ScaleKD: Distilling Scale-Aware Knowledge in Small Object Detector

Yichen Zhu[1], Qiqi Zhou[2,1], Ning Liu[1], Zhiyuan Xu[1],
Zhicai Ou[1], Xiaofeng Mou[1], Jian Tang[1,*]
[1]Midea Group,
[2]Shanghai University of Electric Power
{zhuyc25, zhouqq31, lining22, xuzy70, zhicai.ou, mouxf, jiantang22}@midea.com

Abstract

Despite the prominent success of general object detection, the performance and efficiency of Small Object Detection (SOD) are still unsatisfactory. Unlike existing works that struggle to balance the trade-off between inference speed and SOD performance, in this paper, we propose a novel Scale-aware Knowledge Distillation (ScaleKD), which transfers knowledge of a complex teacher model to a compact student model. We design two novel modules to boost the quality of knowledge transfer in distillation for SOD: 1) a scale-decoupled feature distillation module that disentangled teacher's feature representation into multi-scale embedding that enables explicit feature mimicking of the student model on small objects. 2) a cross-scale assistant to refine the noisy and uninformative bounding boxes prediction student models, which can mislead the student model and impair the efficacy of knowledge distillation. A multi-scale cross-attention layer is established to capture the multi-scale semantic information to improve the student model. We conduct experiments on COCO and VisDrone datasets with diverse types of models, i.e., two-stage and one-stage detectors, to evaluate our proposed method. Our ScaleKD achieves superior performance on general detection performance and obtains spectacular improvement regarding the SOD performance.

1. Introduction

Object detection is a fundamental task that has been developed over the past twenty-year in the computer vision community. Despite the state-of-the-art performance for general object detection having been conspicuously improved since the rise of deep learning, balancing the complexity-precision for small object detection is still an open question. Current works strive to refine feature fusion modules [9, 21], devise novel training schemes [32, 33] to explicitly train on small objects, design new neural architectures [20, 39] to better extract small objects' features, and leverage increased input resolution to enhance representation quality [1, 49]. However, these approaches struggle to balance detection quality on small objects with computational costs at the inference stage.

The above reasons incentivize us to design a cost-free technique at test time to improve SOD performance. In the spirit of the eminent success of knowledge distillation (KD) on image data [14], we explore distillation methods for SOD. Typically, knowledge distillation opts for a complex, high-performance model (teacher) that transfers its knowledge to a compact, low-performance model (student). The student model can harness instructive information to enhance its representation learning ability. Nevertheless, unlocking this potential in SOD involves overcoming two challenges: 1) SOD usually suffers from noisy feature representations. Due to the nature of small objects, which generally take over a small region in the whole image, the feature representations of these small objects can be contaminated by the background and other instances with relatively larger sizes. 2) Object detectors have a low tolerance for noisy bounding boxes on small objects. It is inevitable that teacher models make incorrect predictions. Usually, student models can extract informative dark knowledge [14, 28] from imperfect predictions from the teacher. However, in SOD, small perturbations on the teacher's bounding box can dramatically impair SOD performance on the student detector (§3.2).

To this end, we propose Scale-aware Knowledge Distillation for small object detection (ScaleKD). Our proposed ScaleKD consists of two modules, a Scale-Decoupled Feature (SDF) distillation module and a Cross-Scale Assistant (CSA), to address the aforementioned two challenges correspondingly. The SDF is inspired by the crucial shortcoming of existing feature distillation methods, where the feature representations of objects with varying scales are coupled in a single embedding. It poses difficulty for the student to mimic small objects' features from the teacher model.

[*] Corresponding author

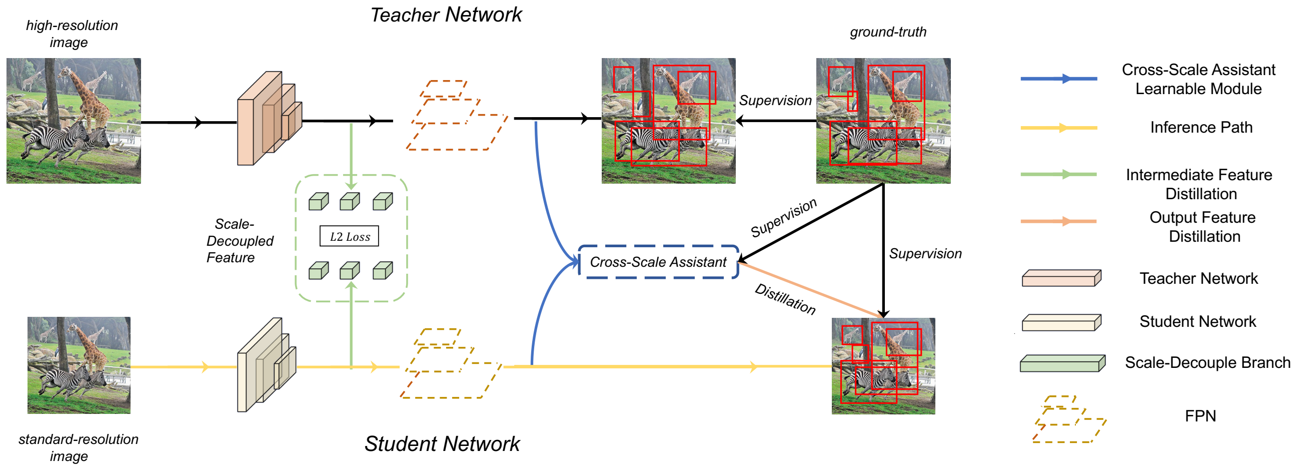

Figure 1. The overview of Scale-aware Knowledge Distillation. It consists of a Scale-Decoupled Feature distillation module and a Cross-Scale Assistant module to improve small object detection.

As a result, the proposed SDF aims to decouple a single-scale feature embedding into a multi-scale feature embedding. The multi-scale embedding is obtained by a parallel multi-branch convolutional block, where each branch deals with one scale. Our SDF allows the student model to better understand the feature knowledge from the perspective of object scale. Furthermore, we propose a learnable CSA to resolve the adverse effect of teachers' noisy bounding box prediction on small objects. The CSA comprises a multi-scale cross-attention module, where representations from the teacher and student models are mapped into a single feature embedding. The multi-scale query-key pair projects the teacher's features into multiple sizes, such that the fine-grained and low-level details can be preserved in CSA, which helps to produce suitable bounding box supervision for the student model.

We demonstrate the effectiveness of our approach on COCO object detection and VisDrone datasets. The experiments are conducted on multiple types of detectors, including two-stage detectors, anchor-based detectors, and anchor-free detectors, and have proven the generalizability of our approach. Our work offers a practical approach for industrial application on SOD as well as introduces a new perspective on designing scale-aware KD modules to improve object detectors. We further extend our method on instance-level detection tasks, such as instance segmentation and keypoint detection, demonstrating the superiority of our approach to dealing with small objects in vision tasks.

In summary, our contributions are the following:

- We propose Scale-Aware Knowledge Distillation (ScaleKD), a novel knowledge distillation framework to improve general detection and SOD performance without bringing extra computational costs at test time.
- Our proposed ScaleKD not only exceeds state-of-the-art KD for object detection methods on general detection performance but also surpasses existing approaches on SOD by a large margin. Extended experiments on instance segmentation and keypoint detection further strength our method.

2. Related Work

Knowledge Distillation. Knowledge distillation has become one of the most effective techniques for model compression [14]. It first trains a cumbersome teacher model and then transfers its knowledge to a lightweight student model. The common knowledge distillation approaches include distillation on output, logits [14], bounding box [16], and features [30, 52, 53].

Our work is also closely correlated to object detection, where knowledge distillation has shown effectiveness in detection tasks [3, 15, 35, 40, 47, 48]. In particular, Fine-Grained [5] is the first work to comprehensively present a KD framework for the object detection task. FKD [44] is the earliest work to adopt attention mechanisms to do feature distillation on object detectors. DeFeat [12] decoupled the foreground feature and background features based on the ground-truth binary mask and performed feature distillation on two features separately. FGFI [41] presents focal distillation, which combines DeFeat and Zhang et al. [44] works. ICD [17] considered conditional distillation of both classification and localization on very instances. Guo [12] investigates distilling image classifier instead of object detector to student detector, which complements existing feature-based distillation methods. Similar to ICD, GID [7] proposed an instance-selection module to transfer the teacher's most informative locations. LGD [45] proposed a teacher-free framework to distill object detectors without a concrete teacher model. Nevertheless, existing methods have largely overlooked the small objects. Our paper is the first work to design knowledge distillation specifically targeted to improve general object detection and small objects' performance.

Small Object Detection. Recognizing small objects in the image is challenging, especially when localization and per-pixel classification are required. A naive approach is to increase the resolution [26]. Nevertheless, it brings tremendous computational costs in inference. Existing approaches mainly focus on applying strong data augmentation [18] or oversampling [4], incorporating context information [9, 21], fusing the features across layers [10, 38], performing scale-aware training [32, 33], adjusting resolution at test-time [37], or applying dilated convolution/large convolution [20, 39] to increase the receptive field. We refer interested readers for a more in-depth survey [6].

Our approach is very different from previous methods, where we seek help with knowledge distillation to boost the performance of SOD. Our method benefits from the advantage of the distillation method, where 1) no extra computational costs are introduced at test time, and 2) neural architectures of the existing model do not need to be modified, a critical benefit in the practical scenario where engineering do not want to alter their networks architecture due to inconvenience on deployment.

3. Scale-aware Knowledge Distillation

This section provides a detailed description of our proposed distillation methods. Figure 1 gives a brief overview of ScaledKD, in which we illustrate the modules as follows: (1) A scale-decoupled feature distillation module that explicitly transfers representation on diverse scales to the student detector. (2) A cross-scale assistant refines the knowledge of object size between complex teacher and compact student.

Definition. In this section, we provide a notion of the components that we will use to describe our method. Considering an object detector $\mathcal{G}^S : \mathbb{R}^d \rightarrow \mathbb{R}^k$ as a student and another predictor $\mathcal{G}^T : \mathbb{R}^d \rightarrow \mathbb{R}^k$ as a teacher, where d and k are two feature dimensions. The former is a computationally efficient network with a relatively lower detection performance. The latter is a heavy model with relatively higher detection performance. Giving a training dataset $S = \{(x_i, y_i)_{i=1}^n\} \sim \mathbb{P}^n$ for distribution $\mathbb{P}$ over a set of instance $\mathcal{X}$. For small object detection, a common approach to improve SOD is to leverage a high-resolution image as input. In our case, we only consider using high-resolution images $\mathcal{X}^{hr}$ for the teacher model, while standard-resolution images are used for the student model.

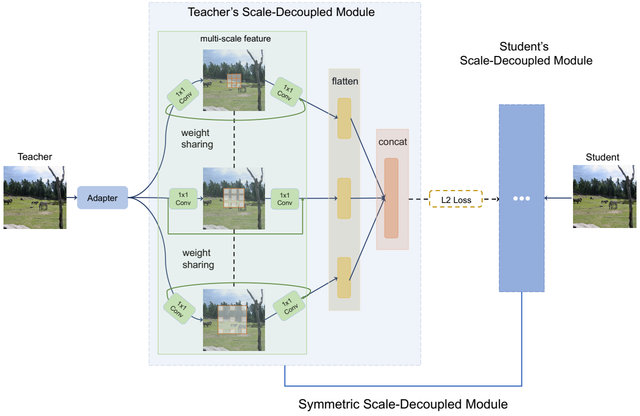

Figure 2. The scale-decouple feature module in ScaleKD. Note that the details of the student's scale-decoupled feature module are omitted due to limited space. In practice, it is symmetric to the teacher's counterpart.

3.1. Scale-Decoupled Feature Distillation

Preliminary. Modern object detectors harness FPN [21] to obtain the multi-scale semantic information from different levels of the backbone to enhance the model's feature learning ability on diverse object scales. A typical distillation approach is to transfer the knowledge of these features from teacher to student. Generally, we can formulate such distillation methods as:

$$\mathcal{L}_{feat} = l(\mathcal{F}^T, f(\mathcal{F}^S)) \tag{1}$$

where F^T and F^S are corresponding feature layers in the teacher and student model. $f(\cdot)$ is a mapping function to align the dimension of the feature map in student to teacher, and $l(\cdot)$ is any bounded loss, i.e., l_2 norm distance.

Motivation. Although state-of-the-art KDs for object detection have developed various techniques, such as focal distillation [41], to enhance the quality of transferred knowledge, such methods treat all objects that have varying sizes equally during distillation. We identify that the key shortcoming of these methods on SOD is that the feature representation on diverse objects is entangled. It is a viable solution when object sizes are not small. However, the size of objects in the detection task varies significantly. The feature representation of small objects can be contaminated by large-region backgrounds and other instances with a relatively larger size. To resolve the issue, we propose a Scale-Decoupled Feature (SDF) distillation module to explicitly disentangled the teacher's feature map.

Methodology. As aforementioned, our goal is to disentangle the whole feature representation of the teacher

Method	Perturbation	AP	AP$_S$	AP$_M$	AP$_L$
Baseline	0	36.9	21.2	40.8	48.4
Chen et al. [5]	0	38.1	21.9	42.2	50.1
	6 pixels	37.7	21.1	42.4	50.2
	12 pixels	37.0	19.9	42.1	50.0

Table 1. Noisy and uninformative bounding boxes from the teacher can severely hurt students' performance. The perturbation is measured by a number of pixels, and it is added along the diagonal direction.

object detector into multiple parts, where each part only deals with similar object sizes. Presumably, such operations can force student detectors to comprehend not only the global knowledge of the entire image but also the scale-specific knowledge.

To be specific, we obtain a feature embedding Z^T and Z^S in the last stage of the backbone for both the teacher and student networks. We intend to fully utilize the feature representation on diverse input scales. Thus, we adopt a multi-branch structure, where each branch uses a convolutional layer with different dilated rates. It is worth noting that the model tends to focus on small objects for the kernel with a small dilated rate and vice versa. In practice, we use residual blocks as in ResNet, which consist of three consecutive convolutional layers, i.e., 1x1, 3x3, and 1x1, along with a skip connection. For the 3x3 Conv layer, we use three different dilated rates $\{1, 2, 3\}$ on three individual branches. An adapter layer, typically a multi-layer perception (MLP) layer, is stacked before this scale-decoupled feature module to align the dimensionality of feature presentation between the teacher and student model.

Usually, one can match the knowledge of the teacher model at a designated branch, i.e., the 3x3 Conv layer with a dilated rate of 1, to the corresponding feature branch in the student model via any distance minimization loss. One drawback is that these operations can be memory intensive. Therefore, we take inspiration from the weight-sharing network in neural architecture search [25, 34] and adopt a weight-sharing scale-decoupled feature. This is based on the fact that all three branches have the same operators. In practice, we only save a set of weights for three branches, dramatically reducing the training memory cost.

We also notice that using three individual losses to match three parallel branches between the teacher and student model can cause the practicer to spend unnecessary effort on hyper-parameter tuning. As a result, for each branch, we use a flattened layer, namely a multi-layer perceptron, and concatenate three flattened layers together. During the distillation process, we adopt a sole l2 loss (denoted as L_{feat} in the later section) to minimize the distance of this concatenated flattened layer between the teacher and student model. the overview of the proposed module can be viewed in Figure 2. Notably, the student's scale-decoupled feature module mirrors the teacher's counterpart.

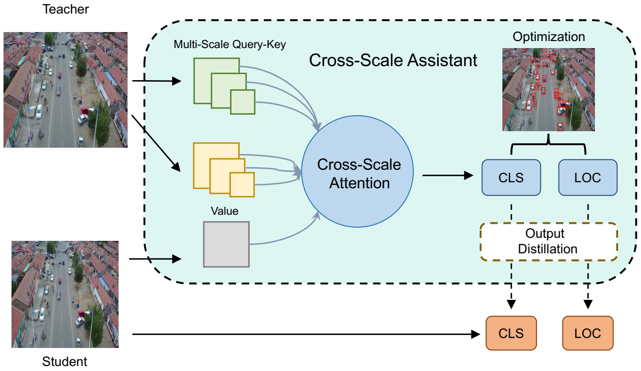

Figure 3. The Cross-Scale Assistant (CSA) module. We adopt a multi-scale query-key pair to perform cross-attention with feature embedding in the student model. The learnable weights in CSA are updated via two separate classification and regression branches which are supervised by the ground truth label.

3.2. Cross-Scale Assistant

Preliminary. Our previous section introduces our proposed feature distillation to resolve the challenge of SOD. Other than feature distillation, another practical approach is the output-based KD, which transfers the teacher's prediction on classification and localization to the student as sources for auxiliary supervision. Our work primarily focuses on bounding box distillation, which can be considered a regression problem. In general, we can write such regression distillation as follow:

$$\mathcal{L}_{bbox} = l_{bbox}(\mathcal{R}^S, \mathcal{R}^T) \tag{2}$$

where $\mathcal{R}^S$ is the regression output of the student network, $\mathcal{R}^T$ is the predication of teacher network. This $l(\cdot)$ is the same as Equation 1, where any bounded loss can be applied. Conventionally, this bounded loss can be either l_1, smoothed l_1 or l_2 loss, depending on the level of penalty that we wise to weight the error between the student's predication and the teacher's output.

Motivation. A key difference between small object detection and general object detection is that SOD is sensitive to noisy bounding boxes. Apparently, the teacher detector is incapable of making perfect predictions on every object. In general object detection, despite being inaccurate, student models can still retrieve informative knowledge from teachers' predictions on bounding boxes. Nevertheless, for small objects, noisy bounding box prediction in the teacher model can confuse the student model, which decreases the SOD performance. As proof of concept, we deliberately add a slight deviation (6 and 12 pixels along the diagonal direction) and compare the mAP on small-scale objects of a vanilla teacher with a teacher that is slightly perturbed. As shown in Table 1, on RetinaNet baseline, the AP of small objects (AP_S) of students drops consistently when perturbations are added, showing the sensitivity of student detector on teacher's bounding box prediction in SOD. Therefore, to build a trustworthy regression distillation module for SOD, one needs to minimize the adverse effect of poison bounding box knowledge in teachers. This does not mean that the teacher's predictions have to be perfect - otherwise, we can directly supervise the student with ground truth - we only need to refine the teacher's output to ensure that their knowledge is informative to the student [27, 50].

Methodology. To address the aforementioned issue, we propose a Cross-Scale Assistant (CSA), which refines the teacher's knowledge and enables the student model to fetch instructive knowledge on objects with different scales.

Our method is simple - we establish the CSA by a cross-attention module. During the cross-attention, a sequence of key and query tokens are generated in calculating KQ-attention within the teacher's knowledge and then mapped with the value tensor, the outputs from the student model, to obtain attentive regions in the feature with each corresponding query. This process is executed at every student pyramid scale to retrieve informative region-based features.

A naive choice is to use plain cross-attention [8]. Albeit, previous studies [43] identified that the standard cross-attention could focus on salient regions repetitively on different heads. Consequently, when large objects appear in the image, the cross-attention will redirect its attention to these large objects and ignore small objects. To this end, in contrast to the plain cross-attention, we developed a multi-scale cross-attention layer, as shown in Figure 3. Notice that cross-attention extracts global information - for each query-key pair, a value is generated to highlight the most responsive region. We then split the query-key pair into multiple sub-pairs, where each sub-pair represents a set of object scales. As a result, our multi-scale query-key can enforce the attention module to focus on regions with diverse scales, such that all objects, especially the small objects, can attend to the feature learning process.

Particularly, give an input sequence from teacher $F^T \in \mathbb{R}^{h\times w\times c}$ and another input sequence from student $F^S \in \mathbb{R}^{h\times w\times c}$, for simplicity, we assume two tensors have the same size. The F^T is projected into a query (Q) and key (K), and F^S is projected into value (V). The keys K and values V are down-sampled to different sizes for different heads indexed by i. Thus, we formulate our multi-scale cross-attention (MSC) as follows:

$$Q_i = F^S W_i^Q \tag{3}$$

$$K_i = MSC(F^T, r_i)W_i^K, V_i = MSC(F^T, r_i)W_i^V, \tag{4}$$

$$V_i = V_i + P(V_i) \tag{5}$$

where $MSC(\cdot, r_i)$ is a MLP layer for aggregation in the $i-$th head with the down-sampling rate of r_i, and $P(\cdot)$ is a depth-wise convolutional layer for projection. Compared with the standard cross-attention, more fine-grained and low-level details that are beneficial to SOD are preserved. Finally, we calculated the attention tensor by:

$$h_i = Softmax(\frac{Q_i K_i^T}{\sqrt{d_h}} V_i) \tag{6}$$

where d_h is the dimension. Remember that the purpose of CSA is to bridge the cross-scale information between the teacher and student model to refine the bounding box supervision in KD. Therefore, we stacked head layers with the classification branch and regression branch to update the weights of these learnable modules. Our empirical analysis shows that the proposed CSA can provide more appropriate supervision on small objects bounding boxes during distillation. Because this is a learnable module, during distillation, we perform a weights update prior to the student model at each iteration through a standard training scheme and detection objectives.

In distillation, instead of transferring the teacher's output knowledge, we transfer the CSA's knowledge on both classification and regression branches to students. For the output-based distillation objectives, we follow [5] to have two loss functions L_{cls} and L_{bbox}. We note that our method is also complementary to other output-based methods, such as LD [46], where all we need to do is simply replace the distillation objective.

In summary, the total training objective for the student model is:

$$L_{total} = \alpha L_{feat} + \beta L_{cls} + \gamma L_{bbox} + L_{det} \tag{7}$$

where L_{det} is the standard training loss for the detector.

Besides the distillation loss and detection loss for optimizing student detectors, we further ensure the instructive representation quality and consistency with student representations by sharing the detection head for supervision. Moreover, the CSA combines the feature of both teacher and student. As a result, a randomly initialized student detector can cause unstable training of CSA. Thus, we first warm up the student model for 30k iterations since it could be detrimental when the instructive knowledge is optimized insufficiently. The student detector backbone is frozen in early 10k iterations under 1× training schedule and 20k for 2× training schedule.

Method	Backbone	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$	FPS	Params (M)
Faster-RCNN [29]	ResNet50	38.4	59.0	42.0	21.5	42.1	50.3	20.1	43.57
+ ScaleKD		**43.1**	**64.0**	**48.2**	**26.1**	**48.0**	**58.8**		
Faster-RCNN [29]	MobileNetV2	27.3	44.7	28.9	14.6	29.6	35.7	12.9	32.61
+ ScaleKD		**30.8**	**48.7**	**32.9**	**18.8**	**34.2**	**42.2**		
Cascade-RCNN [2]	ResNet50	41.0	59.4	44.4	22.7	44.4	54.3	18.9	71.22
+ ScaleKD		**44.9**	**63.2**	**48.8**	**25.7**	**48.1**	**59.2**		
RetinaNet [22]	ResNet50	37.4	56.7	39.6	20.0	40.7	49.7	17.4	37.74
+ ScaleKD		**42.8**	**61.1**	**44.6**	**25.9**	**45.4**	**54.9**		
FOCS [36]	ResNet50	38.5	57.7	41.0	21.9	42.8	48.6	23.9	36.15
+ ScaleKD		**44.0**	**61.6**	**41.6**	**29.1**	**46.9**	**56.2**		
RepPoints [42]	ResNet50	38.6	59.6	41.6	22.5	42.2	50.4	18.2	36.62
+ ScaleKD		**42.9**	**63.8**	**46.8**	**25.7**	**47.1**	**56.7**		

Table 2. Comparison with state-of-the-art methods KD-Det on COCO val2017 datasets. Our approach appear to improve the detection performance drastically, especially on AP_S. The FPS is evaluated on a single V100 GPU.

Student	Method	AP	AP$_{50}$	AP$_{75}$	AR$_1$	AR$_{10}$	AR$_{100}$	AR$_{500}$
RetinaNet-ResNet50	Baseline	26.21	44.90	27.10	0.52	5.35	34.63	37.21
	ZoomInNet [24]	27.11	46.02	28.09	0.53	5.47	34.98	37.66
	ScaleKD	**29.45**	**49.28**	**29.97**	**0.53**	**5.98**	**36.74**	**38.65**
RetinaNet-MobileNetV2	Baseline	21.73	40.01	22.95	0.49	5.11	31.24	34.05
	ZoomInNet [24]	22.49	41.14	24.21	0.50	5.15	31.78	34.91
	ScaleKD	**26.08**	**44.76**	**26.98**	**0.52**	**5.33**	**34.48**	**37.03**

Table 3. Experimental results on VisDrone. We use ResNet101 as the teacher model.

4. Experiment

We conduct quantitative analysis on two object detection datasets: COCO [23] and VisDrone [51]. COCO is a challenging object detection dataset with 80 categories. VisDrone is a dataset specialized to drone-shot image detection, in which most of the objects in this dataset are small.

4.1. Implementation Details

For COCO, we adopt the standard 1× schedule and kept all the hyperparameters for the student model training unchanged for fair comparisons. The models are evaluated on validation set 2017. On the VisDrone dataset, we follow the which equally split one image into four non-overlapping patches and process them independently during training. The training procedure is also kept the same as in previous literature. We use 2x of standard input resolution for all experiments for the teacher model, useless otherwise indicated. Mean average precision (AP) is used as the major metric for all experiments. We use an Adam optimizer with an initial learning rate of 3e-4 and a weight-decay of 1e-4 to update the CSA module. There are three hyper-parameters in the total training objectives, we set $\alpha = 0.07$, $\beta = 0.5$, $\gamma = 0.2$ for all two-stage models, and $\alpha = 0.01$, $\beta = 0.2$, $\gamma = 0.05$ for all one-stage models.

4.2. Main Results

In this section, we show the experimental results of the baseline detectors and our method on COCO and VisDrone datasets. We compare our approach with state-of-the-art distillation methods in the next section.

COCO. We conduct extensive experiments on COCO to validate the effectiveness of ScaleKD. We adopt our method on multiple mainstream baselines, including two-stage detection models (i.e., Faster RCNN [29], Cascade RCNN [2]), and one-stage detection models (i.e., RetinaNet [22], FCOS [36], Reppoints [42]). For all experiments, we use the ResNet101 backbone as a teacher model. The ScaleKD achieves noticeable improvement on all methods without introducing any extra computational cost at inference. It is worth noting that the performance boost on AP_S is overwhelming. For instance, on FCOS the AP_S is increased by 7.2, and on RetinaNet the AP_S is increased by 5.9. Also, our approach is not only effective for SOD but improves the AP in general.

Task	Method	Type	Resolution	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}_{S}	AP^{mask}_{M}	AP^{mask}_{L}
Instance Segmentation	Mask-RCNN	Teacher	High	35.4	56.5	37.9	19.2	38.6	48.4
		Baseline	Low	31.6	51.2	33.4	10.8	33.6	50.9
		Ours	Low	$33.8_{+2.2}$	$53.2_{+2.0}$	$35.7_{+2.3}$	$13.1_{+2.3}$	$35.5_{+1.9}$	$52.3_{+2.4}$
Keypoint Detection	Mask-RCNN	Teacher	High	66.4	87.1	72.8	63.5	72.0	73.6
		Baseline	Low	62.5	86.1	68.2	56.5	72.9	70.1
		Ours	Low	$64.1_{+1.6}$	$87.6_{+1.5}$	$70.1_{+1.9}$	$58.7_{+2.2}$	$74.2_{+1.7}$	$71.2_{+1.1}$

Table 4. Performance comparisons with baselines on instance segmentation and keypoint detection. The high resolution means 800px, and the low resolution represents 400px, both on the short side. The baseline refers to the standard training.

Method	AP	AP$_S$	ΔAP$_S$	AP$_M$	ΔAP$_M$	AP$_L$	ΔAP$_L$
GFL ResNet50							
Baseline	40.1	23.3	-	44.4	-	52.5	-
FitNet [30]	40.7	23.7	+0.4	44.4	+0.0	53.2	+0.7
OD [5]	41.1	23.3	+0.0	45.4	+1.0	53.1	+0.6
DeFeat [11]	40.8	24.3	+1.0	44.6	+0.2	53.7	+0.5
GID [7]	41.5	24.3	+1.0	45.7	+1.3	53.6	+1.1
LD [46]	42.1	24.5	+1.2	46.2	+1.8	54.8	+2.3
ScaleKD	**42.5**	**25.9**	**+2.6**	**46.2**	**+1.8**	54.6	+2.1
RetinaNet ResNet50							
Baseline	37.4	20.0	-	40.7	-	49.7	-
FKD [44]	39.6	21.4	+1.4	42.5	+1.8	51.5	+1.8
LGD [45]	38.3	23.2	+3.2	42.0	+1.3	50.0	+ 0.3
FGFI [41]	38.6	21.4	+1.4	42.5	+1.8	51.5	+1.8
GID [7]	39.1	22.8	+2.8	43.1	+2.4	52.3	+2.4
CLSD [12]	40.7	23.1	+3.1	44.7	+4.0	53.8	+4.1
ScaleKD	**41.7**	**24.8**	**+4.8**	**44.7**	**+4.0**	**53.9**	**+4.2**
FCOS ResNet50							
Baseline	38.5	21.9	-	42.8	-	49.7	-
LD [46]	40.4	23.7	+1.8	44.3	+1.5	52.2	+2.5
GID [7]	42.0	25.6	+3.7	45.8	+3.0	54.2	+4.5
FGFI [41]	42.1	27.0	+5.1	**46.0**	**+3.2**	54.6	+4.9
ScaleKD	**42.7**	**27.8**	**+5.9**	45.7	+2.9	**54.7**	**+5.0**

Table 5. Comparison with state-of-the-art methods on COCO val2017. Our ScaleKD achieves pronounced improvement on AP_S compared to SOTA.

VisDrone. We demonstrate the experimental results on the VisDrone dataset in Table 3. We use ResNet-101 as the backbone for both experiments. We can clearly observe that our approach significantly improves the performance of the student model. Notably, after applying ScaleKD to RetinaNet-MobileNetV2 [22, 31], we achieve detection performance that is on par with the RetinaNet-ResNet50 (26.08 AP versus 26.21 AP), which adopt ResNet50, an 8x heavier backbone than MobileNetV2. Our method also increases both AP for RetinaNet-ResNet-50 [13]. Since VisDrone is a dataset that contains mostly small objects, these results support our claim that the proposed ScaleKD is extremely effective on SOD.

We also make a comparison with ZoomInNet [24], which combines feature level alignment and layer adaptation to distill a standard teacher to small object detectors. We evaluate ZoomInNet on both settings, similar to our approach. Indeed, ZoomInNet can slightly improve the AP on VisDrone, but the improvement is very marginal. In contrast, the ScaleKD achieve drastically higher results than their approach.

Results on Other Instance-Level Tasks with Small Input Resolution. Input resolution is a critical factor in achieving good performance on SOD. Typically, reduced input resolution can result in a considerable performance drop. Here we verify if our proposed ScaleKD can improve the SOD performance on instance-level tasks, even when the input resolution of the student model is low. The evaluation results are reported in Table 4. We notice that the low-resolution models trained with our approach outperform the baseline by a large margin, on both instance segmentation and keypoint detection, based on Mask-RCNN. This shows our ScaleKD is particularly good at predicting small objects, even when the number of effective pixels is small.

4.3. Ablation Study

Comparison with State-of-the-art knowledge distillation. In the last section, we validate the effectiveness of our proposed ScaleKD on two datasets compared to the train-from-scratch counterpart. This section compares ScaleKD to state-of-the-art KD methods on the COCO benchmark. For a fair comparison, we use the *standard image resolution on teacher* and the same teacher architecture, such that the teacher's performance is the same for all approaches. We compare ScaleKD to these SOTA methods on Generalized focal loss (GFL) [19], RetinaNet [22], and FCOS [36], respectively. All student models use ResNet50 as the backbone. We compare eight distillation methods, seven of them are KD methods that are designed for object detec-

Method	CSA	MSC	AP	AP_S	AP_M	AP_L
Baseline	-	-	36.9	21.2	40.8	48.4
ScaleKD	✓	✗	40.2	23.5	43.7	52.6
	✓	✓	41.6	24.8	44.5	53.9

Table 6. Ablation study for Cross-Scale Assistant. MSC: Multi-scale Cross-Attention.

tion, including FitNet [30], OD [5], LGD [45], DeFeat [11], GID [7], LD [46], FGFI [41], and CLSD [12].

Table 5 presents the experimental results. We can observe that ScaleKD achieves superior performance on most evaluation metrics, i.e., on GFL-ResNet50, ScaleKD obtained a 2.8 higher AP score compared with baseline and 0.7 higher AP score compared to second best SOTA method (LD [46]). Furthermore, as we look at the AP_S, the one that measures AP on small objects, we notice that our proposed method obtains a significantly better score than SOTA. For instance, ScaleKD improves AP_S by 3.4, 4.8, and 5.9 on GFL, RetinaNet, and FCOS, respectively. These results indicate that ScaleKD helps compact student detectors on detecting small objects, and it is much more effective than KD methods designed for general object detection.

Impact of Cross-Scale Assistant. We investigate the effectiveness of our proposed Cross-Scale Assistant module. We use standard image resolution for teachers in this section. We are primarily eager to know if CSA improves SOD performance and does the multi-scale cross-attention helps the student model. Table 6 gives empirical results. We can observe the CSA dramatically increase the AP score as well as AP_S for the student RetinaNet model. Moreover, the MSC is critical to the performance. Compared to plain cross-attention, the MSC can improve the AP by 1.4 and AP_S by 1.3. It verifies the advantage of our proposed multi-scale cross-attention over plain cross-attention in our KD framework.

Ablation study of Scale-Decoupled Feature. Table 7 presents the ablation study of the SDF module. We use standard image resolution for teachers in this section. We mainly discuss two aspects: whether SDF improves the performance of SOD and whether the weight-sharing method is worth using. First, we evaluate SDF on COCO with RetinaNet-ResNet50 as the baseline model. The experimental results show that the SDF is indeed effective in improving overall AP and AP for small objects. Besides, using separate weights for parallel branches does not bring noticeable improvement in terms of average AP. Thus, it is worth using the weight-sharing module to save memory costs.

Method	SDF	WS	AP	AP_S	AP_M	AP_L
Baseline	-	-	36.9	21.2	40.8	48.4
ScaleKD	✓	✗	41.6	24.9	44.8	53.7
	✓	✓	41.6	24.8	44.6	53.9

Table 7. Ablation study for Scale-Decoupled Feature module. WS: weight-sharing.

Method	ScaleKD	AP	AP_S	AP_M	AP_L
ResNet-18	✗	35.8	18.9	38.9	47.9
ResNet-18	✓	37.2	20.1	40.2	49.3
ResNet-34	✗	38.9	21.5	42.8	51.4
ResNet-34	✓	40.8	23.1	45.1	53.7

Table 8. Quantitative results of ScaleKD for diverse model complexity. We use ResNet101 as the teacher model.

ScaleKD for Diverse Model Complexity. We further validate the effectiveness of our approach to diverse model complexity. Specifically, we evaluate COCO for two different student backbones, ResNet-18 and ResNet-34. Each model is distilled by a ResNet-101 detector. The results are presented in Table 8, which show that our method can effectively and stably improve the student model, regardless of the model complexity. This observation is contradicted by previous studies on image classification but consistent with the report on detection tasks. In particular, our approach improves the average AP by 1.2 and 2.5 on ResNet-18 and ResNet-34, respectively. Notably, the performance on small objects (AP_S) is boosted tremendously by 2.3 and 2.1. These results verified that our method is generally effective on varying model sizes.

5. Discussion

Balancing inference speed and detection performance for small object detection is challenging. In this work, we propose a Scale-aware knowledge distillation that targets improving the performance of SOD via designed scale-decoupled feature distillation and cross-scale assistant. The former explicitly decouples multi-scale features, and the latter refines the teacher's bounding box noise for more informative knowledge in distillation. We evaluate two benchmarks, COCO 2017 and VisDrone, to demonstrate the effectiveness of our approach.

Acknowledgements: This work was partially supported by the Science and Technology Innovation Action Plan of Shanghai under Grant 2251110540, and Shanghai Pujiang Program under Grant 21PJ1420300.

References

[1] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 206–221, 2018. 1

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 6

[3] Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. Pkd: General distillation framework for object detectors via pearson correlation coefficient. *arXiv preprint arXiv:2207.02039*, 2022. 2

[4] Changrui Chen, Yu Zhang, Qingxuan Lv, Shuo Wei, Xiaorui Wang, Xin Sun, and Junyu Dong. Rrnet: A hybrid detector for object detection in drone-captured images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3

[5] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017. 2, 4, 5, 7, 8

[6] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *arXiv preprint arXiv:2207.14096*, 2022. 3

[7] Xing Dai, Zeren Jiang, Zhao Wu, Yiping Bao, Zhicheng Wang, Si Liu, and Erjin Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2021. 2, 7, 8

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[9] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7036–7045, 2019. 1, 3

[10] Yuqi Gong, Xuehui Yu, Yao Ding, Xiaoke Peng, Jian Zhao, and Zhenjun Han. Effective fusion factor in fpn for tiny object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1160–1168, 2021. 3

[11] Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2154–2164, 2021. 7, 8

[12] Shuxuan Guo, Jose M Alvarez, and Mathieu Salzmann. Distilling image classifiers in object detectors. *Advances in Neural Information Processing Systems*, 34:1036–1047, 2021. 2, 7, 8

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[14] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 1, 2

[15] Tao Huang, Yuan Zhang, Shan You, Fei Wang, Chen Qian, Jian Cao, and Chang Xu. Masked distillation with receptive tokens. *arXiv preprint arXiv:2205.14589*, 2022. 2

[16] Yaomin Huang, Xinmei Liu, Yichen Zhu, Zhiyuan Xu, Chaomin Shen, Zhengping Che, Guixu Zhang, Yaxin Peng, Feifei Feng, and Jian Tang. Label-guided auxiliary training improves 3d object detector. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 684–700. Springer, 2022. 2

[17] Zijian Kang, Peizhen Zhang, Xiangyu Zhang, Jian Sun, and Nanning Zheng. Instance-conditional knowledge distillation for object detection. *Advances in Neural Information Processing Systems*, 34:16468–16480, 2021. 2

[18] Mate Kisantal, Zbigniew Wojna, Jakub Murawski, Jacek Naruniec, and Kyunghyun Cho. Augmentation for small object detection. *arXiv preprint arXiv:1902.07296*, 2019. 3

[19] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020. 7

[20] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6054–6063, 2019. 1, 3

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1, 3

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6, 7

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[24] Bi-Yuan Liu, Huai-Xin Chen, Zhou Huang, Xing Liu, and Yun-Zhi Yang. Zoominnet: A novel small object detector in drone images with cross-scale knowledge distillation. *Remote Sensing*, 13(6):1198, 2021. 6, 7

[25] Marcin Mozejko, Tomasz Latkowski, Lukasz Treszczotko, Michal Szafraniuk, and Krzysztof Trojanowski. Superkernel neural architecture search for image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 484–485, 2020. 4

[26] F Ozge Unel, Burak O Ozkalayci, and Cevahir Cigla. The power of tiling for small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[27] Dae Young Park, Moon-Hyun Cha, Daesin Kim, Bohyung Han, et al. Learning student-friendly teacher networks for knowledge distillation. *Advances in Neural Information Processing Systems*, 34:13292–13303, 2021. 5

[28] Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *International Conference on Machine Learning*, pages 5142–5151. PMLR, 2019. 1

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 6

[30] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014. 2, 7, 8

[31] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 7

[32] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3578–3587, 2018. 1, 3

[33] Bharat Singh, Mahyar Najibi, and Larry S Davis. Sniper: Efficient multi-scale training. *Advances in neural information processing systems*, 31, 2018. 1, 3

[34] Dimitrios Stamoulis, Ruizhou Ding, Di Wang, Dimitrios Lymberopoulos, Bodhi Priyantha, Jie Liu, and Diana Marculescu. Single-path nas: Designing hardware-efficient convnets in less than 4 hours. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 481–497. Springer, 2019. 4

[35] Ruining Tang, Zhenyu Liu, Yangguang Li, Yiguo Song, Hui Liu, Qide Wang, Jing Shao, Guifang Duan, and Jianrong Tan. Task-balanced distillation for object detection. *arXiv preprint arXiv:2208.03006*, 2022. 2

[36] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 6, 7

[37] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *Advances in neural information processing systems*, 32, 2019. 3

[38] Sanghyun Woo, Soonmin Hwang, and In So Kweon. Stairnet: Top-down semantic aggregation for accurate one shot detection. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 1093–1102. IEEE, 2018. 3

[39] Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13668–13677, 2022. 1, 3

[40] Chenhongyi Yang, Mateusz Ochal, Amos Storkey, and Elliot J Crowley. Prediction-guided distillation for dense object detection. *arXiv preprint arXiv:2203.05469*, 2022. 2

[41] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4643–4652, 2022. 2, 3, 7, 8

[42] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9657–9666, 2019. 6

[43] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020. 5

[44] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2020. 2, 7

[45] Peizhen Zhang, Zijian Kang, Tong Yang, Xiangyu Zhang, Nanning Zheng, and Jian Sun. Lgd: Label-guided self-distillation for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3309–3317, 2022. 2, 7, 8

[46] Zhaohui Zheng, Rongguang Ye, Ping Wang, Dongwei Ren, Wangmeng Zuo, Qibin Hou, and Ming-Ming Cheng. Localization distillation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9407–9416, 2022. 5, 7, 8

[47] Du Zhixing, Rui Zhang, Ming Chang, Shaoli Liu, Tianshi Chen, Yunji Chen, et al. Distilling object detectors with feature richness. *Advances in Neural Information Processing Systems*, 34:5213–5224, 2021. 2

[48] Hongyu Zhou, Zheng Ge, Songtao Liu, Weixin Mao, Zeming Li, Haiyan Yu, and Jian Sun. Dense teacher: Dense pseudo-labels for semi-supervised object detection. In *European Conference on Computer Vision*, pages 35–50. Springer, 2022. 2

[49] Peng Zhou, Bingbing Ni, Cong Geng, Jianguo Hu, and Yi Xu. Scale-transferrable object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 528–537, 2018. 1

[50] Wangchunshu Zhou, Canwen Xu, and Julian McAuley. Bert learns to teach: Knowledge distillation with meta learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7037–7049, 2022. 5

[51] Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Haibin Ling, Qinghua Hu, Qinqin Nie, Hao Cheng, Chenfeng Liu, Xiaoyu Liu, et al. Visdrone-det2018: The vision meets drone object detection in image challenge results. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 6

[52] Yichen Zhu, Ning Liu, Zhiyuan Xu, Xin Liu, Weibin Meng, Louis Wang, Zhicai Ou, and Jian Tang. Teach less, learn more: On the undistillable classes in knowledge distillation. In *Advances in Neural Information Processing Systems*, 2022. 2

[53] Yichen Zhu and Yi Wang. Student customized knowledge distillation: Bridging the gap between student and teacher.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5057–5066, 2021. 2