# Structured Knowledge Distillation for Accurate and Efficient Object Detection

Linfeng Zhang and Kaisheng Ma

***Abstract*—Knowledge distillation, which aims to transfer the knowledge learned by a cumbersome teacher model to a lightweight student model, has become one of the most popular and effective techniques in computer vision. However, many previous knowledge distillation methods are designed for image classification and fail in more challenging tasks such as object detection. In this paper, we first suggest that the failure of knowledge distillation on object detection is mainly caused by two reasons: (1) the imbalance between pixels of foreground and background and (2) lack of knowledge distillation on the relation among different pixels. Then, we propose a structured knowledge distillation scheme, including *attention-guided distillation* and *non-local distillation* to address the two issues, respectively. Attention-guided distillation is proposed to find the crucial pixels of foreground objects with an attention mechanism and then make the students take more effort to learn their features. Non-local distillation is proposed to enable students to learn not only the feature of an individual pixel but also the relation between different pixels captured by non-local modules. Experimental results have demonstrated the effectiveness of our method on thirteen kinds of object detection models with twelve comparison methods for both object detection and instance segmentation. For instance, Faster RCNN with our distillation achieves 43.9 mAP on MS COCO2017, which is 4.1 higher than the baseline. Additionally, we show that our method is also beneficial to the robustness and domain generalization ability of detectors. Codes and model weights have been released on GitHub[1].**

***Index Terms*—Attention, instance segmentation, knowledge distillation, model acceleration and compression, non-local module, object detection, student-teacher learning.**

## I. Introduction

DEEP learning has witnessed significant advancements in various computer vision tasks [1], [2], [3], [4]. However, the high computational and memory requirements of state-of-the-art deep neural networks have hindered their deployment in resource-limited edge devices like self-driving cars and mobile phones. To address this problem, abundant techniques have been proposed, including pruning [5], [6], [7], [8], quantization [9], [10], compact model design [11], [12], [13], [14] and knowledge distillation [15], [16]. Knowledge distillation, also referred to as student-teacher learning, aims to transfer the knowledge from a cumbersome teacher model to a lightweight student model. By mimicking the prediction results and features of the teacher, the student can inherit the dark knowledge from the teacher and thus often achieves much higher accuracy. Due to its simplicity and effectiveness, knowledge distillation has emerged as a popular technique for both model compression and accuracy boosting.[1]

As one of the most crucial challenges in computer vision, object detection demands models that are not only accurate but also efficient. Regrettably, many existing knowledge distillation methods in computer vision are primarily designed for image classification and often yield insignificant or even negative effects when applied to object detection [17]. In this paper, we attribute the unsatisfactory performance of knowledge distillation on object detection to the following two issues: (1) imbalance between foreground and background, and (2) lack of knowledge distillation on the relation among different pixels.

*Imbalance Between Foreground and Background.* The number of background pixels in images often greatly exceeds the number of pixels associated with foreground objects. However, only the pixels belonging to foreground objects are truly informative for object detection. In conventional knowledge distillation methods, the student model is typically trained to mimic the features of all pixels equally. Consequently, the students allocate a significant portion of their attention to learning the teacher's knowledge from the background pixels, which hampers their ability to learn the distinctive features of the foreground objects. As a result, this imbalance severely diminishes the effectiveness of knowledge distillation.

To address this issue, we propose *attention-guided distillation*, which selectively distills knowledge from the essential foreground pixels. Previous studies have demonstrated that the attention value of a pixel reflects its significance in the image [18], [19]. Building upon this insight, our attention-guided distillation employs the attention map as a metric to determine whether a pixel belongs to a foreground object. Consequently, knowledge distillation is exclusively applied to these foreground objects, rather than considering all the pixels in the image. This approach allows the student model to focus its learning efforts on the most relevant foreground features, effectively addressing the imbalance issue.

Manuscript received 15 January 2022; revised 30 June 2023; accepted 24 July 2023. Date of publication 1 August 2023; date of current version 3 November 2023. This work was supported in part by the IIISCT (Institute for interdisciplinary Information Core Technology), in part by the National Natural Sciences Foundation of China under Grants 31970972 and 11901338, and in part by the Tsinghua University Initiative Scientific Research Program. Recommended for acceptance by A. van den Hengel. *(Corresponding author: Kaisheng Ma.)*

The authors are with the Institute of Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China (e-mail: zhang-lf19@mails.tsinghua.edu.cn; merrydoudou@gmail.com).

Digital Object Identifier 10.1109/TPAMI.2023.3300470

[1†] https://github.com/ArchipLab-LinfengZhang/Object-Detection-Knowledge-Distillation

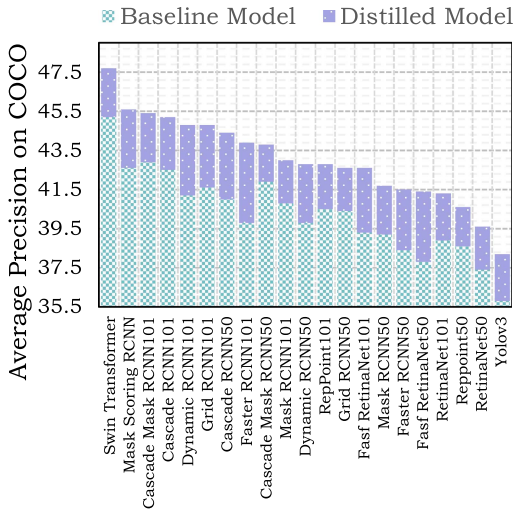



Fig. 1. Main results (mAP) of our method on MS COCO2017. Numbers after detector names indicate the depth of backbone networks.

*Lack of Distillation Among Relation Information.* It is widely recognized that the relation information among different objects holds significant value in object detection. Recent advancements, such as non-local modules [20] and relation networks [21], have demonstrated successful improvements in detector performance by facilitating the capture and utilization of these object relations. Despite these advancements, existing knowledge distillation methods for object detection primarily focus on distilling information from individual pixels while neglecting the crucial aspect of learning inter-pixel relation.

To address this issue, we propose non-local distillation, which aims to capture the relation information of students and teachers with non-local modules and subsequently distill this from teachers to students. Since the non-local modules and attention mechanisms in our method are only required during training, they can be discarded during inference to avoid additional computation and storage costs.

Since the proposed methods are feature-based distillation methods which do not depend on a specific detection model, they can be directly utilized in all kinds of detectors without any modification. Given that the features of the teacher detector contain richer semantic information compared to its prediction results, our method demonstrates significantly better effectiveness than prediction-based knowledge distillation. Furthermore, as detailed in Section V-B5, by using a two-stage knowledge distillation pipeline, our method is able to fully exploit the knowledge learned by a powerful teacher and achieves better student performance than one-stage knowledge distillation.

As shown in Fig. 1, extensive experiments validate the remarkable performance of our method in the domains of object detection and instance segmentation. On MS COCO2017, 3.1, 2.6, and 2.1 AP improvements can be observed on two-stage, one-stage, and anchor-free object detectors on average, respectively. On Mask RCNN, our method improves the performance of object detection and instance segmentation by 2.5 and 2.0 AP on average, respectively. Moreover, we also show that knowledge distillation is beneficial to the robustness and domain generalization abilities of object detectors.

In the discussion section, we have conducted a comprehensive ablation study and sensitivity analysis to demonstrate the effectiveness and stability of each knowledge distillation loss employed in our approach. Furthermore, we have investigated the relation between teachers and students in the context of object detection. Our findings indicate that knowledge distillation in object detection necessitates a teacher model with a high Average Precision (AP), which differs from the conclusions drawn in the field of image classification, where a high-accuracy teacher may harm student performance [22], [23]. These results highlight the need for further exploration of knowledge distillation in tasks beyond image classification. To sum up, the contribution of this paper can be summarized as follows.

- We propose attention-guided distillation, which emphasizes student learning on the foreground objects and suppresses student learning on the background pixels.
- We propose non-local distillation, which enables the students to learn not only the information of the individual pixel but also the relation among different pixels from teachers.
- We evaluate the proposed method against twelve comparison methods using thirteen models for both object detection and instance segmentation. Qualitative analysis and error distribution analysis have demonstrated the benefits of knowledge distillation on both localization and classification.
- We show that in knowledge distillation for object detection, a teacher with a higher AP tends to be more effective, which differs from the previous conclusion in the field of image classification.

## II. Related Work

### A. Knowledge Distillation

Knowledge distillation, initially proposed for compressing ensemble models, has become one of the most popular and effective techniques in neural network training [15], [16]. The concept of knowledge distillation was first introduced by Hinton et al. where students are trained to mimic the softmax outputs of teachers. Since then, numerous methods have been proposed to transfer knowledge from teachers to students, focusing on various aspects such as teacher features [24], attention mechanisms [18], [25], FSP (Flow of Solution Procedure) [26], mutual information [27], positive features [28], task-oriented features [29], relations [30], [31], [32], [33], self-supervised learning knowledge [34], and more. Following its success in image classification, researchers have applied knowledge distillation to various domains and tasks, including object detection [17], [35], [36], [37], semantic segmentation [38], face recognition [39], few-shot learning [40], [41], [42], incremental learning [40], [43], [44], distributed learning [45], scene parsing [46], data augmentation [47], pre-trained language models [48], [49], video recognition [50], [51], [52], image-to-image translation [53], [54], [55], [56], [57], [58], multi-exit networks training [59], [60], multi-modal learning [61], [62], [63], model robustness [64], and federated learning [65].

Recently, there has been a growing interest in investigating the relation between student and teacher performance in knowledge distillation. Mirzadeh et al. [22] found that the teacher with the highest accuracy may not necessarily be the most suitable teacher for knowledge distillation, as a significant accuracy gap between teachers and students can impede student training. Cho et al. [23] discovered that teachers trained with early stopping tend to be more effective in knowledge distillation. Furthermore, Müller et al. [66] demonstrated that label smoothing may have a negative impact on the efficiency of knowledge distillation. Additionally, neural network search methods have been proposed to automatically identify the optimal teacher-student pairing [67], [68]. However, it is important to note that all of the aforementioned research on the student-teacher relation is primarily focused on image classification tasks. The generalizability of these findings and experimental results to more challenging visual tasks, such as object detection, remains largely unknown.

Knowledge distillation has gained significant attention in the context of object detection, aiming to improve the performance of object detectors. Chen et al. introduced the first knowledge distillation method specifically for object detection, which involved distillation losses on the backbone feature, classification head, and regression head [35]. Wang et al. utilized GANs to distill the backbone feature of teacher detectors [69]. Hao et al. and Chen et al. applied knowledge distillation to incremental learning in object detection [43], [44]. Additionally, several studies have focused on distilling teacher knowledge to improve the localization ability of object detectors, resulting in notable performance improvements [70], [71].

Recently, many researchers have found that the imbalance between foreground objects and background is a crucial problem in detection distillation. Dai et al. introduced instance knowledge distillation, which distills feature-based, relation-based, and response-based information in object detection [72]. Li et al. proposed a method where only the features sampled by the region proposal network are subject to $L_2$ distillation loss [17]. Bajestani and Yang presented temporal knowledge distillation for video object detection, introducing a hyperparameter to balance the distillation loss between foreground and background pixels [37]. Wang et al. proposed fine-grained feature imitation, which distills features specifically near object anchor locations [36]. Guo et al. used gradients to identify foreground object pixels, while Du et al. localized the pixels to be distilled based on a feature richness score from the classification head [73]. However, many of these approaches rely on annotations in ground truth, anchors, or bounding boxes, making them less transferable across different detectors.

In contrast, our attention-guided distillation addresses this challenge by adaptively identifying foreground object pixels using a parameter-free attention mechanism. This attention map can be easily generated from features with minimal computational cost. As a result, our approach can be directly applied to various detectors and tasks without the need for modification.

A comparison between the previous object detection knowledge distillation method [36] and our attention-guided distillation method is illustrated in Fig. 3. We highlight the advantages of our method in the following four aspects: (i) Our attention-guided distillation method utilizes a parameter-free attention mechanism to identify foreground object pixels without relying on ground truth annotations, bounding boxes, anchor priors, or gradient propagation. This makes our method easily transferable to different types of detectors. (ii) Unlike previous methods that rely on bounding boxes, our method assigns attention scores to individual pixels, allowing it to be applied to objects of arbitrary shapes. (iii) While previous methods only determine whether a pixel should be distilled or not, our method assigns each pixel a learning priority ranging from 0 to 1, providing more informative guidance for the distillation process. (iv) In addition to identifying crucial pixels in the image, our method also identifies crucial channels. Our ablation study demonstrates that the inclusion of channel masks significantly enhances the performance of knowledge distillation, enabling the identification of both critical pixels and important channels in the image. These advantages highlight the effectiveness and versatility of our attention-guided distillation method in object detection knowledge transfer.

### B. *Other Model Compression Techniques*

Besides knowledge distillation, there are various other model compression techniques, including neural network pruning, quantization, compact model design, adaptive inference, neural architecture search, and more. Neural network pruning aims to iteratively remove unimportant neurons [5], filters [74], channels [75], or layers [76] from an over-parameterized model. These unimportant units can be identified using methods such as L1-norm [74], geometric median [77], meta networks [78], reinforcement learning methods [79], and others. Neural network quantization involves quantizing the weights [5], [9], [10] and activations [80] of neural networks to low-bit or even binary representations [81], [82]. Adaptive inference allows neural networks to selectively skip the computation of certain layers [59], [83], [84], [85], channels [86], [87], and regions [88], [89], [90], [91], [92] based on the input images. Furthermore, there have been numerous advancements in the development of compact and efficient model architectures, such as the MobileNet family [11], [12], [93], ShuffleNet [13], [94], EfficientNet [95], [96], and others.

These model compression techniques have also demonstrated significant improvements in object detection. Li et al. investigated effective quantization schemes with batch normalization freezing and activation calibration [97]. Xie et al. proposed a localization-aware pruning scheme to preserve object localization knowledge in detectors [98]. In addition to compressing pre-trained models, there have been notable efforts in training lightweight detectors directly. The YOLO family of models achieves efficient detection in a one-stage manner [99], [100], [101], [102]. ThunderNet accelerates two-stage detection using low-resolution images, lightweight detection heads, and attention modules [103]. EfficientDet replaces the backbone with EfficientNet and incorporates a Bidirectional FPN layer [104]. Chen et al. propose YOLOF to reduce computation overhead in

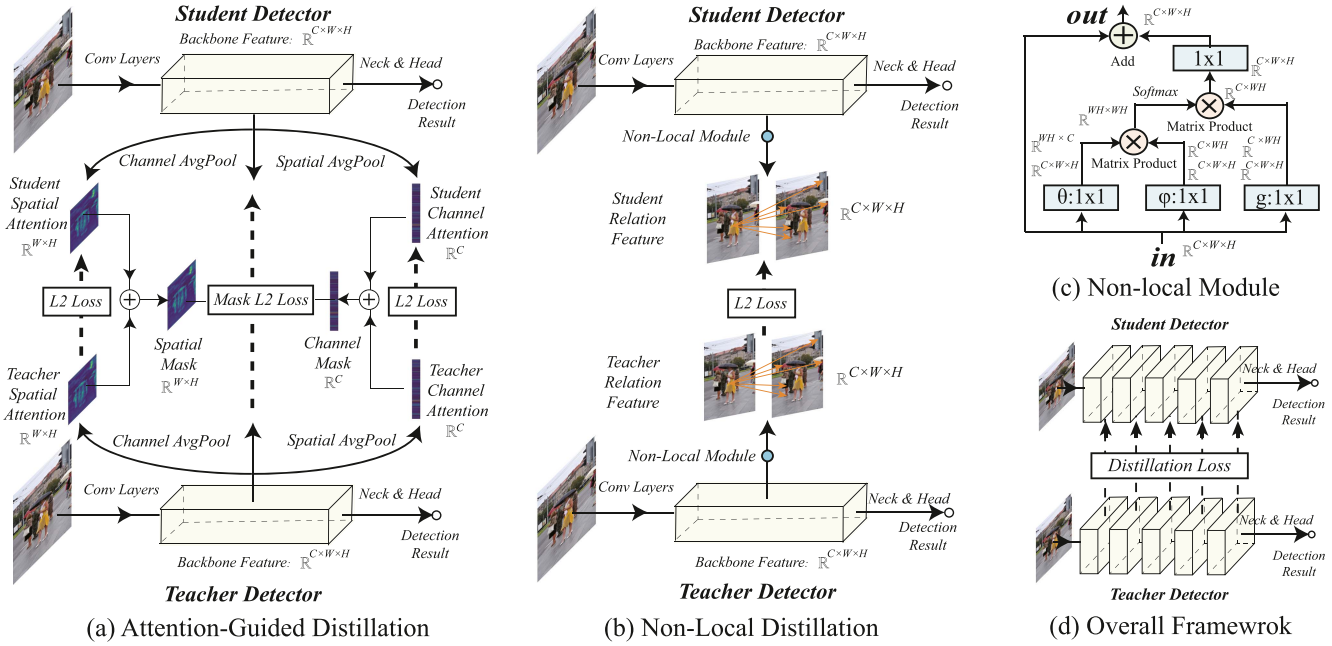



Fig. 2. Details of our method: (a) Attention-guided distillation generates the spatial and channel attention with average pooling in the channel dimension and the spatial dimension, respectively. Then, students are trained to learn teacher features in the pixels and channels with high attention values. Besides, students are encouraged to mimic the attention of teachers. (b) Non-local distillation captures the relation of pixels in an image with non-local modules. Then, the relation information of teachers is transferred to the students by minimizing $L_2$ norm loss. (c) The architecture of non-local modules. '1x1' is a convolution layer with a 1x1 kernel. (d) Distillation loss is applied to backbone features in different layers with various resolutions. The detection head and neck are not involved in our method.

FPN layers through a dilated encoder and a uniform matching policy [105].

### C. Non-Local Module

Non-local mean is a classical image denoising algorithm that computes a weighted mean of all pixels in an image [106]. Building upon its success in capturing long-range dependencies, Wang et al. introduced non-local neural networks. These networks utilize non-local modules to capture relational information between pixels and frames, enhancing the representation ability of features [106]. Since then, several works have successfully applied non-local modules in various domains, including point cloud processing [107], image denoising [108], [109], [110], image super-resolution [111], and video processing [112], [113]. Despite their success, non-local modules introduce noticeable computational and storage overhead. To mitigate this issue, researchers have proposed numerous lightweight non-local modules, such as CCNet [114], GCNet [115], and RCCA [116]. Furthermore, Liu et al. recently applied neural architecture search to discover optimal lightweight non-local modules [117].

## III. Methodology

### A. Overall Illustration

The details of our method are illustrated in Fig. 2. Our approach consists of two distinct distillation methods: attention-guided distillation and non-local distillation. In attention-guided distillation, we initially generate spatial attention and channel attention maps for both teachers and students. This is achieved by applying average pooling on the absolute values of the features in the channel and spatial dimensions, respectively. Subsequently, we normalize the spatial and channel attention maps of teachers and students using a temperature-parameterized softmax function. Next, we combine the normalized attention maps of teachers and students by adding them together and dividing the result by 2. This operation yields the masks for attention-guided distillation. It is important to note that each element within the masks ranges from 0 to 1, indicating the relative importance of different pixels and channels. When calculating the feature distillation loss, we leverage the spatial and channel masks to reweight the loss in different pixels and channels. Consequently,


Fig. 3. Comparison between the proposed attention-guided distillation and other methods. Our method can find the crucial pixels of objects in arbitrary shapes. It not only tells whether a pixel should be distilled but also shows the numerical learning priority of this pixel.

the knowledge distillation loss emphasizes the crucial pixels and channels while suppressing others.

In non-local distillation, we incorporate additional non-local modules to capture the relational information in the backbone features of both teachers and students. Throughout the training process, the student network learns to extract and leverage the relational information from the teachers. As depicted in Fig. 6, it is worth noting that our approach differs from previous non-local neural networks. In previous approaches, non-local modules were used to enhance the backbone features. However, in our method, the non-local modules are solely employed for knowledge distillation. Consequently, during the inference phase, these non-local modules can be discarded to avoid the additional computational and storage costs.

### B. Why Students and Teachers in Our Method Can Have Different Architectures

Our method allows the students to have non-identical architectures to their teachers for several reasons. Firstly, in many of our experiments, both the student and the teacher detectors utilize the Faster RCNN-style detection paradigm, which involves the extraction of image features through backbone networks, object proposal computation with region proposal networks (RPNs), and object localization and classification with regression and classification heads. Given the similarity in their detection pipelines, the features learned by the teacher detector are also similar to the features learned by the student detector, and hence knowledge distillation can be applied. Secondly, our method focuses on the image feature extraction stage of the backbone network. The majority of differences between detectors lie in other stages, such as proposal generation and label assignment, which do not directly impact our approach. Thirdly, previous knowledge distillation works [24], [118], [119], [120] have demonstrated that differences between student and teacher features in terms of channel dimension, width, and height can be harmonized using a linear feature reshaping layer (adaptation layer). This allows our approach to generalize well across different student-teacher configurations. However, as discussed in Section V-B2, our method may not be as effective when the student and teacher detectors employ entirely different detection pipelines (e.g., RetinaNet students versus Faster RCNN teachers). The disparities in their detection pipelines result in the extraction of different types of image features by their respective backbones. Training the student detector with features learned from the teacher detector in such cases could potentially mislead the student training process.

### C. Formulation

*1) Attention-Guided Distillation:* We use $A \in \mathbb{R}^{C,H,W}$ to denote the feature of the backbone in an object detection model, where $C, H, W$ denotes its channel number, height and width, respectively. Then, the generation of the spatial attention map and channel attention map is equivalent to finding the mapping function $\mathcal{G}^s : \mathbb{R}^{C,H,W} \rightarrow \mathbb{R}^{H,W}$ and $\mathcal{G}^c : \mathbb{R}^{C,H,W} \rightarrow \mathbb{R}^{C}$, respectively. Note that the superscripts $s$ and $c$ here are utilized to discriminate 'spatial' and 'channel'. Since the absolute value of each element in the feature implies its importance, we construct $\mathcal{G}^s$ by averaging the absolute values across the channel dimension and construct $\mathcal{G}^c$ by averaging the absolute values across the width and height dimension, which can be formulated as

$$\mathcal{G}^c(A) = \frac{1}{HW}\sum_{H}^{i=1}\sum_{W}^{j=1}|A_{\cdot,i,j}| \quad \text{and}$$
$$\mathcal{G}^s(A) = \frac{1}{C}\sum_{C}^{k=1}|A_{k,\cdot,\cdot}|, \tag{1}$$

where $i, j, k$ denotes the $i_{th}$, $j_{th}$, $k_{th}$ slice of $A$ in the height, width, and channel dimension, respectively. Then, the spatial attention mask $M^s$, and the channel attention mask $M^c$ used in attention-guided distillation can be obtained by summing the attention maps from the teacher and the student detector, which can be formulated as

$$M^s = HW \cdot \text{softmax}\left(\left(\mathcal{G}^s\left(A^{\mathcal{S}}\right) + \mathcal{G}^s\left(A^{\mathcal{T}}\right)\right)/T\right) \quad \text{and}$$
$$M^c = C \cdot \text{softmax}\left(\left(\mathcal{G}^c\left(A^{\mathcal{S}}\right) + \mathcal{G}^c\left(A^{\mathcal{T}}\right)\right)/T\right). \tag{2}$$

Note that the superscripts $\mathcal{S}$ and $\mathcal{T}$ here are used to discriminate students and teachers. $T$ is a hyper-parameter in softmax introduced by Hinton et al. [15] to adjust the distribution of elements in attention masks (see Fig. 4 and Fig. 5). The attention-guided distillation loss $\mathcal{L}_{\text{AGD}}$ is composed of two sub-modules – attention transfer loss $\mathcal{L}_{\text{AT}}$ and attention-masked loss $\mathcal{L}_{\text{AM}}$. $\mathcal{L}_{\text{AT}}$ is utilized to encourage the student model to mimic the spatial and channel attention of the teacher model, which can be formulated as

$$\mathcal{L}_{\text{AT}} = \mathcal{L}_2(\mathcal{G}^s(A^{\mathcal{S}}), \mathcal{G}^s(A^{\mathcal{T}})) + \mathcal{L}_2(\mathcal{G}^c(A^{\mathcal{S}}), \mathcal{G}^c(A^{\mathcal{T}})). \tag{3}$$

$\mathcal{L}_{\text{AM}}$ is utilized to encourage the student to mimic the features of teacher models by a $\mathcal{L}_2$ norm loss masked by $M^s$ and $M^c$, which can be formulated as

$$\mathcal{L}_{\text{AM}} = \left(\sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}(A^{\mathcal{T}}_{k,i,j} - A^{\mathcal{S}}_{k,i,j})^2 \cdot M^s_{i,j} \cdot M^c_k\right)^{\frac{1}{2}}. \tag{4}$$

*2) Non-Local Distillation:* Non-local module [20] is an effective method to improve the performance of neural networks by capturing the global relation information. In this paper, we apply non-local modules to capture the relation between pixels in an image, which can be formulated as

$$r_{i,j} = \frac{1}{WH}\sum_{H}^{i'=1}\sum_{W}^{j'=1} f(A_{\cdot,i,j}, A_{\cdot,i',j'})g(A_{\cdot,i',j'}), \tag{5}$$

where $r$ denotes the obtained relation information. $i, j$ are the spatial indexes of an output position whose response is to be computed. $i', j'$ are the spatial indexes that enumerate all possible positions in an image. $f$ is a pairwise function for computing the relation of two pixels and $g$ is an unary function for computing the representation of an individual pixel. Now, we can introduce the proposed non-local distillation loss $\mathcal{L}_{\text{NLD}}$ as the $\mathcal{L}_2$ loss between the relation information of the students and teachers, which can be formulated as $\mathcal{L}_{\text{NLD}} = \mathcal{L}_2(r^{\mathcal{S}}, r^{\mathcal{T}})$.


Fig. 4. Visualization and distribution of the spatial attention with different $T$ (temperatures) in attention-guided knowledge distillation. With a smaller $T$, the pixels of high and low attention values are emphasized and suppressed more in knowledge distillation, respectively.

Fig. 5. Visualization of spatial attention on more images with $T = 1$.

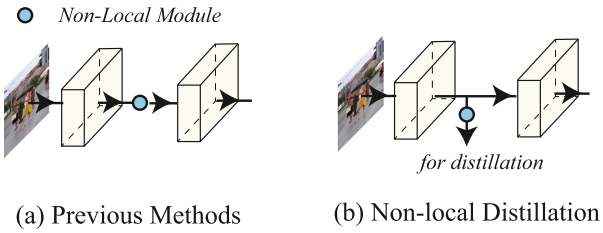



Fig. 6. Comparison between the usage of non-local modules in previous methods and the proposed non-local distillation. The non-local module in our method is not involved in the forward computation. Thus, it can be discarded during inference.

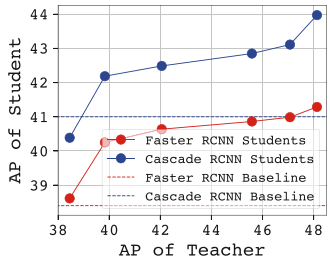



Fig. 7. Relation between the mean average precision of students and teachers on Faster RCNN and Cascade RCNN.

*3) Overall Loss Function:* We introduce three hyper-parameters $\alpha, \beta, \gamma$ to balance different distillation loss functions in our method. The overall distillation loss can be formulated as

$$\mathcal{L}_{\text{Distill}}(A^{\mathcal{T}}, A^{\mathcal{S}}) = \underbrace{\alpha \cdot \mathcal{L}_{\text{AT}} + \beta \cdot \mathcal{L}_{\text{AM}}}_{\text{Attention-guided distillation}} + \underbrace{\gamma \cdot \mathcal{L}_{\text{NLD}}}_{\text{Non-local distillation}}. \quad (6)$$

The sensitivity study of each hyper-parameter and the ablation study on each loss are shown in Fig. 11 and Table VII, respectively. The overall distillation loss is a model-agnostic loss, which can be added to the original training loss of any detection model directly. Hence, by denoting the original training loss of the detector (e.g., classification loss and regression loss) as $\mathcal{L}_{\text{Origin}}$, the overall training loss of a student detector $\mathcal{L}_{\text{Stdent}}$ can be written as

$$\mathcal{L}_{\text{Stdent}} = \mathcal{L}_{\text{Origin}} + \mathcal{L}_{\text{Distill}}. \quad (7)$$

Taking Faster RCNN as an example, $\mathcal{L}_{\text{Origin}}$ can be formulated as

$$\mathcal{L}_{\text{Origin}}(\{p_i\}, \{t_i\}) = \frac{1}{N_{\text{cls}}} \sum_i \mathcal{L}_{\text{cls}}(p_i, p_i^*) + \lambda \frac{1}{N_{\text{reg}}} \sum_i p_i^* \mathcal{L}_{\text{reg}}(t_i, t_i^*), \quad (8)$$

where $i$ is the index of an anchor in a mini-batch and $p_i$ is the predicted probability of anchor $i$ being an object. The ground-truth label $p_i^* = 1$ when the anchor is positive, and $p_i^* = 0$ when the anchor is negative. $t_i$ is a vector representing the 4 parameterized coordinates of the predicated bounding box. $t_i^*$ is that of the

ground-truth box assigned with a positive anchor. $\mathcal{L}_{\text{cls}}$ is the log loss for binary classification (object v.s. not object). $\mathcal{L}_{\text{reg}}$ indicates the smooth L1 loss for regression. $N_{\text{cls}}$ and $N_{\text{reg}}$ are the number of samples in a mini-batch and the number of possible anchor localizations, respectively. They are utilized to normalize the classification and regression loss. Then, the overall loss function of Faster RCNN students trained with our methods can be formulated as

$$\begin{aligned}\mathcal{L}_{\text{Stdent}} &= \mathcal{L}_{\text{Origin}} + \mathcal{L}_{\text{Distill}} \\ &= \frac{1}{N_{\text{cls}}}\sum_i \mathcal{L}_{\text{cls}}(p_i, p_i^*) + \lambda \frac{1}{N_{\text{reg}}}\sum_i p_i^* \mathcal{L}_{\text{reg}}(t_i, t_i^*) \\ &\quad + \alpha \cdot \mathcal{L}_{\text{AT}} + \beta \cdot \mathcal{L}_{\text{AM}} + \gamma \cdot \mathcal{L}_{\text{NLD}}. \end{aligned} \tag{9}$$

## IV. Experiment

### A. Experiment Settings

The proposed knowledge distillation method has primarily been evaluated on the MS COCO2017 dataset, which is a large-scale dataset containing over 120 k images spanning 80 categories [121]. Following the common practice [122], [123], we train the detectors on the COCO `train` split (approximately 118 K images) and evaluate their performance on the `validation` split (5 k images) using official performance metrics [121], including AP (average precision), AP50, AP75, $\text{AP}_S$, $\text{AP}_M$, and $\text{AP}_L$. In addition to MS COCO, we also evaluate our method on the Cityscapes dataset [124], which focuses on semantic urban scene understanding, and the COCO-C dataset [125], designed to benchmark model robustness and domain generalization ability using corrupted images. Our experiments are conducted with various two-stage detectors, including Faster RCNN [2], Cascade RCNN [126], Dynamic RCNN [127], and Grid RCNN [128], as well as one-stage detectors such as Yolov3 [101], SSD [129], RetinaNet [123], and Fsaf RetinaNet [130]. Additionally, we evaluate our method on the RepPoint [131] anchor-free detector. Furthermore, we evaluate our method on instance segmentation instance segmentation models, including Mask RCNN [122], Cascade Mask RCNN [126], Mask Scoring RCNN [132], and Swin Transformer [133]. For comparison, we consider twelve existing knowledge distillation methods [17], [24], [28], [35], [36], [71], [72], [73], [134], [134], [135], [136]. Mean average precision (mAP) serves as the performance metric for all the conducted experiments. We initialize the backbone network of each detector using ImageNet pre-trained weights of VGG, DarkNet, ResNet18, RegNet800 M, ResNet50, ResNet101, ResNeXt101, Vision Transformer, and Swin Transformer [99], [133], [137], [138], [139], [140], [141], [142]. The models are trained for 24 epochs unless otherwise specified, with learning rate decayed at the 18th and 22nd epochs. Data augmentation techniques, including image resizing and random flipping, are applied in all experiments. For two-stage models, we use the hyper-parameter settings $\alpha = \gamma = 7 \times 10^{-5}$, $\beta = 4 \times 10^{-3}$, and $T = 0.1$. For one-stage models, including Cascade Mask RCNN, RetinaNet, and Reppoint with ResNeXt101 backbones, we use $\alpha = \gamma = 4 \times 10^{-4}$, $\beta = 2 \times 10^{-2}$, and $T = 0.5$. We adopt Cascade Mask RCNN, RetinaNet, and Reppoint with ResNeXt101 backbones as the teacher models for two-stage, one-stage, and Reppoint students, respectively. All experiments are implemented using PyTorch [143] with the mmdetection2 framework [144]. The reported frames per second (FPS) is measured on an RTX 2080Ti GPU.

TABLE I
Performance of the Teacher Model Utilized in Our Experiments

| Teacher Model | Backbone | Params | FPS | Bounding box AP | | | | Mask AP | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | AP | $\text{AP}_S$ | $\text{AP}_M$ | $\text{AP}_L$ | AP | $\text{AP}_S$ | $\text{AP}_M$ | $\text{AP}_L$ |
| Cascade Mask RCNN | ResNetX101-DCN | 100.35 | 4.2 | 47.3 | 28.2 | 51.7 | 62.7 | 41.1 | 22.9 | 44.9 | 56.3 |
| Swin-S | Swin Transformer-S | 99.52 | 3.6 | 48.2 | 32.1 | 51.8 | 62.7 | 43.2 | 24.8 | 46.3 | 62.1 |
| RetinaNet | ResNext101 | 102.8 | 8.7 | 40.8 | 22.9 | 44.5 | 54.6 | – | – | – | – |
| Reppoints | ResNext101-DCN | 65.12 | 9.3 | 44.2 | 26.2 | 48.4 | 58.5 | – | – | – | – |
| Yolo v3 | DarkNet53-Channel × 2 | 236.2 | 14.1 | 36.0 | 21.2 | 39.6 | 45.4 | – | – | – | – |
| SSD | VGG16-Channel × 2 | 144.3 | 9.2 | 30.8 | 14.3 | 33.4 | 44.7 | – | – | – | – |
| Faster RCNN | ViT-H | 648.34 | 0.4 | 42.3 | 26.4 | 45.1 | 57.0 | – | – | – | – |

For two-stage detectors such as faster RCNN, cascade RCNN, dynamic RCNN, grid RCNN, mask RCNN, cascade mask RCNN, and mask scoring RCNN, the teacher model is cascade mask RCNN. Faster RCNN with VIT-H serves as the teacher model for faster RCNN with VIT-B students. The remaining teachers are employed to teach students with the same detection pipeline but fewer layers or channels.

### B. Architectures and Performance of Teachers

The performance of the teacher detectors used in the knowledge distillation experiments in this paper is presented in Table I. For the standard two-stage student detectors, including Faster RCNN, Cascade RCNN, Dynamic RCNN, Grid RCNN, Mask RCNN, Cascade Mask RCNN, and Mask Scoring RCNN, we utilize Cascade Mask RCNN with ResNeXt101-DCN backbone as their teachers. Faster RCNN with ViT-H serves as the teacher for Faster RCNN with VIT-B students. For the RetinaNet-like student detectors, namely RetinaNet and Fsaf RetinaNet, we employ RetinaNet with the ResNeXt101 backbone as their teacher. In the case of Reppoints, YoloV3, SSD, and Swin-T students, we use Reppoints with ResNeXt101-DCN backbone, YoloV3 with double convolution channels, SSD with double convolution channels, and Swin-S (Swin-T with double layers) as their respective teachers. Please refer to Table I for the detailed performance of the teacher detectors.

### C. Results on Detection and Instance Segmentation

In this subsection, we present the experimental results of detectors trained with and without our proposed method on MS COCO2017, as shown in Tables II and III. Additionally, a comparison is made between our method and twelve other knowledge distillation methods, as illustrated in Table V. Furthermore, an evaluation of our method on Cityscapes, as presented in Table VI, is also provided. The key observations from the results are as follows: (i) Consistent and significant improvements in average precision (AP) are observed across all nine types of detectors listed in Table II. On average, there are 3.1, 2.6, and 2.1 AP improvements for the two-stage, one-stage, and anchor-free detectors, respectively. (ii) The proposed method enables a student model with a ResNet50 backbone to outperform the same model with a ResNet101 backbone by an average of 1.2 AP. (iii) Notably, on instance segmentation models presented in Table III, there are average improvements of 2.5 AP for bounding box AP and 2.0 AP for mask AP, indicating the effectiveness of our method in both object detection and instance segmentation. (iv) Our method achieves notable AP improvements of 3.5 and 4.1 on Faster RCNN models with ViT-B-32 and ViT-B-16 backbones, respectively, demonstrating its effectiveness in

TABLE II
EXPERIMENTS ON MS COCO2017 WITH THE PROPOSED KNOWLEDGE DISTILLATION METHOD

| Model | Backbone | FPS | Params | Distill | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Faster RCNN | ResNet18 | 28.1 | 30.57 | × | 34.6 | 55.0 | 37.1 | 19.3 | 36.9 | 45.9 |
| | | | | ✓ | 37.0$_{+2.4}$ | 57.2$_{+2.2}$ | 39.7$_{+2.6}$ | 19.9$_{+0.6}$ | 39.7$_{+0.8}$ | 50.3$_{+4.4}$ |
| | ResNet50 | 18.1 | 43.57 | × | 38.4 | 59.0 | 42.0 | 21.5 | 42.1 | 50.3 |
| | | | | ✓ | 41.5$_{+3.1}$ | 62.2$_{+3.2}$ | 45.1$_{+3.1}$ | 23.5$_{+2.0}$ | 45.0$_{+2.9}$ | 55.3$_{+5.0}$ |
| | ResNet101 | 14.2 | 62.57 | × | 39.8 | 60.1 | 43.3 | 22.5 | 43.6 | 52.8 |
| | | | | ✓ | 43.9$_{+4.1}$ | 64.2$_{+4.1}$ | 48.1$_{+3.9}$ | 25.3$_{+2.8}$ | 48.0$_{+4.4}$ | 58.7$_{+5.9}$ |
| | ViT-B-32 | 11.8 | 104.49 | × | 30.9 | 50.5 | 31.7 | 9.7 | 33.7 | 51.5 |
| | | | | ✓ | 34.4$_{+3.5}$ | 54.9$_{+4.4}$ | 34.7$_{+3.0}$ | 11.9$_{+2.2}$ | 37.5$_{+3.8}$ | 56.0$_{+4.5}$ |
| | ViT-B-16 | 3.4 | 104.49 | × | 37.8 | 57.4 | 40.1 | 17.8 | 41.4 | 57.3 |
| | | | | ✓ | 41.9$_{+4.1}$ | 61.0$_{+3.6}$ | 43.4$_{+3.3}$ | 21.3$_{+3.5}$ | 44.9$_{+3.5}$ | 62.9$_{+5.6}$ |
| Cascade RCNN | ResNet50 | 15.4 | 71.22 | × | 41.0 | 59.4 | 44.4 | 22.7 | 44.4 | 54.3 |
| | | | | ✓ | 44.4$_{+3.4}$ | 62.7$_{+3.3}$ | 48.3$_{+3.9}$ | 24.8$_{+2.1}$ | 48.0$_{+3.6}$ | 59.3$_{+5.0}$ |
| | ResNet101 | 11.7 | 90.21 | × | 42.5 | 60.7 | 46.4 | 23.5 | 46.5 | 56.4 |
| | | | | ✓ | 45.2$_{+2.7}$ | 63.5$_{+2.8}$ | 49.4$_{+3.0}$ | 26.2$_{+2.7}$ | 48.7$_{+2.2}$ | 60.8$_{+4.4}$ |
| Dynamic RCNN | ResNet18 | 28.1 | 30.57 | × | 35.0 | 55.2 | 37.4 | 20.1 | 37.4 | 45.8 |
| | | | | ✓ | 38.2$_{+3.2}$ | 56.5$_{+1.3}$ | 41.8$_{+4.4}$ | 20.1$_{+0.0}$ | 40.7$_{+3.3}$ | 53.2$_{+7.4}$ |
| | ResNet50 | 18.1 | 43.57 | × | 39.8 | 58.3 | 43.2 | 23.0 | 42.8 | 52.4 |
| | | | | ✓ | 42.8$_{+3.0}$ | 61.2$_{+2.9}$ | 47.0$_{+3.8}$ | 23.9$_{+0.9}$ | 46.2$_{+3.4}$ | 57.7$_{+5.3}$ |
| | ResNet101 | 14.2 | 62.57 | × | 41.2 | 59.7 | 45.3 | 24.0 | 44.9 | 54.3 |
| | | | | ✓ | 44.8$_{+3.6}$ | 63.0$_{+3.3}$ | 48.9$_{+3.6}$ | 25.0$_{+1.0}$ | 48.9$_{+4.0}$ | 60.4$_{+6.1}$ |
| Grid RCNN | ResNet18 | 26.7 | 66.37 | × | 36.6 | 54.2 | 39.7 | 20.1 | 39.8 | 48.2 |
| | | | | ✓ | 38.8$_{+2.2}$ | 56.7$_{+2.5}$ | 41.5$_{+1.8}$ | 21.1$_{+1.0}$ | 41.6$_{+1.8}$ | 52.7$_{+4.5}$ |
| | ResNet50 | 14.0 | 66.37 | × | 40.4 | 58.4 | 43.6 | 22.8 | 43.9 | 53.3 |
| | | | | ✓ | 42.6$_{+2.2}$ | 61.1$_{+2.7}$ | 46.1$_{+2.5}$ | 24.2$_{+1.4}$ | 46.6$_{+2.7}$ | 55.8$_{+2.5}$ |
| | ResNet101 | 11.0 | 85.36 | × | 41.6 | 59.8 | 45.0 | 23.7 | 45.7 | 54.7 |
| | | | | ✓ | 44.8$_{+3.2}$ | 63.6$_{+3.8}$ | 48.9$_{+3.9}$ | 26.5$_{+2.8}$ | 48.9$_{+3.2}$ | 59.6$_{+1.9}$ |
| RetinaNet | RegNet-800M | 22.4 | 19.27 | × | 35.6 | 54.7 | 37.7 | 19.7 | 39.0 | 47.8 |
| | | | | ✓ | 38.4$_{+2.8}$ | 57.4$_{+3.3}$ | 40.7$_{+3.0}$ | 21.4$_{+1.7}$ | 42.0$_{+3.0}$ | 52.3$_{+4.5}$ |
| | ResNet18 | 25.8 | 23.30 | × | 33.4 | 51.8 | 35.1 | 16.9 | 35.6 | 44.9 |
| | | | | ✓ | 35.9$_{+2.5}$ | 54.4$_{+3.3}$ | 38.0$_{+2.9}$ | 17.9$_{+1.0}$ | 39.1$_{+3.5}$ | 49.4$_{+4.5}$ |
| | ResNet50 | 17.7 | 37.74 | × | 37.4 | 56.7 | 39.6 | 20.0 | 40.7 | 49.7 |
| | | | | ✓ | 39.6$_{+2.2}$ | 58.8$_{+2.1}$ | 42.1$_{+2.5}$ | 22.7$_{+2.7}$ | 43.3$_{+2.6}$ | 52.5$_{+2.8}$ |
| | ResNet101 | 13.5 | 56.74 | × | 38.9 | 58.0 | 41.5 | 21.0 | 42.8 | 52.4 |
| | | | | ✓ | 41.3$_{+2.4}$ | 60.8$_{+2.8}$ | 44.3$_{+2.8}$ | 22.7$_{+1.7}$ | 46.0$_{+3.2}$ | 55.2$_{+2.8}$ |
| Fsaf RetinaNet | ResNet50 | 20.0 | 36.19 | × | 37.8 | 56.8 | 39.8 | 20.4 | 41.1 | 48.8 |
| | | | | ✓ | 41.4$_{+3.6}$ | 61.0$_{+3.2}$ | 44.2$_{+4.4}$ | 23.1$_{+2.7}$ | 45.2$_{+4.1}$ | 55.2$_{+6.4}$ |
| | ResNet101 | 15.0 | 55.19 | × | 39.3 | 58.6 | 42.1 | 22.1 | 43.4 | 51.2 |
| | | | | ✓ | 42.6$_{+3.3}$ | 62.0$_{+3.4}$ | 45.5$_{+3.4}$ | 24.5$_{+2.4}$ | 47.0$_{+2.6}$ | 56.2$_{+5.0}$ |
| RepPoints | ResNet50 | 18.2 | 36.62 | × | 38.6 | 59.6 | 41.6 | 22.5 | 42.2 | 50.4 |
| | | | | ✓ | 40.6$_{+2.0}$ | 61.7$_{+2.1}$ | 43.8$_{+2.2}$ | 23.4$_{+0.9}$ | 44.6$_{+2.4}$ | 53.0$_{+2.6}$ |
| | ResNet101 | 13.2 | 55.62 | × | 40.5 | 61.3 | 43.5 | 23.4 | 44.7 | 53.2 |
| | | | | ✓ | 42.7$_{+2.2}$ | 63.7$_{+2.4}$ | 46.4$_{+2.9}$ | 24.9$_{+1.5}$ | 47.2$_{+2.5}$ | 56.4$_{+3.2}$ |
| Yolo v3 | DarkNet53 | 42.2 | 61.95 | × | 33.4 | 56.3 | 35.2 | 19.5 | 36.4 | 43.6 |
| | | | | ✓ | 35.8$_{+2.4}$ | 58.2$_{+1.9}$ | 38.1$_{+2.9}$ | 21.2$_{+1.7}$ | 39.0$_{+2.6}$ | 45.6$_{+2.0}$ |
| SSD | VGG16 | 26.1 | 38.08 | × | 29.4 | 49.3 | 31.0 | 11.7 | 34.1 | 44.9 |
| | | | | ✓ | 31.2$_{+1.8}$ | 52.1$_{+2.8}$ | 32.8$_{+1.8}$ | 12.6$_{+0.9}$ | 37.4$_{+3.3}$ | 46.2$_{+1.3}$ |

The architecture and performance of the teacher detectors utilized in these experiments are reported in Table I and introduced in Section IV-B.

both convolutional networks and transformers. (v) Compared to twelve other knowledge distillation methods, our method consistently outperforms them by a significant margin. For instance, on Faster RCNN, our method achieves an AP that is 0.6 higher than the second-best distillation method. Interestingly, traditional prediction-based knowledge distillation only yields a modest 0.5 AP improvement compared to the student without knowledge distillation, indicating its limited effectiveness in object detection. We attribute this to the fact that prediction KD primarily enhances classification ability but falls short in improving localization. (vi) Deeper detectors with ResNet50 and ResNet101 backbones benefit from knowledge distillation,

TABLE III
EXPERIMENTS ON MS COCO2017 WITH THE PROPOSED DISTILLATION METHOD ON INSTANCE SEGMENTATION

| Model | Backbone | Params | FPS | Distill | Bounding box AP | | | | Mask AP | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | AP | $AP_S$ | $AP_M$ | $AP_L$ | AP | $AP_S$ | $AP_M$ | $AP_L$ |
| Mask RCNN | ResNet50 | 44.17 | 17.4 | × | 39.2 | 22.9 | 42.6 | 51.2 | 35.4 | 19.1 | 38.6 | 48.4 |
| | | | | ✓ | 41.7$_{+2.5}$ | 23.4$_{+0.5}$ | 45.3$_{+2.7}$ | 55.8$_{+4.6}$ | 37.4$_{+2.0}$ | 19.7$_{+0.6}$ | 40.5$_{+1.9}$ | 52.1$_{+3.7}$ |
| | ResNet101 | 63.16 | 13.5 | × | 40.8 | 23.0 | 45.0 | 54.1 | 36.6 | 19.2 | 40.2 | 50.5 |
| | | | | ✓ | 43.0$_{+2.2}$ | 24.7$_{+1.7}$ | 47.2$_{+2.2}$ | 57.1$_{+3.0}$ | 38.7$_{+2.1}$ | 20.7$_{+1.5}$ | 42.3$_{+2.1}$ | 53.3$_{+2.8}$ |
| Cascade Mask RCNN | ResNet50 | 77.10 | 16.1 | × | 41.9 | 23.2 | 44.9 | 55.9 | 36.5 | 18.9 | 39.2 | 50.7 |
| | | | | ✓ | 43.8$_{+1.9}$ | 24.9$_{+1.7}$ | 47.2$_{+2.3}$ | 58.4$_{+2.5}$ | 38.0$_{+1.5}$ | 20.2$_{+1.3}$ | 40.9$_{+1.7}$ | 52.8$_{+2.1}$ |
| | ResNet101 | 96.09 | 13.1 | × | 42.9 | 24.4 | 46.5 | 57.0 | 37.3 | 19.7 | 40.6 | 51.5 |
| | | | | ✓ | 45.4$_{+2.5}$ | 26.3$_{+1.9}$ | 49.0$_{+2.5}$ | 60.9$_{+3.9}$ | 39.6$_{+2.3}$ | 21.3$_{+2.6}$ | 42.8$_{+2.2}$ | 55.0$_{+3.5}$ |
| Mask Scoring RCNN | ResNet50 | 60.51 | 18.0 | × | 38.8 | 21.7 | 41.9 | 51.8 | 36.3 | 18.8 | 39.3 | 50.8 |
| | | | | ✓ | 41.5$_{+2.7}$ | 24.3$_{+2.6}$ | 45.5$_{+3.6}$ | 53.8$_{+2.0}$ | 38.5$_{+2.2}$ | 20.7$_{+1.9}$ | 42.0$_{+2.7}$ | 52.3$_{+1.5}$ |
| | ResNet101 | 79.40 | 15.0 | × | 42.6 | 24.4 | 46.2 | 56.6 | 38.1 | 20.0 | 41.5 | 53.5 |
| | | | | ✓ | 45.6$_{+3.0}$ | 28.4$_{+4.0}$ | 49.0$_{+2.8}$ | 58.3$_{+1.7}$ | 39.7$_{+1.6}$ | 20.6$_{+0.6}$ | 42.6$_{+1.1}$ | 56.0$_{+2.5}$ |
| Swin Transformer | Swin-T | 69.11 | 10.4 | × | 42.7 | 26.5 | 45.9 | 56.6 | 39.3 | 20.5 | 41.8 | 57.8 |
| | | | | ✓ | 45.2$_{+2.5}$ | 27.6$_{+1.1}$ | 48.9$_{+3.0}$ | 58.5$_{+1.9}$ | 41.6$_{+2.3}$ | 22.9$_{+2.4}$ | 44.8$_{+3.0}$ | 59.4$_{+1.6}$ |

The architecture and performance of the teacher detectors utilized in these experiments are reported in Table I and introduced in Section IV-B.

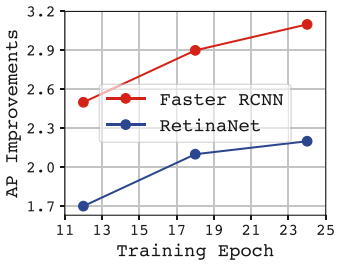



Fig. 8. Relation between the benefits of knowledge distillation and the number of training epochs on Faster RCNN and RetinaNet.

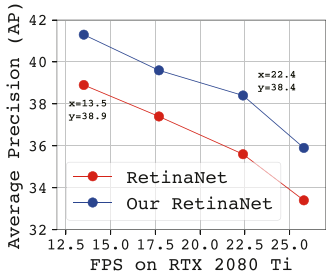



Fig. 10. Knowledge distillation on RetinaNet with different backbones (Resnet18, Regnet800 M, Resnet50, and ResNet101).

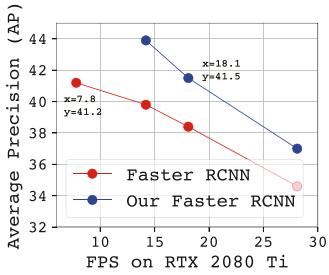



Fig. 9. Knowledge distillation on Faster RCNN with different backbones (Resnet18, Resnet50, Resnet101, and ResNeXt101).

exhibiting AP improvements of 2.7 and 2.9, respectively. (vii) The FPS-AP curves of detectors trained with and without knowledge distillation are depicted in Figs. 9 and 10, respectively. Notably, on RetinaNet, replacing a larger detector (ResNet101 backbone) trained without knowledge distillation with a smaller detector (RegNet800 M backbone) yields a 1.67× acceleration, 2.94× compression, and a 0.5 mAP drop. Similarly, on Faster RCNN, replacing a larger detector (ResNeXt101 backbone) trained without knowledge distillation with a smaller detector (ResNet50 backbone) trained with knowledge distillation leads to a 2.32× acceleration, 2.27× compression, and a 0.3 mAP improvement. These observations highlight the ability of our proposed knowledge distillation method to significantly accelerate and compress neural networks while maintaining or even improving mean average precision.

### D. KD Improves Model Robustness

In addition to accuracy and computational efficiency, another crucial metric in object detection is the ability to handle various image corruptions, such as noise, blurring, and adverse weather conditions [125], [145]. A robust detector is expected to process these corrupted images without the need for additional data augmentation during training, thereby exhibiting better domain generalization capabilities [146]. In this subsection, we evaluate the robustness of detectors trained with and without knowledge distillation using the COCO-C dataset [125]. COCO-C is an evaluation dataset derived from the validation set of COCO, enriched with four types of image corruption, including noise, blurring, weather, and digital corruption. Each corruption type further comprises several fine-grained corruptions. The mAP of Faster RCNN models trained with and without knowledge distillation on corrupted images is compared in Table IV. The results consistently demonstrate that the distilled detector achieves higher mAP scores across all types of corruption, indicating the

TABLE IV
EXPERIMENTAL RESULTS (MAP) OF FASTER RCNN WITH RESNET50 BACKBONE ON COCO2017 WITH DIFFERENT TYPES OF IMAGE CORRUPTION

| Distill | Noise Corruption | | | Blur Corruption | | | | Weather Corruption | | | | Digital Corruption | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Gaussian | Shot | Impulse | Defocus | Glass | Motion | Zoom | Snow | Frost | Fog | Bright | Contrast | Elastic | Pixelate | Jpeg |
| × | 16.8 | 16.8 | 13.4 | 17.6 | 10.9 | 16.4 | 7.4 | 17.6 | 22.1 | 30.8 | 33.6 | 22.4 | 21.2 | 14.2 | 15.4 |
| ✓ | 19.9$_{+3.1}$ | 20.1$_{+3.3}$ | 16.5$_{+3.1}$ | 20.1$_{+2.5}$ | 12.6$_{+1.7}$ | 18.2$_{+1.8}$ | 8.3$_{+0.9}$ | 19.4$_{+1.8}$ | 24.2$_{+2.1}$ | 33.7$_{+2.9}$ | 36.4$_{+2.8}$ | 24.2$_{+1.8}$ | 23.9$_{+2.7}$ | 16.2$_{+2.0}$ | 18.1$_{+2.7}$ |

TABLE V
COMPARISON BETWEEN OUR METHOD AND OTHER DISTILLATION METHODS ON FASTER RCNN WITH RESNET50 BACKBONE

| KD Method | AP | AP$_{50}$ | AP$_{75}$ | AP$_S$ | AP$_M$ | AP$_L$ |
|---|---|---|---|---|---|---|
| without KD | 38.4 | 59.0 | 42.0 | 21.5 | 42.1 | 50.3 |
| Prediction KD [15] | 38.9 | 59.6 | 42.3 | 21.0 | 42.8 | 52.4 |
| Adriana *et al.* [24] | 39.6 | 60.1 | 43.3 | 22.5 | 42.8 | 52.2 |
| Kang *et al.*† [71] | 40.9 | – | – | 24.5 | 44.0 | 52.0 |
| Sun *et al.*† [134] | 40.1 | – | – | 23.0 | 43.6 | 53.0 |
| Chen *et al.* [35] | 38.7 | 59.0 | 42.1 | 22.0 | 41.9 | 51.0 |
| Wang *et al.* [36] | 39.1 | 59.8 | 42.8 | 22.2 | 42.9 | 51.1 |
| Li *et al.* [17] | 39.6 | 60.1 | 43.3 | 22.5 | 42.8 | 52.2 |
| Heo *et al.* [28] | 38.9 | 60.1 | 42.6 | 21.8 | 42.7 | 50.7 |
| Guo *et al.*† [73] | 40.9 | – | – | 23.6 | 44.8 | 53.3 |
| Du *et al.*† [135] | 39.5 | 60.1 | 43.3 | 22.3 | 43.6 | 51.7 |
| Zhang *et al.* [136] | 39.2 | – | – | – | – | – |
| Dai *et al.*† [72] | 40.2 | 60.7 | 43.8 | 22.7 | 44.0 | 53.2 |
| **Our method** | **41.5** | **62.2** | **45.1** | **23.5** | **45.0** | **55.3** |

Methods marked with † are proposed after our conference version. Some results are missing because their origin papers does not report them.
The numbers in bold indicate the highest average precision.

TABLE VI
EXPERIMENTAL RESULTS ON CITYSCAPES

| Model | Backbone | Distill | Box AP | Mask AP |
|---|---|---|---|---|
| Faster RCNN | ResNet50 | × | 40.3 | – |
| | | ✓ | 43.5$_{+3.2}$ | – |
| Mask RCNN | ResNet50 | × | 41.0 | 35.8 |
| | | ✓ | 43.0$_{+2.0}$ | 37.5$_{+1.7}$ |

TABLE VII
ABLATION STUDY ON THE THREE DISTILLATION LOSS IN OUR METHOD

| Loss | | | | | | | |
|---|---|---|---|---|---|---|---|
| | $\mathcal{L}_{\text{AT}}$ | × | ✓ | × | × | ✓ | ✓ |
| | $\mathcal{L}_{\text{AM}}$ | × | × | ✓ | × | ✓ | ✓ |
| | $\mathcal{L}_{\text{NLD}}$ | × | × | × | ✓ | × | ✓ |
| **Result** | **AP** | 38.4 | 39.6 | 40.8 | 39.8 | 41.2 | **41.5** |
| | **AP**$_S$ | 21.5 | 22.7 | 22.8 | 22.7 | 23.0 | **23.5** |
| | **AP**$_M$ | 42.1 | 42.9 | 44.3 | 43.1 | 44.6 | **45.0** |
| | **AP**$_L$ | 50.3 | 52.5 | 54.3 | 52.3 | 55.3 | **55.3** |

Experiments are conducted with faster rcnn students with RESNET50 backbones on MS COCO2017.
The numbers in bold indicate the highest average precision.

effectiveness of our method in enhancing the robustness and domain generalization abilities of object detection.

## E. Ablation Study

*1) Ablation Study on Knowledge Distillation Loss:* Table VII shows the ablation study on the proposed attention-guided distillation ($\mathcal{L}_{\text{AT}}$ and $\mathcal{L}_{\text{AM}}$) and non-local distillation ($\mathcal{L}_{\text{NLD}}$).

TABLE VIII
ABLATION STUDY ON THE SPATIAL ATTENTION AND CHANNEL ATTENTION

| Attention Type | Spatial | × | ✓ | × | ✓ |
|---|---|---|---|---|---|
| | Channel | × | × | ✓ | ✓ |
| **Result** | **AP** | 38.4 | 41.0 | 40.7 | **41.2** |
| | **AP**$_S$ | 21.5 | 22.7 | 22.9 | **23.0** |
| | **AP**$_M$ | 42.1 | 44.7 | 44.1 | **44.6** |
| | **AP**$_L$ | 50.3 | 54.2 | 54.1 | **55.3** |

The numbers in bold indicate the highest average precision.

It is observed that: (i) Attention-guided distillation and non-local distillation lead to 2.8 and 1.4 AP improvements, respectively. (ii) $\mathcal{L}_{\text{AT}}$ and $\mathcal{L}_{\text{AM}}$ lead to 1.2 and 2.4 AP improvements respectively, indicating that most of the benefits of attention-guided distillation are obtained from the feature loss masked by the attention maps ($\mathcal{L}_{\text{AM}}$). (iii) There are 3.1 AP improvements with the combination of attention-guided distillation and non-local distillation. These observations indicate that each distillation loss in our method has its individual effectiveness, and they can be utilized together to achieve better performance.

*2) Ablation Study on Attention Types:* In contrast to previous attention-based knowledge distillation methods, our method incorporates both spatial attention and channel attention in the attention-guided distillation process. In this subsection, we conducted an ablation study to assess the individual effectiveness of these two types of attention using Faster RCNN (ResNet50 backbone) on the MS COCO2017 dataset. Table VIII presents the results of the ablation study, revealing that spatial attention and channel attention yield improvements of 2.6 and 2.3 in AP (average precision), respectively. Notably, the combination of the two types of attention leads to a further improvement of 2.8 in AP. These findings indicate that both spatial attention and channel attention exhibit their own effectiveness and can be synergistically employed to achieve enhanced performance.

## F. Sensitivity Study

*1) Sensitivity Study on Hyperparameters:* This subsection investigates the sensitivity of the proposed method to four hyperparameters: $\alpha$, $\beta$, $\gamma$, and $T$. These hyperparameters are employed to balance the magnitude of different distillation loss functions and adjust the distribution of attention masks. The sensitivity study is conducted on the MS COCO2017 dataset using Faster RCNN with a ResNet50 backbone. The results, illustrated in Fig. 11, reveal that even the worst hyperparameter settings lead to a mere 0.3 AP drop compared to the highest AP, while still maintaining a remarkable 2.9 AP improvement

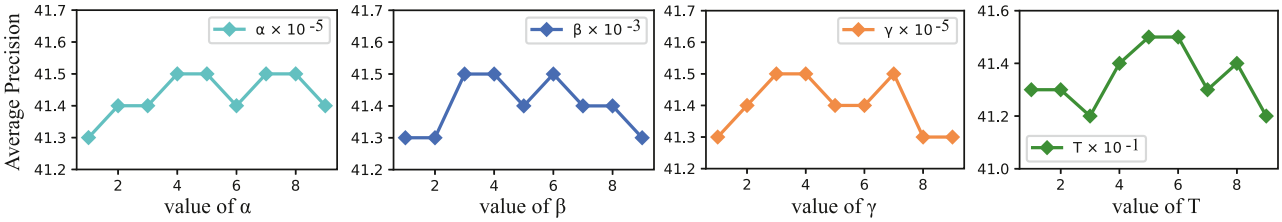



Fig. 11. Hyper-parameter sensitivity study of $\alpha, \beta, \gamma, T$ with Faster RCNN with ResNet50 backbones on MS COCO2017.

TABLE IX
RESULTS OF DIFFERENT TYPES OF NON-LOCAL MODULES WITH FASTER RCNN (RESNET50 BACKBONE) ON MS COCO2017

| Non-Local Type | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|
| **Embedded Gaussian** | **41.5** | **62.2** | **45.2** | 23.4 | **45.0** | 55.1 |
| Dot Production | 41.4 | 59.6 | 44.7 | 23.6 | 45.3 | 53.6 |
| **Concatenation** | **41.5** | **62.2** | 45.1 | **23.5** | **45.0** | 55.3 |
| Gaussian | 41.3 | 59.9 | 44.6 | **23.5** | 45.0 | 54.8 |
| Criss Cross | 41.3 | 61.8 | 44.9 | 22.6 | 44.8 | **56.1** |
| RCCA | 41.4 | 61.9 | 45.1 | 22.8 | 44.8 | 56.0 |

The numbers in bold indicate the highest average precision.

over the baseline model. These findings indicate that our method exhibits robustness to the choice of hyperparameters.

*2) Sensitivity Study on Types of Non-Local Modules:* There are four original types of non-local modules: Gaussian, embedded Gaussian, dot product, and concatenation. Additionally, recent efficient non-local modules, such as the Criss Cross module and the RCCA module [114], [116], have been proposed. The performance of our method with different types of non-local modules is summarized in Table IX. The results indicate that the lowest-performing non-local type (Gaussian) exhibits only a 0.2 AP difference compared to the highest-performing types (Embedded Gaussian and Concatenation). Furthermore, the utilization of efficient non-local modules does not significantly impact the performance. These findings suggest that our method demonstrates robustness to the choice of non-local modules.

### G. Influence From Adaptation Layers

The adaptation layer, also known as the regression layer, plays a crucial role in knowledge distillation. Originally, knowledge distillation focused on distilling knowledge from predicted probability distributions that have the same shape for both students and teachers [15]. However, in feature-based knowledge distillation, the features of students and teachers often have different shapes, especially when they have different architectures or varying numbers of channels. As a result, their distance cannot be directly computed. To address this challenge, Romero et al. proposed the use of an adaptation layer in FitNet [24] to adjust the shape of student and teacher features for computing their $L_2$-norm distance in knowledge distillation. Subsequently, this technique has been adopted in a wide range of knowledge distillation methods [29], [36], [118], [119], [120]. Recent research suggests that proper regularization of adaptation layers can lead to slight performance improvements [29]. Notably, Chen et al. investigated the influence of adaptation layers in knowledge distillation for object detection and demonstrated that even when students and teachers have the same feature shape, adaptation layers can significantly improve the precision of students [35]. Motivated by these observations, we thoroughly examine the impact of adaptation layers in our proposed method. The adaptation layers in our method can be summarized as follows:

- For non-local distillation and attention-mask distillation, the features of students to be distilled are represented as $\mathbb{R}^{C\times H\times W}$, and we employ a 1x1 convolutional layer for each of them, respectively.
- For channel-wise attention in attention-guided knowledge distillation, the to-be-distilled features of students and teachers are $\mathbb{R}^{C}$. Thus, we employ a fully-connected layer as the adaptation layer.
- For spatial-wise attention in attention-guided knowledge distillation, the to-be-distilled features of students and teachers are $\mathbb{R}^{W\times H}$. To transfer spatial knowledge, we first reshape them into $\mathbb{R}^{1\times W\times H}$ and then employ a 3x3 convolutional layer as the adaptation layer.

All adaptation layers are initialized and trained simultaneously with the weights of students, employing the same optimizer and learning rate. Our experiments conducted on MS COCO2017 with Faster RCNN demonstrate that the removal of the aforementioned three adaptation layers results in decrements of 0.4, 0.2, and 0.1 AP, respectively, indicating the positive influence of adaptation layers on model performance. We argue that this influence stems from the ability of adaptation layers to alleviate the learning difficulty of students in the context of knowledge distillation. Without adaptation layers, the student detector is expected to generate features identical to those of the teacher. Conversely, with the utilization of adaptation layers, the student detector is expected to generate features that undergo a linear transformation identical to the teacher features. This relaxation of optimization constraints in knowledge distillation eases the training process of student detectors. Experimental results further demonstrate that the student models equipped with adaptation layers achieve lower knowledge distillation losses, providing empirical validation for the aforementioned conjecture.

### H. Impact of Training Time on Knowledge Distillation

In this subsection, we investigate the influence of training time on the effectiveness of knowledge distillation. Experimental results obtained from training Faster RCNN and RetinaNet for different durations (12, 18, and 24 epochs) are depicted in


Fig. 12. Qualitative analysis of MS COCO2017 with Faster RCNN (ResNet50 backbone) trained with and without knowledge distillation. (a–k) shows the detection results of eleven different input images.

Fig. 8. It is evident from the results that increasing the training time leads to a substantial improvement in mAP (mean average precision) achieved through knowledge distillation for both detectors. Specifically, the mAP improvements attained with 24 training epochs exceed those obtained with 12 epochs by 24% and 29% on Faster RCNN and RetinaNet, respectively. These findings clearly indicate that knowledge distillation benefits from longer training durations. We posit that the rationale behind this observation lies in the nature of the knowledge transferred from the teachers. In comparison to the supervision provided by ground truth labels, the knowledge conveyed by teachers exhibits higher dimensionality and greater variance. Consequently, an extended training time is necessary for the detectors to converge and effectively exploit the knowledge distilled from the teachers.

## V. Discussion

### A. Analysis of the Benefits of Knowledge Distillation

*1) Qualitative Analysis:* Fig. 12 illustrates the comparison of detection results between detectors trained with and without knowledge distillation. The following observations are made: (i) Knowledge distillation enhances the detection ability of small objects. Subfigures (a-c) demonstrate that the distilled model accurately detects cars, a handbag, and a person inside a car, respectively. (ii) Knowledge distillation improves the generation of bounding boxes. Subfigures (d-e) reveal that the baseline model generates multiple bounding boxes for the boat and the train, while the distilled model avoids these errors. (iii) Knowledge distillation enables detectors to classify objects more accurately. In subfigure (j), the baseline model incorrectly classifies the woman's hand as a cat, whereas the distilled model does not exhibit this issue. (iv) Knowledge distillation enhances the detection ability of dense objects. Subfigure (h) illustrates that the distilled model can detect a significantly larger number of carrots on a plate compared to the baseline model. In summary, these observations demonstrate the positive impact of knowledge distillation on various aspects of object detection performance, including the detection of small objects, improved bounding box generation, enhanced object classification, and improved detection of dense objects.

*2) Analysis of Detection Error Types:* We conducted an analysis of different types of detection errors in both the baseline and distilled detectors, as depicted in Fig. 13. The distribution of error types across all categories in MS COCO2017 is presented in four columns, specifically focusing on the categories of 'Person', 'Car', and 'Stop Sign'. The numbers in the figure legends represent the AUC (area under the curve). The analysis reveals that our distillation method effectively reduces errors across various categories, including errors in localization (Loc) as well as classification errors for both similar (Sim) and dissimilar (Oth) categories. In summary, our method demonstrates improved performance in terms of both localization and classification abilities.

### B. Relation Between Students and Teachers

*1) An Accurate Teacher is Usually a Good Teacher:* A considerable body of research has focused on investigating the

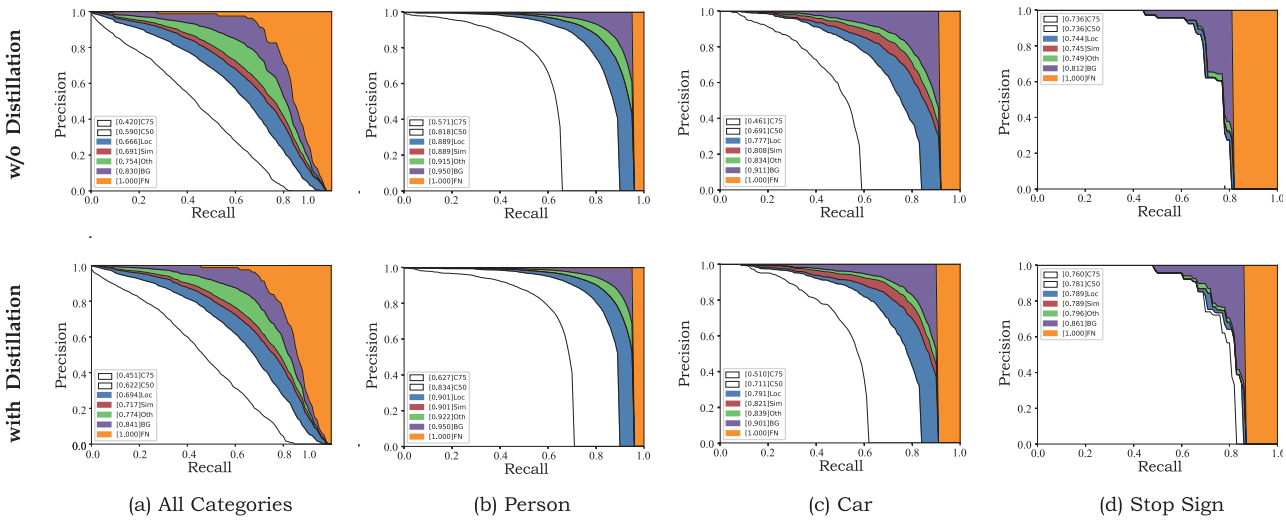



Fig. 13. Distribution of error types on Faster RCNN50 trained with and without knowledge distillation on (a) All Categories, (b) Category of Person, (c) Car and (d) Stop Sign. **C75** - Area under curve corresponds to AP IoU=0.75 metric. **C50** - Area under curve corresponds to AP IoU=.50 metric. **Loc** - Localization error; **Sim** - Classification error on similar classes; **Oth** - Classification error on not similar classes; **BG** - False positive prediction fired on background. **FN** - False Negative prediction.

relation between students and teachers. Mirzadeh et al. and Cho et al. [22], [23] have demonstrated that a teacher with higher accuracy may not necessarily be the most effective teacher for knowledge distillation. In fact, there are instances where a teacher with excessively high accuracy can potentially impede the performance of students. Furthermore, Hossein et al. [147] and Li et al. [148] have shown that even the same model, or a model with lower accuracy compared to the student, can be employed as a suitable teacher for knowledge distillation. However, it is important to note that these experiments were primarily conducted within the domain of image classification. In this subsection, we investigate whether these observations remain consistent in the context of object detection.

Experiments were conducted on Faster RCNN and Cascade RCNN students, both utilizing ResNet50 backbones, with teacher models exhibiting different average precision (AP) values, as depicted in Fig. 7. The results revealed the following observations: (i) Across all conducted experiments, students trained with higher AP teachers consistently achieved higher AP themselves. This positive relation is supported by Pearson correlation coefficients of 0.86 and 0.96 between the AP values of students and teachers for Faster RCNN and Cascade RCNN, respectively, indicating a strong positive correlation. (ii) When the teacher's AP is lower or equal to that of the student, the improvements attained through knowledge distillation are limited and may even have a negative impact. These findings suggest that an inaccurate teacher has the potential to adversely affect the performance of students. Taken together, these observations indicate that the relation between students and teachers in the context of object detection differs from that observed in image classification. Our experimental results strongly suggest a significant positive correlation between the AP values of students and teachers. Consequently, employing a high AP teacher is likely to yield substantial performance improvements for students.

We suggest that the reason why a high AP teacher model is crucial in object detection but not very necessary in image classification is that object detection is a more challenging task. As a result, a weaker teacher model may introduce more negative influence on students, which prevents students from achieving higher AP. In contrast, on image classification, most teacher models can achieve a very high training accuracy, so they do not introduce much error to students during knowledge distillation.

*2) KD From Two-Stage Teachers to One-Stage Students:* In object detection, two-stage and one-stage detectors represent the most widely recognized paradigms. In this subsection, we investigate the feasibility of distilling knowledge from a two-stage teacher to a one-stage student. We employ a Faster RCNN with ResNeXt101 backbone as the teacher and a RetinaNet with ResNet50 as the student. Experimental results reveal that in this scenario, knowledge distillation leads to a decrease in the mean average precision of RetinaNet from 37.4 to 35.2. This suggests that the knowledge from the two-stage teacher is detrimental rather than beneficial for a one-stage student.

*3) KD From Multi-Task Teachers to Single-Task Students:* We further investigate the potential for knowledge distillation from a multi-task teacher to a single-task student. Initially, we train a Mask RCNN teacher on both object detection and instance segmentation using MS COCO2017 dataset. Subsequently, we distill this knowledge to a Faster RCNN student, which solely focuses on object detection. Experimental results reveal that the student achieves a 3.1 mAP improvement on MS COCO (from 38.4 to 41.5), surpassing even a single-task teacher (e.g., Faster RCNN teacher), with a margin of 0.2 mAP. These findings indicate that knowledge can be effectively distilled from a multi-task teacher to a single-task student, particularly when the tasks are closely related (e.g., detection and segmentation).

TABLE X
COMPARISON OF OUR METHOD IN ONE-STAGE AND TWO-STAGE KNOWLEDGE DISTILLATION SETTING

| KD Framework | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|
| without KD | 38.4 | 59.0 | 42.0 | 21.5 | 42.1 | 50.3 |
| One-Stage KD | 40.7 | 60.9 | 44.5 | 22.6 | 44.4 | 53.9 |
| Two-Stage KD, Weak Teacher | 40.5 | 60.4 | 43.0 | 22.4 | 44.7 | **56.5** |
| **Two-stage KD, Strong Teacher** | **41.5** | **62.2** | **45.1** | **23.5** | **45.0** | 55.3 |

Experiments are conducted on MS COCO2017 with faster RCNN students with RESNET50 backbones.
The numbers in bold indicate the highest average precision.

TABLE XI
ABLATION STUDY ON SUPERVISION FROM LABELS

| Supervision | | | mAP |
|---|---|---|---|
| Feature KD | Predication KD | Labels | |
| × | × | ✓ | 38.4 |
| ✓ | × | × | 0.0 |
| × | ✓ | × | 38.1 |
| × | ✓ | ✓ | 38.9 |
| ✓ | ✓ | × | 41.4 |
| ✓ | × | ✓ | 41.5 |
| ✓ | ✓ | ✓ | 41.7 |

*4) KD From Single-Task Teachers to Multi-Task Students:* This subsection investigates the potential of utilizing knowledge from a single-task teacher to enhance the performance of multi-task students. Initially, we train a Faster RCNN teacher specifically for object detection. Subsequently, we distill this knowledge to a Mask RCNN student, which is trained for both object detection and instance segmentation. Our experimental results demonstrate a notable improvement of 1.6 mask AP (from 35.4 to 37.0) in instance segmentation, even when the teacher is not explicitly trained for this task. We argue that this improvement arises from the non-task-biased nature of the knowledge transferred from teachers, thereby benefiting relevant visual tasks.

*5) Two-Stage KD versus One-Stage KD:* Traditional knowledge distillation follows a two-stage training pipeline, where a large teacher model is initially trained and then distilled to a lightweight student model. Recently, researchers have proposed a one-stage knowledge distillation approach called online knowledge distillation or deep mutual learning. These methods involve training two or multiple student models and distilling their knowledge to each other. In this subsection, we compare the effectiveness of our method in both the two-stage and one-stage knowledge distillation frameworks. Experiments are conducted on MS COCO2017 using Faster RCNN students with ResNet50 backbones. For one-stage knowledge distillation experiments, we adopt the deep mutual learning approach and train two student detectors to mimic each other using our methods. For two-stage knowledge distillation experiments, we explore the following two schemes.

- *Strong Teacher Scheme:* distilling knowledge from a pre-trained Cascade Mask RCNN teacher with ResNetX101-DCN backbone, which achieves 47.3 AP.
- *Weak Teacher Scheme:* distilling knowledge from a pre-trained teacher which has the identical architecture with the student (i.e., Faster RCNN with ResNet50 backbone) and achieves 38.4 AP.

Experimental results are shown in Table X. It is observed that one-stage knowledge distillation leads to 2.3 AP improvements over the baseline, which is 0.2 higher than two-stage knowledge distillation with a weak teacher, but still 0.8 AP lower than the two-stage knowledge distillation with a strong teacher. This observation indicates that one-stage knowledge distillation can achieve comparable and even better performance than two-stage knowledge distillation when no powerful teacher is available. Moreover, two-stage knowledge distillation can make more use of the knowledge from a powerful teacher and achieve better performance.

### C. Ablation Study on Supervision From Labels

There are usually three kinds of supervision in knowledge distillation, including the supervision from labels in datasets, feature knowledge distillation, and prediction knowledge distillation. In this subsection, we study whether the two knowledge distillation supervision can replace the label supervision in object detection.

- *Label Supervision*: The supervision from the annotations in datasets, which is utilized to compute the original training loss of detectors.
- *Feature KD Supervision*: The supervision from distilling the feature of teachers with the KD methods proposed in this paper.
- *Prediction KD Supervision*: The supervision from distilling the prediction of teachers (*i.e.*, training the student with the pesudo labels from the teacher).

The performance of Faster RCNN students with ResNet50 backbones on MS COCO2017 trained with different supervision is shown in Table XI. It is observed that: (i) With only supervision from feature distillation, the training of the student detector fails. We argue that this is because the feature distillation loss is only applied to the backbone of student detectors and the detection heads of the student in this setting can not be trained by any supervision. (ii) With only the supervision from prediction distillation, the student detector has 0.3 lower AP than the baseline (i.e., the student trained with only label supervision). We suggest that this is because the prediction of the teacher contains some error that may mislead student training and thus harms student performance. (iii) By using both the supervision from labels and prediction distillation, 0.5 AP improvements can be obtained. We suggest that this benefit comes from the uncertainty in teacher prediction. (iv) By using both prediction KD and feature KD while not using label supervision, the student achieves 41.4 AP, which is 0.1 lower than using both label supervision and feature KD. This observation indicates that supervision from labels is still necessary even if the knowledge distillation loss is applied. (v) The best performance (41.7 AP) can be obtained with the combination of all three supervision, indicating that the benefits of knowledge distillation on features and prediction are orthogonal.

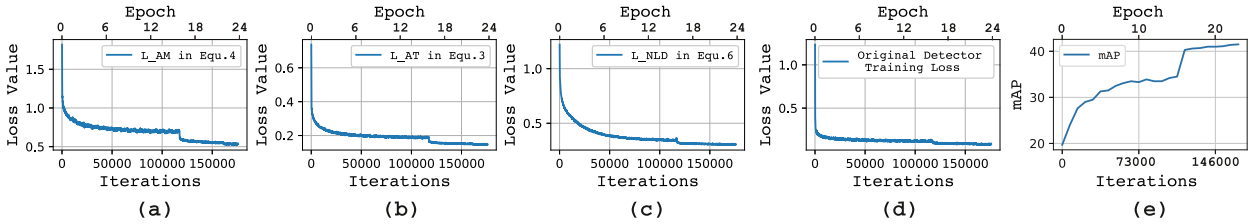



Fig. 14. Curves of different loss and mAP during training of distilled Faster RCNN with a ResNet50 backbone on MS COCO2017. (a-c) are curves of knowledge distillation loss. (d) is the original training loss of Faster RCNN. Note that there are significant mAP improvements and loss reduction at the $16_{th}$ epoch because the learning rate is decayed at this time. (e) shows the mAP on validation set in different training time.

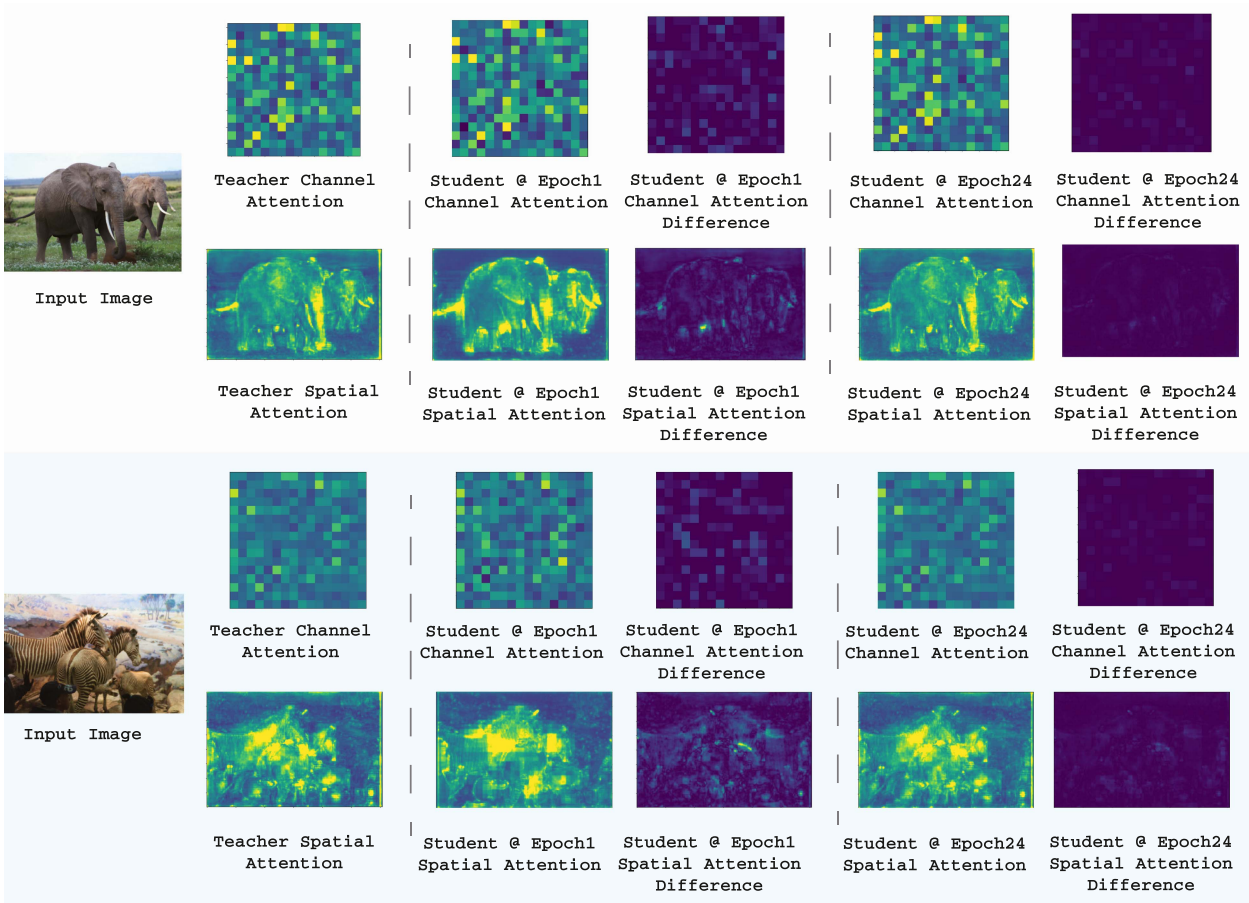



Fig. 15. Comparison of the spatial attention and channel attention between the teacher detector and the student detectors. Note that "student @ Epoch1" and "student @ Epoch24" indicates the Faster RCNN students (ResNet50 backbones) trained for 1 epoch and 24 epochs on MS COCO2017, respectively. Note that we reshape the channel attention from $\mathbb{R}^{256}$ to $\mathbb{R}^{16\times16}$ for better visualization. The subfigures of attention difference show the difference between student attention and teacher attention.

### D. Visualization of Training Behavior

Fig. 14 illustrates the training loss and mAP of Faster RCNN students on MS COCO2017. It is observed that: (i) In the initial iterations (0–20 k), all three knowledge distillation losses and the original training loss decrease rapidly. Moreover, no significant mAP difference is observed between students trained with and without knowledge distillation during this phase. (ii) In the subsequent iterations (20k-120 k), the original training loss stabilizes, while the knowledge distillation loss continues to decrease noticeably. Consequently, the student trained with knowledge distillation achieves a higher AP than the student trained without it. These findings suggest that knowledge distillation facilitates the student's acquisition of knowledge from the teacher when the supervision provided by the original training loss is insufficient. (iii) In the final iterations, after the learning rate decay, both distilled and non-distilled students exhibit significant AP improvements, with the distilled student maintaining its superiority. In summary, the primary distinction between the original training loss of detectors and the knowledge distillation loss lies in their respective behaviors throughout the

training process. The original training loss decreases rapidly in the initial iterations and then stabilizes during the latter part of the training. In contrast, the knowledge distillation loss exhibits a clear decreasing trend throughout the entire training period, indicating its greater optimization challenge and the provision of additional beneficial information beyond the labels. Similarly, in GID [72], Dai et al. demonstrate a gradual increase in the number of selected semi-positive instances during training, indicating the student's heightened focus on valuable yet challenging-to-learn instances. Their observation aligns with our finding that the knowledge distillation loss is more challenging to optimize while providing additional and valuable knowledge to the students.

### E. Visualization of Spatial and Channel Attention

Fig. 15 presents a detailed visualization of the spatial attention and channel attention of the teacher detector and student detectors trained for 1 epoch and 24 epochs. The student detector is a Faster RCNN with a ResNet50 backbone trained on MS COCO2017. The following observations are made: (i) Both the channel attention and spatial attention of even the student trained for only 1 epoch yield similar results to those of the teacher detector. Additionally, the spatial attention of the student successfully localizes the pixels of foreground objects, such as elephants and zebras. These findings indicate that the attention transfer loss $\mathcal{L}_{\text{AT}}$ effectively minimizes the discrepancy between student attention and teacher attention. (ii) Comparing the student trained for 1 epoch with the one trained for 24 epochs, the latter exhibits significantly less attention difference with the teacher detector. This observation suggests that $\mathcal{L}_{\text{AT}}$ continuously reduces the attention difference between students and teachers throughout the entire training period. This result is consistent with Fig. 14(b), where $\mathcal{L}_{\text{AT}}$ is optimized from 0.8 to 0.1 during training. (iii) Regarding the channel attention of both students and teachers, it is observed that certain channels receive much higher attention while others remain inactive. This disparity signifies that different channels encode distinct semantic information, providing further support for our motivation behind channel-wise masking in knowledge distillation.

## VI. Conclusion

In this paper, we propose two knowledge distillation methods, namely attention-guided distillation and non-local distillation, to improve the performance of object detection models. Attention-guided distillation employs an attention mechanism to identify crucial pixels and channels in the feature map. This allows the student model to focus on these important pixels rather than learning the entire teacher feature map uniformly. Non-local distillation enables students to capture the relational information among pixels using non-local modules. We conduct experiments with twelve knowledge distillation methods and thirteen models to demonstrate the effectiveness of our approach in two-stage, one-stage, anchor-free, and anchor-based models for both object detection and instance segmentation. Furthermore, our analyses of the detection results and error types show that our method significantly improves object detectors in terms of localization and classification abilities. Additionally, experiments on corrupted images demonstrate that our method enhances model robustness and domain generalization ability. We provide detailed ablation studies on different distillation losses and types of attention to highlight the individual effectiveness of each module in our method. Moreover, we investigate the influence of hyperparameters, non-local modules, training time, and adaptation layers through extensive experiments. Furthermore, our study on the relation between students and teachers in object detection reveals that, contrary to previous findings in image classification, a teacher with a higher average precision (AP) is generally a better teacher in object detection. Additionally, we show that knowledge from detectors in different detection paradigms is detrimental to each other, while knowledge from a multi-task teacher is beneficial to a single-task student. To analyze the effectiveness of knowledge distillation, we provide visualizations of the training loss, spatial attention, and channel attention.

## Acknowledgments

Besides, we sincerely appreciate the editors, reviewers, and Yijie Guan for their valuable comments.

## References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.
[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[3] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.
[5] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–8.
[6] T. Zhang et al., "A systematic DNN weight pruning framework using alternating direction method of multipliers," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 184–199.
[7] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, "Rethinking the value of network pruning," in *Proc. Int. Conf. Learn. Representations*, pp. 1–8, 2019.
[8] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–8.
[9] M. Nagel, M. V. Baalen, T. Blankevoort, and M. Welling, "Data-free quantization through weight equalization and bias correction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1325–1334.
[10] A. Zhou, A. Yao, Y. Guo, L. Xu, and Y. Chen, "Incremental network quantization: Towards lossless CNNs with low-precision weights," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–8.
[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.
[12] A. Howard et al., "Searching for MobileNetV3," pp. 1314–1324, 2019.
[13] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 116–131.
[14] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and $<$ 0.5 mb model size," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–8.
[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[16] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2006, pp. 535–541.

[17] Q. Li, S. Jin, and J. Yan, "Mimicking very efficient network for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6356–6364.

[18] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–8.

[19] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[20] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[21] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3588–3597.

[22] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5191–5198.

[23] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4794–4802.

[24] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–8.

[25] Y. Hou, Z. Ma, C. Liu, and C. C. Loy, "Learning lightweight lane detection CNNs by self attention distillation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1013–1021.

[26] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4133–4141.

[27] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9163–9171.

[28] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1921–1930.

[29] L. Zhang, Y. Shi, Z. Shi, K. Ma, and C. Bao, "Task-oriented feature distillation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 14 759–14 771.

[30] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3967–3976.

[31] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1365–1374.

[32] J. Zhu et al., "Complementary relation contrastive distillation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9260–9269.

[33] X. Li, J. Wu, H. Fang, Y. Liao, F. Wang, and C. Qian, "Local correlation consistency for knowledge distillation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 18–33.

[34] G. Xu, Z. Liu, X. Li, and C. C. Loy, "Knowledge distillation meets self-supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 588–604.

[35] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 742–751.

[36] T. Wang, L. Yuan, X. Zhang, and J. Feng, "Distilling object detectors with fine-grained feature imitation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4933–4942.

[37] M. F. Bajestani and Y. Yang, "TKD: Temporal knowledge distillation for active perception," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 953–962.

[38] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2604–2613.

[39] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2051–2062, Apr. 2019.

[40] A. Cheraghian, S. Rahman, P. Fang, S. K. Roy, L. Petersson, and M. Harandi, "Semantic-aware knowledge distillation for few-shot class-incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2534–2543.

[41] X. Hu, K. Tang, C. Miao, X. Hua, and H. Zhang, "Distilling causal effect of data in class-incremental learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3957–3966.

[42] H. Hu, S. Bai, A. Li, J. Cui, and L. Wang, "Dense relation distillation with context-aware aggregation for few-shot object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10 185–10 194.

[43] Y. Hao, Y. Fu, Y.-G. Jiang, and Q. Tian, "An end-to-end architecture for class-incremental object detection with knowledge distillation," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 1–6.

[44] L. Chen, C. Yu, and L. Chen, "A new knowledge distillation for incremental object detection," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–7.

[45] R. Anil, G. Pereyra, A. Passos, R. Ormandi, G. E. Dahl, and G. E. Hinton, "Large scale distributed neural network training through online distillation," 2018, *arXiv: 1804.03235*.

[46] Z. Liu, X. Qi, and C. Fu, "3D-to-2D distillation for indoor scene parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4464–4474.

[47] H. Bagherinezhad, M. Horton, M. Rastegari, and A. Farhadi, "Label refinery: Improving imageNet classification through label progression," 2018, *arXiv:1805.02641*.

[48] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv: 1910.01108*.

[49] C. Xu, W. Zhou, T. Ge, F. Wei, and M. Zhou, "BERT-of-theseus: Compressing BERT by progressive module replacing," 2020, *arXiv: 2002.02925*.

[50] M. Liu, X. Chen, Y. Zhang, Y. Li, and J. M. Rehg, "Attention distillation for learning video representations," in *Proc. Brit. Mach. Vis. Conf.*, 2020, pp. 1–19.

[51] S. Bhardwaj, M. Srinivasan, and M. M. Khapra, "Efficient video classification using fewer frames," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 354–363.

[52] X. Wang, J.-F. Hu, J.-H. Lai, J. Zhang, and W.-S. Zheng, "Progressive teacher-student learning for early action prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3556–3565.

[53] Y. Zhang, H. Chen, X. Chen, Y. Deng, C. Xu, and Y. Wang, "Data-free knowledge distillation for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7852–7861.

[54] Y. Liu, Z. Shu, Y. Li, Z. Lin, F. Perazzi, and S.-Y. Kung, "Content-aware GAN compression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12 156–12 166.

[55] Q. Jin et al., "Teachers do more than teach: Compressing image-to-image models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13 600–13 611.

[56] H. Chen et al., "Distilling portable generative adversarial networks for image translation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 3585–3592.

[57] Z. Li, R. Jiang, and P. Aarabi, "Semantic relation preserving knowledge distillation for image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 648–663.

[58] M. Li, J. Lin, Y. Ding, Z. Liu, J.-Y. Zhu, and S. Han, "GAN compression: Efficient architectures for interactive conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5284–5294.

[59] L. Zhang, Z. Tan, J. Song, J. Chen, C. Bao, and K. Ma, "SCAN: A scalable neural networks framework towards compact and efficient models," 2019, *arXiv: 1906.03951*.

[60] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," 2019, *arXiv: 1905.08094*.

[61] Y. Chen, Y. Xian, A. S. Koepke, Y. Shan, and Z. Akata, "Distilling audio-visual knowledge by compositional contrastive learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7016–7025.

[62] S. Ren, Y. Du, J. Lv, G. Han, and S. He, "Learning from the master: Distilling cross-modal advanced knowledge for lip reading," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13 325–13 333.

[63] L. Wang, J. Huang, Y. Li, K. Xu, Z. Yang, and D. Yu, "Improving weakly supervised visual grounding by contrastive knowledge distillation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14 090–14 100.

[64] L. Zhang, M. Yu, T. Chen, Z. Shi, C. Bao, and K. Ma, "Auxiliary training: Towards accurate and robust models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 372–381.

[65] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 2351–2363.

[66] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 4694–4703.

[67] M. Kang, J. Mun, and B. Han, "Towards oracle knowledge distillation with neural architecture search," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4404–4411.

[68] Y. Liu et al., "Search to distill: Pearls are everywhere but not the eyes," 2019, *arXiv: 1911.09074*.

[69] X. Wang, R. Zhang, Y. Sun, and J. Qi, "KDGAN: Knowledge distillation with generative adversarial networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 775–786.

[70] Z. Zheng, R. Ye, P. Wang, J. Wang, D. Ren, and W. Zuo, "Localization distillation for object detection," 2021, *arXiv:2102.12252*.

[71] Z. Kang, P. Zhang, X. Zhang, J. Sun, and N. Zheng, "Instance-conditional knowledge distillation for object detection," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 16 468–16 480.

[72] X. Dai et al., "General instance distillation for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7842–7851.

[73] J. Guo et al., "Distilling object detectors via decoupled features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2154–2164.

[74] Y. He, G. Kang, X. Dong, Y. Fu, and Y. Yang, "Soft filter pruning for accelerating deep convolutional neural networks," 2018, *arXiv: 1808.06866*.

[75] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1389–1397.

[76] Y. Ro and J. Y. Choi, "Layer-wise pruning and auto-tuning of layer-wise learning rates in fine-tuning of deep networks," 2020, *arXiv: 2002.06048*.

[77] Y. He, P. Liu, Z. Wang, Z. Hu, and Y. Yang, "Filter pruning via geometric median for deep convolutional neural networks acceleration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4340–4349.

[78] Z. Liu et al., "MetaPruning: Meta learning for automatic neural network channel pruning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3296–3305.

[79] Y. He, J. Lin, Z. Liu, H. Wang, L.-J. Li, and S. Han, "AMC: Automl for model compression and acceleration on mobile devices," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 784–800.

[80] J. Choi, Z. Wang, S. Venkataramani, P. I.-J. Chuang, V. Srinivasan, and K. Gopalakrishnan, "PACT: Parameterized clipping activation for quantized neural networks," 2018, *arXiv: 1805.06085*.

[81] M. Courbariaux, Y. Bengio, and J.-P. David, "BinaryConnect: Training deep neural networks with binary weights during propagations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 3123–3131.

[82] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet classification using binary convolutional neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 525–542.

[83] X. Wang, F. Yu, Z.-Y. Dou, T. Darrell, and J. E. Gonzalez, "SkipNet: Learning dynamic routing in convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 409–424.

[84] Z. Wu et al., "BlockDrop: Dynamic inference paths in residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8817–8826.

[85] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger, "Multi-scale dense networks for resource efficient image classification," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–8.

[86] J. Yu, L. Yang, N. Xu, J. Yang, and T. Huang, "Slimmable neural networks," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–8.

[87] J. Yu and T. S. Huang, "Universally slimmable networks and improved training techniques," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1803–1811.

[88] Y. Wang, R. Huang, S. Song, Z. Huang, and G. Huang, "Not all images are worth 16x16 words: Dynamic vision transformers with adaptive sequence length," 2021, *arXiv:2105.15075*.

[89] Y. Wang, Z. Chen, H. Jiang, S. Song, Y. Han, and G. Huang, "Adaptive focus for efficient video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 16 249–16 258.

[90] Y. Wang, K. Lv, R. Huang, S. Song, L. Yang, and G. Huang, "Glance and focus: A dynamic approach to reducing spatial redundancy in image classification," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020, pp. 2432–2444.

[91] Z. Xie, Z. Zhang, X. Zhu, G. Huang, and S. Lin, "Spatially adaptive inference with stochastic feature sampling and interpolation," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, UK, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., 2020, pp. 531–548.

[92] L. Yang, Y. Han, X. Chen, S. Song, J. Dai, and G. Huang, "Resolution adaptive networks for efficient inference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 2366–2375.

[93] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1–8.

[94] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

[95] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[96] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," 2021, *arXiv:2104.00298*.

[97] R. Li, Y. Wang, F. Liang, H. Qin, J. Yan, and R. Fan, "Fully quantized network for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2810–2819.

[98] Z. Xie, L. Zhu, L. Zhao, B. Tao, L. Liu, and W. Tao, "Localization-aware channel pruning for object detection," *Neurocomputing*, vol. 403, pp. 400–408, 2020.

[99] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[100] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.

[101] J. Redmon and A. Farhad, "YOLOv3: An incremental improvement," 2018, *arXiv: 1804.02767*.

[102] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9157–9166.

[103] Z. Qin et al., "ThunderNet: Towards real-time generic object detection on mobile devices," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6718–6727.

[104] M. Tan, R. Pang, and Q. Le, "EfficientDet: Scalable and efficient object detection," 2019, *arXiv: 1911.09070*.

[105] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13 039–13 048.

[106] A. Buades, B. Coll, and J. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, 2005, pp. 60–65.

[107] X. Yan, C. Zheng, Z. Li, S. Wang, and S. Cui, "PointASNL: Robust point clouds processing using nonlocal neural networks with adaptive sampling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5589–5598.

[108] S. Lefkimmiatis, "Non-local color image denoising with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3587–3596.

[109] D. Liu, B. Wen, Y. Fan, C. C. Loy, and T. S. Huang, "Non-local recurrent network for image restoration," 2018, *arXiv: 1806.02919*.

[110] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," 2019, *arXiv: 1903.10082*.

[111] S. Zhou, J. Zhang, W. Zuo, and C. C. Loy, "Cross-scale internal graph neural network for image super-resolution," 2020, *arXiv: 2006.16673*.

[112] P. Yi, Z. Wang, K. Jiang, J. Jiang, and J. Ma, "Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3106–3115.

[113] Y. Xu, L. Gao, K. Tian, S. Zhou, and H. Sun, "Non-local ConvLSTM for video compression artifact reduction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7043–7052.

[114] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 603–612.

[115] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 1–8.

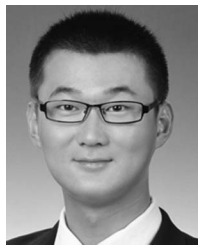
[116] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 593–602.

[117] Y. Li et al., "Neural architecture search for lightweight non-local networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10 297–10 306.

[118] D. Walawalkar, Z. Shen, and M. Savvides, "Online ensemble model compression using knowledge distillation," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., 2020, pp. 18–35.

[119] C. Bian, W. Feng, L. Wan, and S. Wang, "Structural knowledge distillation for efficient skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 2963–2976, 2021.
[120] Z. Yang et al., "Focal and global knowledge distillation for detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4643–4652.
[121] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[122] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
[123] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[124] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.
[125] C. Michaelis et al., "Benchmarking robustness in object detection: Autonomous driving when winter is coming," 2019, *arXiv: 1907.07484*.
[126] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, May 2021, doi: 10.1109/tpami.2019.2956516.
[127] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic R-CNN: Towards high quality object detection via dynamic training," 2020, *arXiv: 2004.06002*.
[128] X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan, "Grid R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7363–7372.
[129] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[130] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 840–849.
[131] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9657–9666.
[132] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6409–6418.
[133] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10 012–10 022.
[134] R. Sun, F. Tang, X. Zhang, H. Xiong, and Q. Tian, "Distilling object detectors with task adaptive regularization," 2020, *arXiv: 2006.13108*.
[135] Z. Du et al., "Distilling object detectors with feature richness," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 5213–5224.
[136] P. Zhang, Z. Kang, T. Yang, X. Zhang, N. Zheng, and J. Sun, "LGD: Label-guided self-distillation for object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 3309–3317.
[137] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–8.
[138] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[139] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10425–10433.
[140] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[141] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5987–5995.
[142] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–8.
[143] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.
[144] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv: 1906.07155*.
[145] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–8.
[146] J. Wang, C. Lan, C. Liu, Y. Ouyang, and T. Qin, "Generalizing to unseen domains: A survey on domain generalization," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Montreal, Canada, 2021, pp. 4627–4635, doi: 10.24963/ijcai.2021/628.
[147] H. Mobahi, M. Farajtabar, and P. Bartlett, "Self-distillation amplifies regularization in Hilbert space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 3351–3361.
[148] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisit knowledge distillation: A teacher-free framework," 2019, *arXiv: 1909.11723*.

**Linfeng Zhang** is currently working toward the PhD degree with the Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University supervised by Kaisheng Ma. His research interests includes computer vision, deep neural network compression, and acceleration. He is also the reviewer in ICCV, CVPR, NeurIPS, ICLR, TIP, IJCAI, and so on.



**Kaisheng Ma** received the PhD degree from the Department of Computer Science and Engineering, Pennsylvania State University, following Dr. Vijaykrishnan Narayanan, Dr. Jack Sampson in PSU, and Dr. Yuan Xie in UCSB. He is an assistant professor of computer science with the Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University.