuracy changes of teachers in Table 4 and Table 5 after acclimation. We can infer that, though the non-target consistencies of them improve significantly after acclimation (see Figure 4), there are only minor accuracy drops. This indicates that the more powerful teachers are easy to adapt student predictions. Besides, the teacher accuracy drops are negilible and our DIST+ can obtain better distillation performance.

Effects of losses in DIST+. In DIST+, we propose teacher acclimation loss $\mathcal{L}_{ta}$ for aligning the teacher from the student, channel relation loss $\mathcal{L}_{cr}$ and spatial relation loss $\mathcal{L}_{sr}$ to conduct relaxed match in intermediate features. We

Table 12
Ablation of the proposed losses in DIST+. The student and teacher models are ResNet-18 and ResNet-34, respectively

Method	$\mathcal{L}_{ta}$	$\mathcal{L}_{cr}$	$\mathcal{L}_{sr}$	ACC (%)
KD	-	-	-	71.21
DIST	-	-	-	72.07
DIST+	✔	✘	✘	72.21
DIST+	✘	✔	✘	72.25
DIST+	✘	✘	✔	72.19
DIST+	✘	✔	✔	72.30
DIST+	✔	✔	✔	**72.39**

Table 13
Ablation of the loss in teacher acclimation. The student and teacher models are ResNet-18 and ResNet-34, respectively

Acclimation loss	Teacher ACC (%)		Student ACC (%)
	before	after	
w/o acclimation	73.31	73.31	72.30
KL divergence	73.31	71.03	70.67
Pearson distance	73.31	72.89	72.18
$\mathcal{L}_{\text{ta}}$	73.31	73.23	**72.39**

conduct experiments to validate the effects of each loss independently. As shown in Table 12, all the losses contribute to performance increments compared to the original DIST. Moreover, by combining them together, our resulting DIST+ performs the best.

Ablation study on teacher acclimation loss. This paper proposes a novel teacher acclimation loss $\mathcal{L}_{\text{ta}}$ that aims to optimize the teacher's non-target class predictions by aligning their correlations with the student's non-target class predictions, as we find the essential discrepancy between the teacher prediction and student prediction is mostly come from the non-target classes. Here we conduct experiments to show the efficacy of $\mathcal{L}_{\text{ta}}$ in comparisons with KL divergence (element-wisely reconstruct predictions in all classes) and Pearson distance (optimize the correlations in all classes).

As the results summarized in Table 13, adapting the teacher with KL divergence causes significant performance drops on both teacher accuracy and student accuracy, showing that the hard restriction of KL divergence will destroy the intrinsic preferences in teacher predictions, which are vital for achieving high accuracy. On the other hand, directly aligning the whole predictions has slight influences on teacher accuracy and student accuracy, as the student model has lower accuracy, and forcing the teacher predictions on target class to match the student's can decrease the teacher accuracy. In contrast, our $\mathcal{L}_{\text{ac}}$, which aligns the non-target classes only, can reduce the teacher-student discrepancy while retaining a similar accuracy. As a result, acclimation with our $\mathcal{L}_{\text{ac}}$ obtains the best distillation performance.

Ablation study on loss weights. We elaborate experiments to tune the hyperparameters of loss weights. For simplicity, we keep the weights of logits distillation loss (λ_2 and λ_3) the same, and so do λ_4 and λ_5 for feature distillation loss. As shown in Table 14, for DIST with logits distillation only, setting $\lambda_2 = \lambda_3 = 2$ achieves a better balance between task loss and distillation loss. While for

Table 14
Effects of different loss weights in DIST+. The student and teacher models are ResNet-18 and ResNet-34, respectively

Method	Task	Logits		Feature		ACC (%)
	λ_1	λ_2	λ_3	λ_4	λ_5	
DIST	1	1	1	0	0	71.85
DIST	1	2	2	0	0	72.07
DIST	1	4	4	0	0	72.02
DIST+	1	2	2	0.5	0.5	72.21
DIST+	1	2	2	1	1	**72.39**
DIST+	1	2	2	2	2	72.31
DIST+	1	2	2	4	4	72.27

Table 15
Average training speed (batches / second) of training ResNet-18 student with ResNet-34 teacher on ImageNet using strategy B1. The speed is tested based on our implementations on 8 NVIDIA V100 GPUs

KD [12]	RKD [45]	SRRL [41]	CRD [15]	DIST	DIST+
14.28	11.11	12.98	8.33	14.19	13.42

the DIST+ combining both logits and feature distillations, smaller feature distillation loss weights $\lambda_4 = \lambda_5 = 1$ are more beneficial to the performance.

6.6 Comparisons of training speed

We compare the training speed of our DIST with vanilla KD [12], RKD [45], CRD [15], and SRRL [41], as summarized in Table 15. Our DIST has almost the same highest training speed as the vanilla KD, outperforming other feature-based KD methods. While the DIST+ can achieve better performance with a marginal decrease of training efficiency.

7 Conclusion

This paper presents a new knowledge distillation (KD) method named DIST to implement better distillation from a stronger teacher. We empirically study the catastrophic discrepancy problem between the student and a stronger teacher, and propose a relation-based loss to relax the exact match of KL divergence in a linear sense. To further enhance the capability of distillation on stronger teachers, we extend DIST to a new variant DIST+, which introduces a novel teacher acclimation mechanism to narrow the teacher-student gap and a feature-level relaxed distillation loss. Our method DIST and DIST+ is simple yet effective in handling strong teachers. Extensive experiments show our superiority in various benchmark tasks.

Acknowledgments

This work was supported in part by the Australian Research Council under Projects DP210101859 and FT230100549.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[3] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[4] S. You, T. Huang, M. Yang, F. Wang, C. Qian, and C. Zhang, "Greedynas: Towards fast one-shot nas with greedy supernet," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1999–2008.

[5] T. Huang, S. You, B. Zhang, Y. Du, F. Wang, C. Qian, and C. Xu, "Dyrep: bootstrapping training with dynamic re-parameterization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 588–597.

[6] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[7] Z. Cai and N. Vasconcelos, "Cascade r-cnn: High quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–1, 2019. [Online]. Available: http://dx.doi.org/10.1109/tpami.2019.2956516

[8] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[12] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[13] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1921–1930.

[14] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5007–5016.

[15] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *International Conference on Learning Representations*, 2019.

[16] S. Du, S. You, X. Li, J. Wu, F. Wang, C. Qian, and C. Zhang, "Agree to disagree: Adaptive ensemble knowledge distillation in gradient space," *advances in neural information processing systems*, vol. 33, pp. 12 345–12 355, 2020.

[17] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4794–4802.

[18] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5191–5198.

[19] W. Son, J. Na, J. Choi, and W. Hwang, "Densely guided knowledge distillation using multiple teacher assistants," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9395–9404.

[20] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[21] K. Pearson, "Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia," *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, no. 187, pp. 253–318, 1896.

[22] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge distillation from a stronger teacher," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 716–33 727, 2022.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[24] D. Y. Park, M.-H. Cha, D. Kim, B. Han *et al.*, "Learning student-friendly teacher networks for knowledge distillation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 292–13 303, 2021.

[25] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," *Advances in neural information processing systems*, vol. 30, 2017.

[26] Q. Li, S. Jin, and J. Yan, "Mimicking very efficient network for object detection," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2017, pp. 6356–6364.

[27] J. Guo, K. Han, Y. Wang, H. Wu, X. Chen, C. Xu, and C. Xu, "Distilling object detectors via decoupled features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2154–2164.

[28] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5311–5320.

[29] Z. Yang, Z. Li, X. Jiang, Y. Gong, Z. Yuan, D. Zhao, and C. Yuan, "Focal and global knowledge distillation for detectors," *arXiv preprint arXiv:2111.11837*, 2021.

[30] T. Huang, Y. Zhang, S. You, F. Wang, C. Qian, J. Cao, and C. Xu, "Masked distillation with receptive tokens," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=mWRngkvIki3

[31] Y. Zhang, W. Chen, Y. Lu, T. Huang, X. Sun, and J. Cao, "Avatar knowledge distillation: Self-ensemble teacher paradigm with uncertainty," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 5272–5280. [Online]. Available: https://doi.org/10.1145/3581783.3611788

[32] Y. Zhang, T. Huang, J. Liu, T. Jiang, K. Cheng, and S. Zhang, "Freekd: Knowledge distillation via semantic frequency prompt," 2023.

[33] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 578–587.

[34] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2604–2613.

[35] Y. Wang, W. Zhou, T. Jiang, X. Bai, and Y. Xu, "Intra-class feature variation distillation for semantic segmentation," in *European Conference on Computer Vision*. Springer, 2020, pp. 346–362.

[36] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross-image relational knowledge distillation for semantic segmentation," *arXiv preprint arXiv:2204.06986*, 2022.

[37] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 702–703.

[38] R. Wightman, H. Touvron, and H. Jégou, "Resnet strikes back: An improved training procedure in timm," *arXiv preprint arXiv:2110.00476*, 2021.

[39] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 11 953–11 962.

[40] S. Marcel and Y. Rodriguez, "Torchvision the machine-vision package of torch," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1485–1488.

[41] J. Yang, B. Martinez, A. Bulat, and G. Tzimiropoulos, "Knowledge distillation via softmax regression representation learning," in *International Conference on Learning Representations*, 2020.

[42] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5008–5017.

[43] T. Huang, Y. Zhang, M. Zheng, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge diffusion for distillation," *arXiv preprint arXiv:2305.15712*, 2023.

[44] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[45] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976.

[46] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[47] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9163–9171.

[48] N. Passalis, M. Tzelepi, and A. Tefas, "Probabilistic knowledge transfer for lightweight deep representation learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2030–2039, 2020.

[49] J. Kim, S. Park, and N. Kwak, "Paraphrasing complex network: Network compression via factor transfer," *arXiv preprint arXiv:1802.04977*, 2018.

[50] L. Zhang and K. Ma, "Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors," in *International Conference on Learning Representations*, 2020.

[51] Y. Liu, C. Shu, J. Wang, and C. Shen, "Structured knowledge distillation for dense prediction," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[53] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[54] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[55] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

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

[57] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[58] Z. Shen, Z. Liu, D. Xu, Z. Chen, K.-T. Cheng, and M. Savvides, "Is label smoothing truly incompatible with knowledge distillation: An empirical study," in *International Conference on Learning Representations*, 2020.

[59] K. Chandrasegaran, N.-T. Tran, Y. ZHAO, and N. man Cheung, "To smooth or not to smooth? on compatibility between label smoothing and knowledge distillation," 2022. [Online]. Available: https://openreview.net/forum?id=Vvmj4zGU_z3

Tao Huang received the BE degree in computer science and technology from Huazhong University of Science and Technology, in 2020. He is currently working toward the PhD degree in computer science with the University of Sydney. His research interests include fundamental algorithms for machine learning and computer vision, such as AutoML, representation learning, model compression and knowledge distillation. He has published his research outcomes in many prestigious journals and top tier conferences.

Shan You (Member, IEEE) received the bachelor of mathematics and applied mathematics (elite class) degree from Xi'an Jiaotong University, and the PhD degree of computer science from Peking University. He is currently a senior researcher with SenseTime, and also a post doc with Tsinghua University. His research interests include fundamental algorithms for machine learning and computer vision, such as AutoML, representation learning, light detector and face analysis. He has published his research outcomes in many top tier conferences and transactions.

Fei Wang received the bachelor's and master's degrees from the Beijing University of Posts and Telecommunications. Currently, he is working toward the PhD degree with the University of Science and Technology of China. He is the director of SenseTime Intelligent Automotive Group. He is the head of SenseAuto-Parking engineering and SenseAuto-Cabin research. He leads a vibrant team of more than 60 people to develop comprehensive solutions for the intelligent vehicle and deliver more than 20 mass production of SenseAuto-Cabin projects in the last 6 years. He has published more than 20 papers with CVPR/NIPS/ICCV during the last few years. His research interests include automotive drive system, AI chip, deep learning, etc.

Chen Qian received the BEng degree from the Institute for Interdisciplinary Information Science, Tsinghua University, in 2012, and the MPhil degree from the Department of Information Engineering, The Chinese University of Hong Kong, in 2014. He is currently working with SenseTime as research director. His research interests include human-related computer vision and machine learning problems.

Chang Xu received the PhD degree from Peking University, China. He is ARC Future Fellow and Associate Professor at the School of Computer Science, University of Sydney. He received the University of Sydney Vice-Chancellor's Award for Outstanding Early Career Research. His research interests lie in machine learning algorithms and applications in computer vision. He has published over 100 papers in prestigious journals and top tier conferences. He has received several paper awards, including Distinguished Paper Award in AAAI 2023 and Distinguished Paper Award in IJCAI 2018. He served as an area chair of NeurIPS, ICML, ICLR, KDD, CVPR, and MM, as well as a Senior PC member of AAAI and IJCAI. In addition, he served as an associate editor at IEEE T-PAMI, IEEE T-MM, and T-MLR. He has been named a Top Ten Distinguished Senior PC Member in IJCAI 2017 and an Outstanding Associate Editor at IEEE T-MM in 2022.