Prediction-Guided Distillation for Dense Object Detection

Chenhongyi Yang[1⋆], Mateusz Ochal[2,3], Amos Storkey[2], and Elliot J. Crowley[1]

[1] School of Engineering, University of Edinburgh, UK
[2] School of Informatics, University of Edinburgh, UK
[3] School of Engineering and Physical Sciences, Heriot-Watt University, UK

Abstract. Real-world object detection models should be cheap and accurate. Knowledge distillation (KD) can boost the accuracy of a small, cheap detection model by leveraging useful information from a larger teacher model. However, a key challenge is identifying the most informative features produced by the teacher for distillation. In this work, we show that only a very small fraction of features within a ground-truth bounding box are responsible for a teacher's high detection performance. Based on this, we propose Prediction-Guided Distillation (PGD), which focuses distillation on these *key predictive regions* of the teacher and yields considerable gains in performance over many existing KD baselines. In addition, we propose an adaptive weighting scheme over the key regions to smooth out their influence and achieve even better performance. Our proposed approach outperforms current state-of-the-art KD baselines on a variety of advanced one-stage detection architectures. Specifically, on the COCO dataset, our method achieves between +3.1% and +4.6% AP improvement using ResNet-101 and ResNet-50 as the teacher and student backbones, respectively. On the CrowdHuman dataset, we achieve +3.2% and +2.0% improvements in MR and AP, also using these backbones. Our code is available at `https://github.com/ChenhongyiYang/PGD`.

Keywords: Dense Object Detection, Knowledge Distillation

1 Introduction

Advances in deep learning have led to considerable performance gains on object detection tasks [2, 6, 11, 15, 18, 25, 26, 27, 31]. However, detectors can be computationally expensive, making it challenging to deploy them on devices with limited resources. Knowledge distillation (KD) [1, 13] has emerged as a promising approach for compressing models. It allows for the direct training of a smaller student model [17, 24, 28, 33] using information from a larger, more powerful teacher model; this helps the student to generalise better than if trained alone.

KD was first popularised for image classification [13] where a student model is trained to mimic the *soft labels* generated by a teacher model. However, this

⋆ Corresponding Author. Email: chenhongyi.yang@ed.ac.uk

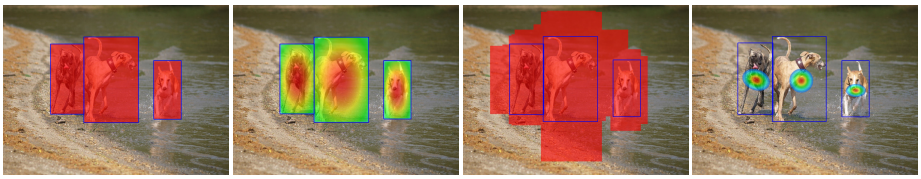

Fig. 1: A comparison between different foreground distillation regions. The ground-truth bounding box is marked in blue. The colour heatmaps indicate the distillation weight for different areas. In contrast to other methods (a)-(c) [9,30,34], Our approach (d) focuses on a few key predictive regions of the teacher.

approach does not work well for object detection [34] which consists of jointly classifying and localising objects. While soft label-based KD can be directly applied for classification, finding an equivalent for localisation remains a challenge. Recent work [8,9,30,34,35,37,41] alleviates this problem by forcing the student model to generate feature maps similar to the teacher counterpart; a process known as *feature imitation*.

However, which features should the student imitate? This question is of the utmost importance for dense object detectors [6,15,18,31,38,42] because, unlike two-stage detectors [2, 11, 27], they do not use the RoIAlign [11] operation to explicitly pool and align object features; instead they output predictions at every location of the feature map [16]. Recent work [30,35] has shown that distilling the whole feature map with equal weighting is sub-optimal because not all features carry equally meaningful information. Therefore, a weighting mechanism that assigns appropriate importance to different regions, particularly to *foreground* regions near the objects, is highly desirable for dense object detectors, and has featured in recent work. For example, in DeFeat [9], foreground features that lie within ground truth (GT) boxes (Fig. 1a) are distilled with equal weighting. In [30] the authors postulate that useful features are located at the centre of GT boxes and weigh the foreground features using a Gaussian (Fig. 1b). In Fine-grained Feature Imitation (FGFI) [34], the authors distil features covered by anchor boxes whose Intersection over Union (IoU) with the GTs are above a certain threshold (Fig. 1c).

In this paper, we treat feature imitation for foreground regions differently. Instead of assigning distillation weights using hand-design policies, we argue that feature imitation should be conducted on a few *key predictive regions*: the locations where the teacher model generates the most accurate predictions. Our intuition is that these regions should be distilled because they hold the information that leads to the best predictions; other areas will be less informative and can contaminate the distillation process by distracting from more essential features. To achieve our goal, we adapt the *quality* measure from [6] to score teacher predictions. Then, we conduct an experiment to visualise how these scores are distributed and verify that high-scoring *key predictive regions* contribute the most to teacher performance. Those findings drive us to propose a

Prediction-Guided Weighting (PGW) module to weight the foreground distillation loss: inspired by recent progress in label assignment [6, 20, 32, 38, 42] for dense detectors, we sample the top-K positions with the highest quality score from the teacher model and use an adaptive Gaussian distribution to fit the *key predictive regions* for smoothly weighting the distillation loss. Fig. 1d shows a visual representation of the regions selected for distillation. We call our method Prediction-Guided Distillation (PGD). Our contributions are as follows:

1. We conduct experiments to study how the *quality* scores of teacher predictions are distributed in the image plane and observe that the locations that make up the top-1% of scores are responsible for most of the teacher's performance in modern state-of-the-art dense detectors.
2. Based on our observations, we propose using the *key predictive regions* of the teacher as foreground features. We show that focusing distillation mainly on these few areas yields significant performance gains for the student model.
3. We introduce a parameterless weighting scheme for foreground distillation pixels and show that when applied to our *key predictive regions*, we achieve even stronger distillation performance.
4. We benchmark our approach on the COCO and CrowdHuman datasets and show its superiority over the state-of-the-art across multiple detectors.

2 Related Work

Dense Object Detection. In the last few years, object detection has seen considerable gains in performance [2, 3, 6, 11, 15, 18, 25, 26, 27, 31]. The demand for simple, fast models has brought one-stage detectors into the spotlight [6, 31]. In contrast to two-stage detectors, one-stage detectors directly regress and classify candidate bounding boxes from a pre-defined set of anchor boxes (or anchor points), alleviating the need for a separate region proposal mechanism. Anchor-based detectors [6,18] achieve good performance by regressing from anchor boxes with pre-defined sizes and ratios. In contrast, anchor-free methods [15, 31, 42] regress directly from anchor points (or locations), eliminating the need for the additional hyper-parameters used in anchor-based models. A vital challenge for detectors is determining which bounding box predictions to label as positive and negative – a problem frequently referred to as *label assignment* [42]. Anchors are commonly labelled as positives when their IoU with the GT is over a certain threshold (e.g. IoU ≥ 0.5) [18, 31], however, more elaborate mechanisms for label assignment have been proposed [6, 31, 38, 42]. For example, FCOS [31] applies a weighting scheme to suppress low-quality positive predictions using a "center-ness" score. Other works dynamically adjust the number of positive instances according to statistical characteristics [38] or by using a differentiable confidence module [42]. In DDOD [6], the authors separate label assignment for the classification and regression branches and balance the influence of positive samples between different scales of the feature pyramid network (FPN).

Knowledge Distillation for Object Detection. Early KD approaches for classification focus on transferring knowledge to student models by forcing their predictions to match those of the teacher [13]. More recent work [34, 35, 41] claims that feature imitation, i.e. forcing the intermediate feature maps of student models to match their teacher counterpart, is more effective for detection. A vital challenge when performing feature imitation for dense object detectors is determining which feature regions to distil from the teacher model. Naively distilling all feature maps equally results in poor performance [9,30,35]. To solve this problem, FGFI [34] distils features that are covered by anchor boxes which have a high IoU with the GT. However, distilling in this manner is still sub-optimal [8,30,35,40,41]. TADF [30] suppresses foreground pixels according to a static 2D Gaussian fitted over the GT. LD [40] gives higher priority to central locations of the GT using DIoU [39]. GID [8] propose to use the top-scoring predictions using L1 distance between the classifications scores of the teacher and the student, but do not account for location quality. In LAD [21], the authors use *label assignment* distillation where the detector's encoded labels are used to train a student. Others weight foreground pixels according to intricate adaptive weighting or attention mechanisms [8, 14, 35, 36, 41]. However, these weighting schemes still heavily rely on the GT dimensions, and they are agnostic to the capabilities of the teacher. In contrast, we focus distillation on only a few *key predictive regions* using a combination of classification and regression scores as a measure of quality. We then smoothly aggregate and weigh the selected locations using an estimated 2D Gaussian, which further focuses distillation and improves performance. This allows us to dynamically adjusts to different sizes and orientations of objects independently of the GT dimensions while accounting for the teacher's predictive abilities.

3 Method

We begin by describing how to measure the predictive quality of a bounding box prediction and find the *key predictive regions* of a teacher network (Sec. 3.1). Then, we introduce our *Prediction-Guided Weighting* (PGW) module that returns a foreground distillation mask based on these regions (Sec. 3.2). Finally, we describe our full Prediction-Guided Distillation pipeline (Sec. 3.3).

3.1 Key Predictive Regions

Our goal is to amplify the distillation signal for the most meaningful features produced by a teacher network. For this purpose, we look at the *quality* of a teacher's bounding box predictions taking both classification and localisation into consideration, as defined in [6]. Formally, the quality score of a box $\hat{b}_{(i,j)}$ predicted from a position $X_i = (x_i, y_i)$ w.r.t. a ground truth b is:

$$q(\hat{b}_{(i,j)}, b) = \underbrace{\mathbb{1}\left[X_i \in b\right]}_{\text{indicator}} \cdot \underbrace{\left(\hat{p}_{(i,j)}(b)\right)^{1-\xi}}_{\text{classification}} \cdot \underbrace{\left(\text{IoU}\big(b, \hat{b}_{(i,j)}\big)\right)^{\xi}}_{\text{localisation}} \tag{1}$$

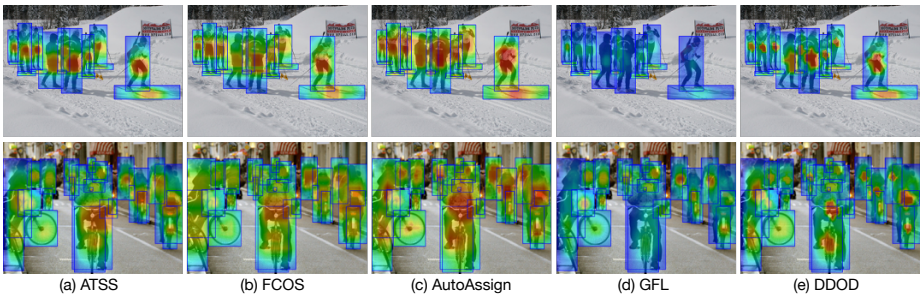

Fig. 2: A visualisation of quality scores for various dense object detectors with $\xi = 0.8$ following [6]. We acquire the quality heatmap by taking the maximum value at each position across FPN layers.

where $\mathbb{1}\left[X_i \in \Omega_b\right]$ is an indicator function that is 1 if X_i lies inside box b and 0 otherwise; $\hat{p}_{(i,j)}(b)$ is the classification probability w.r.t. the GT box's category; $\text{IoU}\big(b, \hat{b}_{(i,j)}\big)$ is the IoU between the predicted and ground-truth box; ξ is a hyper-parameter that balances classification and localisation. We calculate the quality score of location X_i as the maximum value of all prediction scores for that particular location, i.e. $\hat{q}_i = max_{j \in J_i}\ q(\hat{b}_{(i,j)}, b)$, where J_i is the set of predictions at location X_i. While this quality score has been applied for standard object detection [6], we are the first to use it to identify useful regions for distillation.

In Fig. 2 we visualise the heatmaps of prediction quality scores for five state-of-the-art detectors, including anchor-based (ATSS [38] and DDOD [6]) and anchor-free (FCOS [31], GFL [15] and AutoAssign [42]) detectors. Across all detectors, we observe some common characteristics: (1) For the vast majority of objects, high scores are concentrated around a **single region**; (2) The size of this region doesn't necessarily correlate strongly with the size of the actual GT box; (3) Whether or not the centring prior [31, 42] is applied for label assignment during training, this region tends to be close to the centre of the GT box. These observations drive us to develop a *Prediction-Guided Weighting* (PGW) module to focus the distillation on these important regions.

3.2 Prediction-Guided Weighting Module

The purpose of KD is to allow a student to mimic a teacher's strong generalisation ability. To better achieve this goal, we propose to focus foreground distillation on locations where a teacher model can yield predictions with the highest quality scores because those locations contain the most valuable information for detection and are critical to a teacher's high performance. In Fig. 3 we present the results of a pilot experiment to identify how vital these high-scoring locations are for a detector. Specifically, we measure the performance of different pre-trained detectors after masking out their top-$X\%$ predictions before non-maximum suppression (NMS) during inference. We observe that in all cases the mean Averaged Precision (mAP) drops dramatically as the mask-out ratio

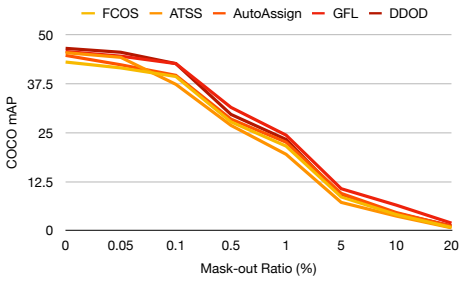

Fig. 3: COCO mAP performance of pre-trained detectors after ignoring predictions in the top-X% of quality scores during inference. We observe that the top-1% predictions within the GT box region are responsible for most performance gains.

increases. Masking out the top-1% of predictions incurs around a 50% drop in AP. This suggests that the *key predictive regions* (responsible for the majority of a dense detector's performance) lie within the top-1% of all anchor positions bounded by the GT box.

Given their significance, how do we incorporate these regions into distillation? We could simply use all feature locations weighted by their quality score, however, as we show in Sec. 4.3 this does not yield the best performance. Inspired by recent advances in label assignment for dense object detectors [6, 32], we instead propose to focus foreground distillation on the top-K positions (feature pixels) with the highest quality scores across all FPN levels. We then smooth the influence of each position according to a 2D Gaussian distribution fitted by Maximum-Likelihood Estimation (MLE) for each GT box. Finally, foreground distillation is conducted only on those K positions with their weights assigned by the Gaussian.

Formally, for an object o with GT box b, we first compute the quality score for each feature pixel inside b, then we select the K pixels with the highest quality score $T^o = \{(X_k^o, l_k^o)|k = 1, ..., K\}$ across all FPN levels, in which X_k^o and l_k^o are the absolute coordinate and the FPN level of the k-th pixel. Based on our observation in Sec. 3.1, we assume the selected pixels T_k^o are drawn as $T_k^o \sim \mathcal{N}(\mu, \Sigma|o)$ defined on the image plane and use MLE to estimate μ and Σ:

$$\hat{\mu} = \frac{1}{K}\sum_{k=1}^{K} X_k^o, \quad \hat{\Sigma} = \frac{1}{K}\sum_{k=1}^{K}(X_k^o - \hat{\mu})(X_k^o - \hat{\mu})^T \tag{2}$$

Then, for every feature pixel $P_{(i,j),l}$ on FPN layer l with absolute coordinate $X_{i,j}$, we compute its distillation importance w.r.t. object o by:

$$I_{(i,j),l}^o = \begin{cases} 0 & P_{(i,j),l} \notin T^o \\ \exp\left(-\frac{1}{2}(X_{i,j} - \hat{\mu})\hat{\Sigma}^{-1}(X_{i,j} - \hat{\mu})^T\right) & P_{(i,j),l} \in T^o \end{cases} \tag{3}$$

If a feature pixel has non-zero importance for multiple objects, we use its maximum: $I_{(i,j),l} = \max_o \{I_{(i,j),l}^o\}$. Finally, for each FPN level l with size $H_l \times W_l$, we assign the distillation weight $M_{(i,j),l}$ by normalising the distillation importance by the number of non-zero importance pixels at that level:

$$\mathbf{M_{(i,j),l}} = \frac{I_{(i,j),l}}{\sum_{i=1}^{H_l}\sum_{j=1}^{W_l} \mathbb{1}_{(i,j),l}} \tag{4}$$

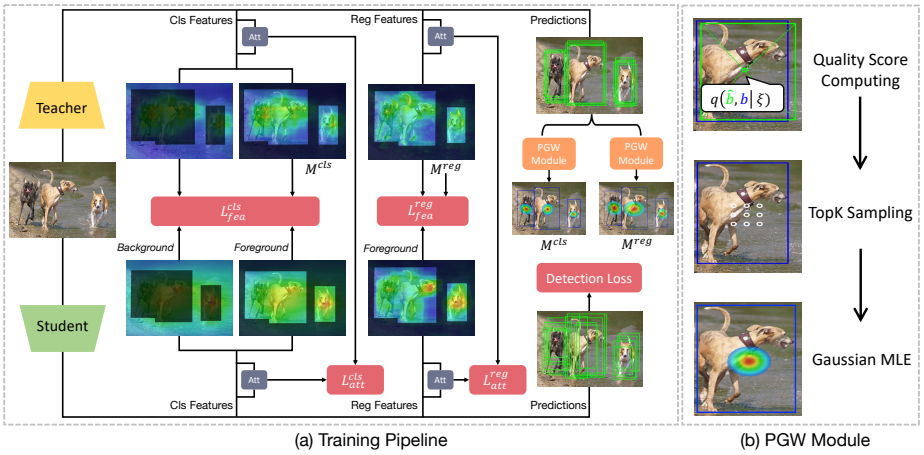

Fig. 4: Our Prediction-Guided Distillation (PGD) pipeline. The *Prediction-Guided Weighting* (PGW) modules find the teacher's *key predictive regions* and generates a foreground distillation weighting mask by fitting a Gaussian over these regions. Our pipeline also adopts the attention masks from FGD [35] and distils them together with the features. We distil the classification and regression heads separately to accommodate for these two distinct tasks. [6].

where $\mathbb{1}_{(i,j),l}$ is an indicator function that outputs 1 if $I_{(i,j),l}$ is not zero. The process above constitutes our *Prediction-Guided Weighting* (PGW) module whose output is a foreground distillation weight $\mathbf{M}$ across all feature levels and pixels.

3.3 Prediction-Guided Distillation

In this section, we introduce our KD pipeline, which is applicable to any dense object detector. We build our work on top of the state-of-the-art Focal and Global Distillation (FGD) [35] and incorporate their spatial and channel-wise attention mechanisms. In contrast to other distillation methods, we use the output mask from our PGW module to focus the distillation loss on the most important foreground regions. Moreover, we decouple the distillation for the classification and regression heads to better suit the two different tasks [6, 22]. An illustration of the pipeline is shown in Fig 4.

Distillation of Features. We perform feature imitation at each FPN level, encouraging feature imitation on the first feature maps of the regression and classifications heads. Taking inspiration from [6], we separate the distillation process for the classification and regression heads – distilling features of each head independently. Formally, at each feature level of the FPN, we generate two foreground distillation masks $\mathbf{M^{cls}}, \mathbf{M^{reg}} \in \mathbb{R}^{H \times W}$ with different ξ^{cls} and ξ^{reg} using PGW. Then, student features $F^{S,cls}, F^{S,reg} \in \mathbb{R}^{C \times H \times W}$ are encouraged to mimic teacher features $F^{T,cls}, F^{T,reg} \in \mathbb{R}^{C \times H \times W}$ as follows:

$$L_{fea}^{cls} = \sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}(\alpha\mathbf{M}_{\mathbf{i,j}}^{\mathbf{cls}} + \beta N_{i,j}^{cls})P_{i,j}^{T,cls}A_{k,i,j}^{T,cls}(F_{k,i,j}^{T,cls} - F_{k,i,j}^{S,cls})^2 \tag{5}$$

$$L_{fea}^{reg} = \sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}\gamma\mathbf{M}_{\mathbf{i,j}}^{\mathbf{reg}}A_{k}^{T,reg}(F_{k,i,j}^{T,reg} - F_{k,i,j}^{S,reg})^2 \tag{6}$$

where α, β, γ are hyperparameters to balance between loss weights; N^{cls} is the normalised mask over background distillation regions: $N_{i,j}^{cls} = \mathbb{1}_{i,j}^{-}/\sum_{h=1,w=1}^{H,W}\mathbb{1}_{w,h}^{-}$ where $\mathbb{1}_{a,b}^{-}$ is the background indicator that becomes 1 if pixel (a, b) does not lie within any GT box. P and A are spatial and channel attention maps from [35] as defined below. Note, we do not use the Global Distillation Module in FGD and the adaptation layer that is commonly used in many KD methods [4, 9, 34, 35, 37, 41] as we find them have negligible impact to the overall performance.

Distillation of Attention. We build on the work in FGD [35] and additionally encourage the student to imitate the attention maps of the teacher. We use spatial attention as defined in [35], but we modify their channel attention by computing it independently for each feature location instead of all spatial locations. Specifically, we define s**p**atial attention $\mathbf{P} \in \mathbb{R}^{1\times H\times W}$ and ch**a**nnel attention $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$ over a single feature map $F \in \mathbb{R}^{C\times H\times W}$ as follows:

$$P_{i,j} = \frac{HW \cdot \exp\left(\sum_{k=1}^{C}|F_{k,i,j}|/\tau\right)}{\sum_{i=1}^{H}\sum_{j=1}^{W}\exp\left(\sum_{k=1}^{C}|F_{k,i,j}|/\tau\right)}, \quad A_{k,i,j} = \frac{C \cdot \exp\left(|F_{k,i,j}|/\tau\right)}{\sum_{k=1}^{C}\exp\left(|F_{k,i,j}|/\tau\right)} \tag{7}$$

Similar to feature distillation, we decouple the attention masks for classification and regression for the teacher and student: $A^{T,cls}$, $A^{T,reg}$, $P^{S,cls}$. The two attention losses are defined as follows:

$$L_{att}^{cls} = \frac{\delta}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}|P_{i,j}^{T,cls} - P_{i,j}^{S,cls}| + \frac{\delta}{CHW}\sum_{k=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}|A_{k,i,j}^{T,cls} - A_{k,i,j}^{S,cls}| \tag{8}$$

$$L_{att}^{reg} = \frac{\delta}{C\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbb{1}_{i,j}}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{C}\mathbb{1}_{i,j}|A_{k,i,j}^{T,reg} - A_{k,i,j}^{S,reg}| \tag{9}$$

where δ is balancing loss weight hyperparameter; and $\mathbb{1}_{i,j}$ is an indicator that becomes 1 when $\mathbf{M}_{\mathbf{i,j}}^{\mathbf{reg}} \neq 0$.

Full Distillation. The full distillation loss is

$$L_{distill} = L_{fea}^{cls} + L_{fea}^{reg} + L_{att}^{cls} + L_{att}^{reg} \tag{10}$$

4 Experiments

4.1 Setup and Implementation Details

We evaluate PGD on two benchmarks: COCO [19] for general object detection and CrowdHuman [29] for crowd scene detection; this contains a large number of

Detector	Setting	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
FCOS [31]	Teacher	43.1	62.4	46.6	25.5	47.1	54.7
	Student	38.2	57.9	40.5	23.1	41.3	49.4
	DeFeat [9]	40.7(+2.5)	60.5(+2.6)	43.5(+3.0)	24.7(+1.6)	44.4(+3.1)	52.4(+3.0)
	FRS [41]	40.9(+2.7)	60.6(+2.7)	44.0(+3.5)	25.0(+1.9)	44.4(+3.1)	52.6(+3.2)
	FKD [37]	41.3(+3.1)	60.9(+3.0)	44.1(+3.6)	23.9(+0.8)	44.9(+3.6)	53.8(+4.4)
	FGD [35]	41.4(+3.2)	61.1(+3.2)	44.2(+3.7)	25.3(+**2.2**)	45.1(+3.8)	53.8(+4.4)
	Ours	42.5(+**4.3**)	62.0(+**4.1**)	45.4(+**4.9**)	24.8(+1.7)	46.1(+**5.8**)	55.5(+**6.1**)
Auto-Assign [42]	Teacher	44.8	64.1	48.9	27.3	48.8	57.5
	Student	40.6	60.1	43.8	23.6	44.3	52.4
	DeFeat [9]	42.3(+1.7)	61.6(+1.5)	46.1(+2.3)	24.1(+0.5)	46.0(+1.7)	54.4(+2.0)
	FRS [41]	42.4(+1.8)	61.9(+1.8)	46.0(+2.2)	24.9(+1.3)	46.0(+1.7)	54.8(+2.4)
	FKD [37]	42.8(+2.2)	62.1(+2.0)	46.5(+2.7)	25.7(+2.1)	46.4(+2.1)	55.5(+3.1)
	FGD [35]	43.2(+2.6)	62.5(+2.4)	46.9(+3.1)	25.2(+1.6)	46.7(+2.4)	56.2(+3.8)
	Ours	43.8(+**3.1**)	62.9(+**2.8**)	47.4(+**3.6**)	25.8(+**2.2**)	47.3(+**3.0**)	57.5(+**5.1**)
ATSS [38]	Teacher	45.5	63.9	49.7	28.7	50.1	57.8
	Student	39.6	57.6	43.2	23.0	42.9	51.2
	DeFeat [9]	41.8(+2.2)	60.3(+2.7)	45.3(+2.1)	24.8(+1.8)	45.6(+2.7)	53.5(+2.3)
	FRS [41]	41.6(+2.0)	60.1(+2.5)	44.8(+1.6)	24.9(+1.9)	45.2(+2.3)	53.2(+2.0)
	FGFI [34]	41.8(+2.2)	60.3(+2.7)	45.3(+2.1)	24.8(+1.8)	45.6(+2.7)	53.5(+2.3)
	FKD [37]	42.3(+2.7)	60.7(+3.1)	46.2(+3.0)	26.3(+3.3)	46.0(+3.1)	54.6(+3.4)
	FGD [35]	42.6(+3.0)	60.9(+3.3)	46.2(+3.0)	25.7(+2.7)	46.7(+3.8)	54.5(+3.3)
	Ours	44.2(+**4.6**)	62.3(+**4.7**)	48.3(+**5.1**)	26.5(+**3.5**)	48.6(+**5.7**)	57.1(+**5.9**)
GFL [15]	Teacher	45.8	64.2	49.8	28.3	50.3	58.6
	Student	40.2	58.4	43.3	22.7	43.6	52.0
	DeFeat [9]	42.1(+1.9)	60.5(+2.1)	45.2(+1.9)	24.4(+1.7)	46.1(+2.5)	54.5(+2.5)
	FRS [41]	42.2(+2.0)	60.6(+2.2)	45.6(+2.3)	24.7(+2.0)	46.0(+2.4)	55.5(+3.5)
	FKD [37]	43.1(+2.9)	61.6(+3.2)	46.6(+3.3)	25.1(+2.4)	47.2(+3.6)	56.5(+4.5)
	FGD [35]	43.2(+3.0)	61.8(+3.4)	46.9(+3.6)	25.2(+2.5)	47.5(+3.9)	56.2(+4.2)
	LD [40]	43.5(+3.3)	61.8(+3.4)	47.4(+**4.1**)	24.7(+2.0)	47.5(+3.9)	57.3(+5.3)
	Ours	43.8(+**3.6**)	62.0(+**3.6**)	47.4(+**4.1**)	25.4(+**2.7**)	47.8(+**4.2**)	57.6(+**5.6**)
DDOD [6]	Teacher	46.6	65.0	50.7	29.0	50.5	60.1
	Student	42.0	60.2	45.5	25.7	45.6	54.9
	DeFeat [9]	43.2(+1.2)	61.6(+1.4)	46.7(+1.2)	25.7(+0.0)	46.5(+0.9)	57.3(+2.4)
	FRS [41]	43.7(+1.7)	62.2(+2.0)	47.6(+2.1)	25.7(+0.0)	46.8(+1.2)	58.1(+3.2)
	FGFI [34]	44.1(+2.1)	62.6(+2.4)	47.9(+2.4)	26.3(+0.6)	47.3(+1.7)	58.5(+3.6)
	FKD [37]	43.6(+1.6)	62.0(+1.8)	47.1(+1.6)	25.9(+0.2)	47.0(+1.4)	58.1(+3.2)
	FGD [35]	44.1(+2.1)	62.4(+2.2)	47.9(+2.4)	26.8(+1.1)	47.2(+1.6)	58.5(+3.6)
	Ours	45.4(+**3.4**)	63.9(+**3.7**)	49.0(+**3.5**)	26.9(+**1.2**)	49.2(+**3.6**)	59.7(+**4.8**)

Table 1: A comparison between our PGD with other state-of-the-art distillation methods on COCO *mini-val* set. All models are trained locally. We set hyper-parameters for competing methods following their paper or open-sourced code bases.

heavily occluded objects. Our codebase is built on PyTorch [23] and the MMDetection [5] toolkit and is available at `https://github.com/ChenhongyiYang/PGD`. All models are trained on 8 Nvidia 2080Ti GPUs. For both COCO and CrowdHuman, all models are trained using batch sizes of 32 and with an initial learning rate of 0.02, we adopt ImageNet pre-trained backbones and freeze all Batch Normalisation layers during training. Unless otherwise specified, on both dataset we train teacher models for 3× schedule (36 epochs) [10] with multi-scale inputs using ResNet-101 [12] as backbone, and train student models for 1× schedule (12 epochs) with single-scale inputs using ResNet-50 as backbone. The

COCO models are trained using the *train2017* set and evaluated on *mini-val* set following the official evaluation protocol [19]. The CrowdHuman models are trained using the CrowdHuman *training* set, which are then evaluated on the CrowdHuman *validation* set following [7]. We set K in the top-K operation to 30 for all detectors and set α to 0.8 and 0.4 for anchor-based and anchor-free detectors respectively. Following [35], we set $\sigma = 0.0008$, $\tau = 0.8$ and $\beta = 0.5\alpha$; we set $\xi^{cls} = 0.8$ and $\xi^{reg} = 0.6$ following [6]. We empirically set $\gamma = 1.6\alpha$ with minimal tuning.

Detector	Setting	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
FCOS	Teacher	43.1	62.4	46.6	25.5	47.1	54.7
	Student	32.8	51.3	34.5	18.4	35.4	42.6
	FGD	34.7(+1.9)	53.0(+1.7)	36.8(+2.3)	19.8(+1.4)	36.8(+1.4)	44.9(+2.3)
	Ours	37.3(+**4.5**)	55.6(+**4.3**)	39.8(+**5.3**)	20.5(+**2.1**)	40.3(+**4.9**)	49.9(+**7.3**)
ATSS	Teacher	45.5	63.9	49.7	28.7	50.1	57.8
	Student	33.5	50.1	36.0	18.7	36.2	43.6
	FGD	35.8(+2.3)	52.6(+2.5)	38.8(+2.8)	20.6(+1.9)	38.4(+2.2)	46.2(+2.6)
	Ours	38.3(+**4.8**)	55.1(+**5.0**)	41.7(+**5.7**)	21.3(+**2.6**)	41.6(+**5.4**)	51.6(+**8.0**)

Table 2: Distillation results on COCO *mini-val* using MobileNetV2 as the student backbone.

4.2 Main Results

Comparison with State-of-the-art. We compare our PGD and other recent state-of-the-art object detection KD approaches for five high-performance dense detectors on COCO; these are a mixture of anchor-based (ATSS and DDOD) and anchor-free (FCOS, GFL and AutoAssign) detectors for COCO. The results are presented in Table 1. We use the same teacher and student models and the same training settings in each case, and all training is conducted locally. For competing distillation methods, we follow the hyper-parameter settings in their corresponding papers or open-sourced code repositories. We observe that our methods surpass other KD methods with a large margin for all five detectors, which validates the effectiveness of our approach. Our approach significantly improvement over the baseline approach FGD [35] and even outperforms LD [40] when applied to GFL [15], which was specifically designed for this detector. We observe PGD is particularly good at improving the AP_{75} of student models, suggesting that the student model's localisation abilities have been largely improved.

Distilling to a Lightweight Backbone. Knowledge Distillation is usually used to transfer useful information from a large model to a lightweight model suitable for deployment on the edge. With this in mind, we apply PGD using a ResNet-101 as the teacher backbone and a MobileNet V2 [28] as the student backbone on anchor-based (ATSS) and anchor-free (FCOS) detectors. The results are provided in Tab. 2. Our method surpasses the baseline by a significant

Setting	MR ↓	AP ↑	JI ↑
Teacher	41.4	90.2	81.4
Student	46.0	88.0	79.0
FKD []	44.3(-1.7)	89.1(+1.1)	80.0(+1.0)
DeFeat [9]	44.2(-1.8)	89.1(+1.1)	79.9(+0.9)
FRS [41]	44.1(-1.9)	89.2(+1.2)	80.3(+1.3)
FGFI [34]	43.8(-2.2)	89.2(+1.2)	80.3(+1.3)
FGD [35]	43.1(-2.9)	89.3(+1.3)	80.4(+1.4)
Ours	42.8(**-3.2**)	90.0(+**2.0**)	80.7(+**1.7**)

Table 3: A comparison between our PGD with other state-of-the-art distillation methods on CrowdHuman *validation* set using DDOD as object detector.

margin, pointing to its potential for resource-limited applications.

Distillation for Crowd Detection. We compare our approach to other KD methods on the challenging CrowdHuman dataset that features heavily crowded scenes. We use the DDOD object detector for this experiment as it achieves the strongest performance. In addition to detection AP, we report the log miss rate (MR) [7] designed for evaluation in crowded scenes as well as the Jaccard Index (JI) that evaluates a detector's counting ability. The results are available in Table 3. Our approach performs better than all competing methods. While FGD achieves comparable MR and JI scores to our method, the AP for our methods is significantly greater. We believe this is because PGD strongly favours highly accurate predictions during distillation, which directly impacts the AP metric.

Detector	Setting	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
FCOS	S & T	38.2	57.9	40.5	23.1	41.3	49.4
	FGD	39.0(+0.8)	58.6(+0.7)	41.4(+0.9)	23.7(+0.6)	42.1(+0.8)	50.6(+1.2)
	Ours	39.5(+1.3)	59.2(+1.3)	41.9(+1.4)	24.4(+1.3)	42.8(+1.5)	50.6(+1.2)
ATSS	S & T	39.6	57.6	43.2	23.0	42.9	51.2
	FGD	40.2(+0.6)	58.6(+1.0)	43.6(+1.4)	23.3(+0.3)	43.7(+0.8)	52.3(+1.1)
	Ours	40.7(+1.1)	58.9(+1.3)	44.2(+2.0)	24.0(+0.9)	44.2(+1.3)	52.9(+1.7)

Table 4: Self-distillation performance on COCO *mini-val*. ResNet-50 is adopted as teacher and student backbone, which are both trained for 1× schedule.

Self-Distillation. Self-distillation is a special case of knowledge distillation where the teacher and student models are exactly same. It is useful as it can boost a model's performance while avoiding introducing extra parameters. We compare the our method's self-distillation performance with the baseline FGD and present results for both anchor-free FCOS and anchor-based ATSS in Tab. 4. The teachers and students use ResNet-50 as backbone and are trained with 1× schedule using single-scale inputs. We can see that our approach achieves a better performance than the baseline, indicating its effectiveness in self-distillation.

Setting	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Teacher	45.5	63.9	49.7	28.7	50.1	57.8
Student	39.6	57.6	43.2	23.0	42.9	51.2
Box	43.3 (+3.7)	61.4(+3.8)	47.2(+4.0)	25.9(+2.9)	47.6(+4.7)	56.4(+5.2)
BoxGauss	43.7(+4.1)	61.9(+4.3)	47.6(+4.4)	26.7(+3.7)	47.8(+4.9)	56.6(+5.4)
Centre	43.1(+3.5)	61.0(+3.4)	46.9(+3.7)	25.9(+2.9)	47.3(+4.4)	56.1(+4.9)
Quality	43.8(+4.2)	61.8(+4.2)	47.8(+4.6)	25.7(+2.7)	48.2(+5.3)	56.8(+5.6)
TopkEq	43.9(+4.3)	62.0(+4.4)	47.7(+4.5)	27.1(+**4.1**)	48.0(+5.1)	56.8(+5.6)
KDE	44.0(+4.4)	62.1(+4.5)	47.8(+4.6)	26.3(+3.3)	48.5(+5.6)	56.8(+5.6)
Ours	44.2(+**4.6**)	62.3(+**4.7**)	48.3(+**5.1**)	26.5(+3.5)	48.6(+**5.7**)	57.1(+**5.9**)

Table 5: Ablation study on different foreground distillation strategies on COCO *mini-val* seet using ATSS as object detector.

K	1	5	9	15	30	45	60
AP	43.2	43.5	43.6	43.9	**44.2**	44.0	43.9

(a) Ablation study on different K in the top-K operation using ATSS as detector.

α	0.005	0.01	0.03	0.05	0.07	0.1	0.2
FCOS	41.7	42.0	**42.5**	**42.5**	42.4	42.2	41.8
ATSS	42.9	43.2	43.7	43.9	**44.2**	44.1	43.2

(b) Ablation study on distillation loss magnitude α using FCOS and ATSS.

Table 6: Hyper-parameter ablation studies on COCO *mini-val*.

4.3 Ablation Study

Comparing Foreground Distillation Strategies. We compare alternative strategies for distilling foreground regions to investigate how important is distilling different foreground regions. We use ATSS as our object detector and present results in Tab. 5. Note here we only modify the foreground distillation strategy while keeping everything else the same. We first evaluate the strategy used in FGD [35] and DeFeat [9], where regions in the GT box are distilled equally. We dub this the *Box* strategy (Fig. 1a). Compared to our method, *Box* achieves 0.9 AP worse performance. A possible reason for this is that it can include sub-optimal prediction locations that distract from more meaningful features. Note that the *Box* strategy still outperforms the baseline FGD, we attribute this improvement to the decoupling of distillation for classification and regression branches. Several works [30, 40] postulate that most meaningful regions lie near the centre of the GT box. We evaluate the *BoxGauss* strategy that was proposed in TADF [30] (Fig. 4b). Specifically, a Gaussian distribution is used to weight the distillation loss, where its mean is the centre of the GT box, and the standard deviation is calculated from the box dimensions. This strategy yields +0.4 AP improvement over vanilla *Box* strategy, suggesting the importance of focusing on the centre area; however, it is still surpassed by our approach. We consider a *Centre* strategy, which distils a $0.2H \times 0.2W$ area at the middle of the GT box. Somewhat surprisingly, this achieves an even worse AP than the

vanilla *Box* strategy in almost all instances, with comparable performance on small objects. A possible explanation is that a fixed ratio region fails to cover the full span of useful regions for different-sized objects and limits the amount of distilled information. Then we compare to an adaptive loss weighting mechanism where we directly use the quality score in Equation 1 to weight features for the distillation loss. The strategy—which we refer to as *Quality*—improves slightly on *BoxGauss*, especially for medium and high scoring boxes. However, it significantly under-performs on small objects. In contrast, the *TopkEq* strategy, where we limit distillation to only the top-K pixels according to the quality score (we set $K = 30$ to match our method), provides a significant improvement to the detection of small objects. A possible explanation for this is that distilling on positions with lower scores still introduces considerable noise, whereas limiting distillation to only the highest-scoring pixels focuses the student towards only the most essential features of the teacher. Finally, we compare our method to one that replaces the Gaussian MLE with kernel density estimation, the *KDE* strategy. It achieves similar performance to our Gaussian MLE approach, but is more complicated.

Hyper-parameter Settings. Here, we examine the effect of changing two important hyper-parameters used in our approach, as presented in Tab. 6. The first is K, which is the number of high-scoring pixels used for distillation. The best performance is obtained for $K = 30$. Small K can cause distillation to neglect important regions, whereas large K can introduce noise that distracts the distillation process from the most essential features. The second hyper-parameter we vary is α which controls the magnitude of the distillation loss. We can see how this affects performance for anchor-based ATSS and anchor-free FCOS. We find that the ATSS's performance is quite robust when α is between 0.05–0.1, and FCOS can achieve good performance when α is between 0.03–0.1. For both types of detectors, a small α will minimise the effect of distillation, and a large α can make training unstable.

Decoupled Distillation. In our pipeline, we decouple the KD loss to distil the classification and regression heads separately (see Section 3.3). This practice differs from previous feature imitation-based approaches where the FPN neck features are distilled. Here we conduct experiments to test this design and present the result in Tab. 7. Firstly, we remove the regression KD loss and only apply the classification KD loss using FPN features. The model achieves 43.6 mAP on COCO. Then we test only applying the classification KD loss using the classification feature map; the performance improves very slightly (by 0.2). Next, we only test the regression KD loss using the regression features, resulting in 41.7 COCO mAP. The performance is significantly harmed because the regression KD loss only considers foreground regions while ignoring background areas. Finally, we come to our design by combining both classification and regression KD losses, which achieves the best performance, at 44.2 COCO mAP.

Fig. 5: Visualisation of the detection results on COCO *mini-val* set using ATSS as detector and PGD for distillation. GTs are shown in blue; plain student detections are shown in red; distilled student predictions are shown in orange.

neck	cls	reg	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
-	-	-	39.6	57.6	43.2	23.0	42.9	51.2
✓			43.6(+4.0)	61.8(+4.2)	47.5(+4.3)	26.1(+**3.1**)	47.8(+4.9)	56.8(+5.6)
	✓		43.8(+4.2)	62.1(+4.5)	47.5(+4.3)	26.5(3.5)	48.0(+5.1)	56.8(+5.6)
		✓	41.7(+2.1)	60.2(+2.6)	45.2(+2.0)	25.3(+2.3)	45.4(+2.5)	53.9(+2.7)
	✓	✓	44.2(+**4.6**)	62.3(+**4.7**)	48.3(+**5.1**)	26.5(+3.5)	48.6(+**5.7**)	57.1(+**5.9**)

Table 7: Comparison between different distillation branches.

Qualitative Studies. We visualise box predictions using ATSS as our object detector in Fig. 5, in which we show GT boxes alongside student predictions with and without distillation using PGD. While the high-performance ATSS is able to accurately detect objects in most cases, we observe some clear advantages of using our distillation approach: it outputs fewer false positives (Fig. 5 b,c), improves detection recall (Fig. 5 a,d), and localises objects better (Fig. 5 b,d,e).

5 Conclusion

In this work, we highlight the need to focus distillation on features of the teacher that are responsible for high-scoring predictions. We find that these *key predictive regions* constitute only a small fraction of all features within the boundaries of the ground-truth bounding box. We use this observation to design a novel distillation technique—PGD—that amplifies the distillation signal from these features. We use an adaptive Gaussian distribution to smoothly aggregate those top locations to further enhance performance. Our approach can significantly improve state-of-the-art detectors on COCO and CrowdHuman, outperforming many existing KD methods. In future, we could investigate the applicability of high-quality regions to two-stage and transformer models for detection.

Acknowledgements

The authors would like to thank Joe Mellor, Kaihong Wang, and Zehui Chen for their useful comments and suggestions. This work was supported by a PhD studentship provided by the School of Engineering, University of Edinburgh as well as the EPSRC Centre for Doctoral Training in Robotics and Autonomous Systems (Grant No. EP/S515061/1) and SeeByte Ltd, Edinburgh, UK.

References

1. Ba, L.J., Caruana, R.: Do deep nets really need to be deep? In: NeurIPS (2014)
2. Cai, Z., Vasconcelos, N.: Cascade R-CNN: Delving into high quality object detection. In: CVPR (2018)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers. In: ECCV (2020)
4. Chen, G., Choi, W., Yu, X., Han, T., Chandraker, M.: Learning efficient object detection models with knowledge distillation. In: NeurIPS (2017)
5. Chen, K., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. In: arXiv preprint arXiv:1906.07155 (2019)
6. Chen, Z., Yang, C., Li, Q., Zhao, F., Zha, Z.J., Wu, F.: Disentangle your dense object detector. In: ACM MM (2021)
7. Chu, X., Zheng, A., Zhang, X., Sun, J.: Detection in Crowded Scenes: One Proposal, Multiple Predictions. In: CVPR (2020)
8. Dai, X., Jiang, Z., Wu, Z., Bao, Y., Wang, Z., Liu, S., Zhou, E.: General instance distillation for object detection. In: CVPR (2021)
9. Guo, J., Han, K., Wang, Y., Wu, H., Chen, X., Xu, C., Xu, C.: Distilling object detectors via decoupled features. In: CVPR (2021)
10. He, K., Girshick, R., Dollár, P.: Rethinking Imagenet Pre-training. In: CVPR (2019)
11. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
13. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. In: NeurIPS 2014 Deep Learning Workshop (2014)
14. Kang, Z., Zhang, P., Zhang, X., Sun, J., Zheng, N.: Instance-Conditional Knowledge Distillation for Object Detection. In: NeurIPS (2021)
15. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In: NeurIPS (2020)
16. Li, Y., Chen, Y., Wang, N., Zhang, Z.: Scale-aware trident networks for object detection. In: ICCV (2019)
17. Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y., Sun, J.: Light-head r-cnn: In defense of two-stage object detector. In: arXiv preprint arXiv:1711.07264 (2017)
18. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
20. Ma, Y., Liu, S., Li, Z., Sun, J.: IQDet: Instance-Wise Quality Distribution Sampling for Object Detection. In: CVPR (2021)
21. Nguyen, C.H., Nguyen, T.C., Tang, T.N., Phan, N.L.: Improving object detection by label assignment distillation. In: WACV (2022)
22. Oksuz, K., Cam, B.C., Akbas, E., Kalkan, S.: A ranking-based, balanced loss function unifying classification and localisation in object detection. Advances in Neural Information Processing Systems **33**, 15534–15545 (2020)
23. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)

24. Qin, Z., Li, Z., Zhang, Z., Bao, Y., Yu, G., Peng, Y., Sun, J.: Thundernet: Towards real-time generic object detection on mobile devices. In: ICCV (2019)
25. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR (2016)
26. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: CVPR (2017)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
28. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In: CVPR (2018)
29. Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. In: arXiv preprint arXiv:1805.00123 (2018)
30. Sun, R., Tang, F., Zhang, X., Xiong, H., Tian, Q.: Distilling object detectors with task adaptive regularization. In: arXiv preprint arXiv:2006.13108 (2020)
31. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully convolutional one-stage object detection. In: ICCV (2019)
32. Wang, J., Song, L., Li, Z., Sun, H., Sun, J., Zheng, N.: End-to-end object detection with fully convolutional network. In: CVPR (2021)
33. Wang, R.J., Li, X., Ling, C.X.: Pelee: A real-time object detection system on mobile devices. In: NeurIPS (2018)
34. Wang, T., Yuan, L., Zhang, X., Feng, J.: Distilling object detectors with fine-grained feature imitation. In: CVPR (2019)
35. Yang, Z., Li, Z., Jiang, X., Gong, Y., Yuan, Z., Zhao, D., Yuan, C.: Focal and global knowledge distillation for detectors. In: arXiv preprint arXiv:2111.11837 (2021)
36. Yao, L., Pi, R., Xu, H., Zhang, W., Li, Z., Zhang, T.: G-detkd: Towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation. In: ICCV (2021)
37. Zhang, L., Ma, K.: Improve Object Detection with Feature-based Knowledge Distillation: Towards Accurate and Efficient Detectors. In: ICLR (2021)
38. Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z.: Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: CVPR (2020)
39. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-iou loss: Faster and better learning for bounding box regression. In: AAAI (2020)
40. Zheng, Z., Ye, R., Wang, P., Wang, J., Ren, D., Zuo, W.: Localization distillation for object detection. In: arXiv preprint arXiv:2102.12252 (2021)
41. Zhixing, D., Zhang, R., Chang, M., Liu, S., Chen, T., Chen, Y., et al.: Distilling object detectors with feature richness. In: NeurIPS (2021)
42. Zhu, B., Wang, J., Jiang, Z., Zong, F., Liu, S., Li, Z., Sun, J.: Autoassign: Differentiable label assignment for dense object detection. In: arXiv preprint arXiv:2007.03496 (2020)