Dive Deeper Into Box for Object Detection

Ran Chen[1], Yong Liu[2], Mengdan Zhang[2], Shu Liu[3],
Bei Yu[1], and Yu-Wing Tai[4]

[1] The Chinese University of Hong Kong
[2] Tencent Youtu Lab
[3] SmartMore
[4] The Hong Kong University of Science and Technology

Abstract. Anchor free methods have defined the new frontier in state-of-the-art object detection researches where accurate bounding box estimation is the key to the success of these methods. However, even the bounding box has the highest confidence score, it is still far from perfect at localization. To this end, we propose a box reorganization method (DDBNet), which can dive deeper into the box for more accurate localization. At the first step, drifted boxes are filtered out because the contents in these boxes are inconsistent with target semantics. Next, the selected boxes are broken into boundaries, and the well-aligned boundaries are searched and grouped into a sort of optimal boxes toward tightening instances more precisely. Experimental results show that our method is effective which leads to state-of-the-art performance for object detection.

1 Introduction

Object detection is an important task in computer vision, which requires predicting a bounding box of an object with a category label for each instance in an image. State-of-the-art techniques can be divided into either anchor-based methods [1,5,7–9,19,21–23] and anchor-free methods [3,12,20,27,30,32]. Recently, the anchor-free methods have increasing popularity over the anchor-based methods in many applications and benchmarks [2,4,6,17]. Despite the success of anchor-free methods, one should note that these methods still have limitations on their accuracy, which are bounded by the way that the bounding boxes are learned in an atomic fashion. Here, we discuss two concerns of existing anchor-free methods which lead to the inaccurate detection.

First, the definition of center key-points [3] is inconsistent with their semantics. As we all know that center key-point is essential for anchor-free detectors. It is a common strategy to embed positive center key-points inside an object bounding box into a Uniform or Gaussian distribution in the training stage of the anchor-free detectors such as FCOS [27] and CornerNet [14]. However, it is inevitable to falsely consider noisy pixels from background as positives, as illustrated in Fig. 1. Namely, exploiting a trivial strategy to define positive targets would lead to a significant semantic inconsistency, degrading the regression accuracy of detectors.

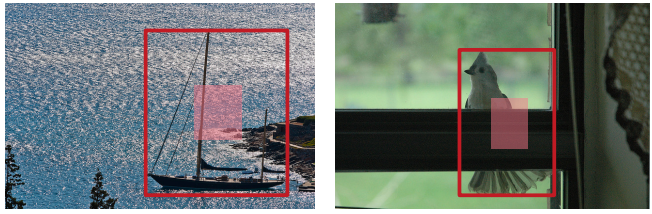


Fig. 1: **An illustration of the inconsistency between the semantics of center key-points inside a bounding box and their annotations.** Pixels of backgrounds in the red central area are considered as positive center key-points, which is incorrect.

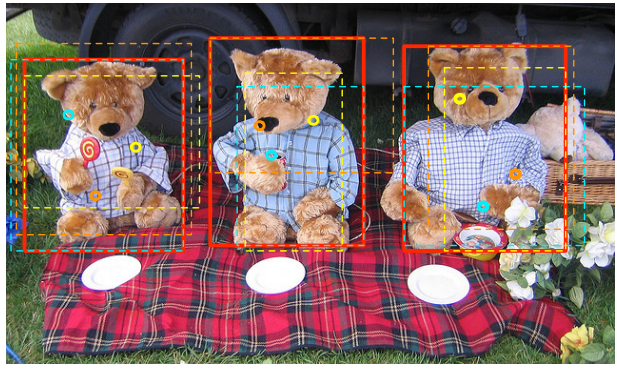


Fig. 2: **An illustration of the boundary drifts in box predictions of general anchor-free detectors.** Limited by regional receptive fields and the design of treating each box prediction as an atomic operation in general detectors, each predicted box with dotted line is imperfect where four boundaries are not well aligned to the ground truth simultaneously. After box decomposition and combination, the reorganized box with red color gets better localization.

Second, local wise regression is limited. Concretely, a center key-point usually provides box predictions in a regional/local-wise manner, which potentially defects the detection accuracy. The local-wise prediction results from the limitation of the receptive fields of convolution kernels, and the design of treating each box prediction from each center key-point as an atomic operation. As shown in Fig. 2, the dotted predicted box and corresponding center key-point are presented in the same color. Although each predicted box is surrounding the object, it is imperfect because four boundaries are not well aligned to the ground truth simultaneously. As a result, choosing a box of high score at inference stage as the final detection result is sometimes inferior.

To tackle the inaccurate detection problem, we present a novel bounding box reorganization method, which dives deeper into box regressions of center key-points and takes care of semantic consistencies of center key-points. This

reorganization method contains two modules, denoted as box decomposition and recombination (D&R) module and semantic consistency module. Specifically, box predictions of center key-points inside an instance form an initial coarse distribution of the instance localization. This distribution is not well aligned to the ideal instance localization, and boundary drifts usually occur. The D&R module is proposed to firstly decompose these box predictions into four sets of boundaries to model an instance localization at a lower refined level, where the confidence of each boundary is evaluated according to the deviation with ground-truth. Next, these boundaries are sorted and recombined to form a sort of more accurate box predictions for each instance, as described in Fig. 2. Then, these refined box predictions contribute to the final evaluation of box regressions.

Meanwhile, the semantic consistency module is proposed to rule out noisy center key-points coming from the background, which allows our method to focus on key-points that are strongly related to the target instance semantically. Thus, box predictions from these semantic consistent key-points can form a more tight and robust distribution of the instance localization, which further boosts the performance of the D&R module. Our semantic consistency module is an adaptive strategy without extra hyper-parameters for predefined spatial constraints, which is superior to existing predefined strategies in [27,28,33].

The main contribution of this work lies in the following aspects.

- We propose a novel box reorganization method in a unified anchor free detection framework. Especially, a D&R module is proposed to take the boundary prediction as an atomic operation, and then reorganize well-aligned boundaries into boxes in a bottom-up fashion with negligible computation overhead. To the best of our knowledge, the idea of breaking boxes into boundaries for training has never been investigated in this task.
- We evaluate the semantic inconsistency between center key-points inside an instance and the annotated labels, which helps boost the convergence of a detection network.
- The proposed method DDBNet obtains a state-of-the-art result of 45.5% in AP. The stable experimental results in all metrics ensure that this method can be effectively extended to typical anchor free detectors.

2 Related Work

Anchor based Object Detectors. In anchor-based detectors, the anchor boxes can be viewed as pre-defined sliding windows or proposals, which are classified as positive or negative samples, with an extra offsets regression to refine the prediction of bounding boxes. The design of anchor boxes is popularized by two-stage approaches such as Faster R-CNN in its RPNs [23], and single-stage approaches such as SSD [19], RetinaNet [16], and YOLO9000 [21], which has become the convention in a modern detector. Anchor boxes make the best use of the feature maps of CNNs and avoid repeated feature computation, speeding up the detection process dramatically. However, anchor boxes result in excessively too many hyper-parameters that are used to describe anchor shapes

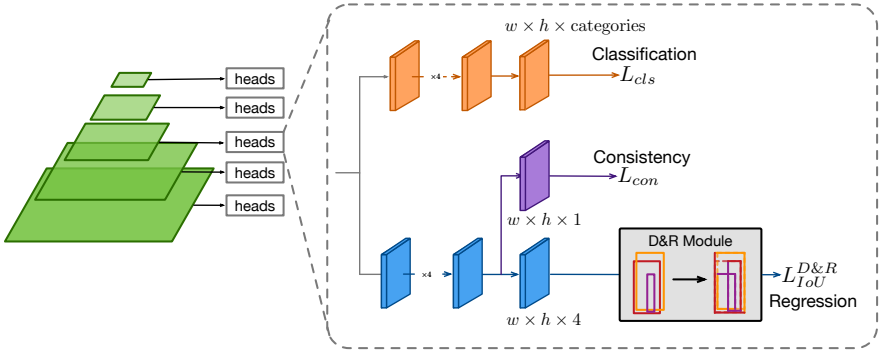


Fig. 3: **An illustration of our network architecture.** Two novel components: the D&R module and the consistency module are incorporated into a general detection network. The D&R module carries out box decomposition and recombination in the training stage regularized by the IoU loss and predicts boundary confidences supervised by the boundary deviation. The consistency module selects meaningful pixels whose semantics is consistent with the instance to improve network convergence in the training stage.

or to label each anchor box as a positive, ignored or negative sample. These hyper-parameters have shown a great impact on the final accuracy, and require heuristic tuning.

Anchor Free Object Detectors. Anchor-free detectors directly learn the object existing possibility and the bounding box coordinates without anchor reference. DenseBox [11] is a pioneer work of anchor-free based detectors. While due to the difficulty of handling overlapping situations, it is not suitable for generic object detection.

One successful family of anchor free works [13, 27, 28, 33] adopts the Feature Pyramid network [15] (FPN) as the backbone network and applies direct regression and classification on multi-scale features. These methods treat the bounding box prediction as an atomic task without any further investigations, which bounds the detection accuracy due to the two concerns we discussed in the introduction. To avoid the drawback of anchors and refine the box presentations, points based box representation becomes popular recently [3, 14, 30, 32]. For example, CornerNet [14] predicts the heatmap of corners and apply an embedding method to group a pair of corners that belong to the same object. [32] presents a bottom-up detection framework inspired by the keypoint estimations. Compared to these points based methods, our proposed method has following innovations: 1) Our method focuses on the mid-level boundary representations to achieve a balance between accuracy and robustness of feature modeling; 2) Our method does not need to learn an embedding explicitly while obtaining a reliable boundary grouping to produce the final bounding box predictions.

Furthermore, it is observed that anchor-free methods may produce a number of low-quality predicted bounding boxes at locations that are far from the

center of a target object. In order to suppress these low-quality detections, a novel "centerness" branch to predict the deviation of a pixel to the center of its corresponding bounding box is exploited in FCOS [27]. This score is then used to down-weight low-quality detected bounding boxes and merge the detection results in NMS. FoveaBox [13] focuses on the object's center motivated by the fovea of human eyes. It is reasonable to degrade the importance of pixels close to boundaries, but the predefined center field may not cover all cases in the real world, as shown in Fig. 1. Thus, we propose an adaptive consistency module to solve the inconsistency issue mentioned above between the semantics of pixels inside an instance and the predefined labels or scores.

3 Our Approach

In this work, we build DDBNet based on FCOS as a demonstration, which is an advanced anchor-free method. As shown in Fig. 3, our innovations lie in the box decomposition and recombination (D&R) module and the semantic consistency module.

To be specific, the D&R module reorganizes the predicted boxes by breaking them into boundaries for training which is concatenated behind the regression branch. In the training stage, once bounding box predictions are regressed at each pixel, the D&R module decomposes each bounding box into four directional boundaries. Then, boundaries of the same kind are ranked by their actual boundary deviations from the ground-truth. Consequently, by recombining ranked boundaries, more accurate box predictions are expected, which are then optimized by the IoU loss [31].

As for the semantic consistency module, a new branch of estimating semantic consistency instead of centerness is incorporated into the framework, which is optimized under the supervision of the semantic consistency module. This module exploits an adaptive filtering strategy based on the outputs of the classification and the regression branches. More details about the two modules are provided in the following subsections.

3.1 Box Decomposition and Recombination

Given an instance I, every pixel i inside of I regresses a box $p_i = \{l_i, t_i, r_i, b_i\}$. The set of predicted boxes is denoted as $B_I = \{p_0, p_1, \ldots, p_n\}$, where l, t, r, b denote the left, the top, the right, and the bottom boundaries respectively.

Normally, an IoU regression loss is expressed as

$$L_{IoU} = -\frac{1}{N_{pos}} \sum_{I} \sum_{i}^{n} \log(IoU(p_i, p_I^*)), \tag{1}$$

where N_{pos} is the number of positive pixels of all instances, p_I^* is the regression target. Simply, the proposed box decomposition and recombination (D&R) module is designed to reproduce more accurate p_i with the optimization of IoU loss.

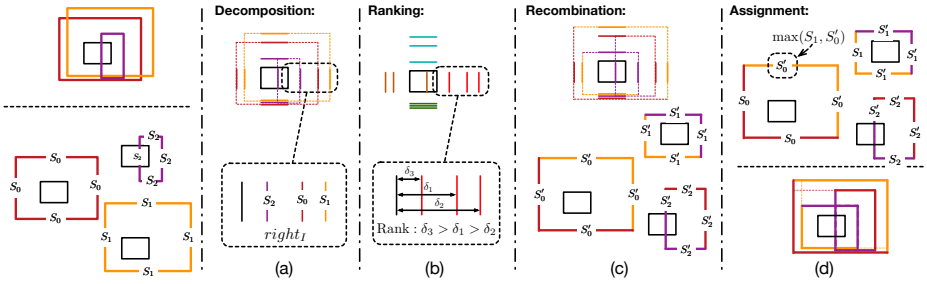


Fig. 4: **An illustration of the work flow of the D&R module.** For a clear visualization, only three predictions in color are provided for the same ground-truth shown in black. (a) **Decomposition:** Break up boxes and assign IoU scores S_0, S_1, S_2 of boxes to boundaries as confidence. (b) **Ranking:** The rule how we recombine boundaries to new boxes. (c) **Recombination:** Regroup boundaries as new boxes and assign new IoU scores S_0', S_1', S_2' to boundaries as confidence. (d) **Assignment:** Choose the winner confidence as final result. The recombined box is shown on the right.

As shown in Fig. 4, the D&R module consists of four steps before regularizing the final box predictions based on the IoU regression. More details are described as follows.

Decomposition: A predicted box p_i is split into boundaries l_i, t_i, r_i, b_i and the IoU s_i between p_i and p_I^* is assigned as the confidences of four boundaries, as shown in Fig. 4(a). For instance I, the confidences of boundaries is denoted as a $N \times 4$ matrix S_I. Then we group four kinds of boundaries into four sets, which are $left_I = \{l_0, l_1, ..., l_n\}$, $right_I = \{r_0, r_1, ..., r_n\}$, $bottom_I = \{b_0, b_1, ..., b_n\}$, $top_I = \{t_0, t_1, ..., t_n\}$.

Ranking: Considering the constraint of the IoU loss [31], where the larger intersection area of prediction boxes with smaller union area is favored, the optimal box prediction is expected to have the lowest IoU loss. Thus, traversing all the boundaries of the instance I to obtain the optimal box rearrangement B_I' is an intuitive choice. However, in this way, the computation complexity is quite expensive, which is $\mathcal{O}(n^4)$. To avoid the heavy computation brought by such brute force method, we apply a simple and efficient ranking strategy. For each boundary set of instance I, the deviations δ_I^l, δ_I^r, δ_I^b, δ_I^t to the targets boundary $p_I^* = \{l_I, r_I, b_I, t_I\}$ are calculated. Then, boundaries in each set are sorted by the corresponding deviations, as shown in Fig. 4(b). The boundary closer to the ground-truth has the higher rank than the boundary farther. We find that this ranking strategy works well and the ranking noise does not affect the stability of the network training.

Recombination: As shown in Fig. 4(c), boundaries of four sets with the same rank are recombined as a new box $B_I' = \{p_0', p_1', \ldots, p_n'\}$. Then the IoU s_i' between p_i' and p_I^* is assigned as the recombination confidence of four boundaries. The

confidences of recombination boundaries is expressed as matrix S'_I with shape $N \times 4$.

Assignment: Now we get two sets of boundaries scores S_I and S'_I. As described as Fig. 4(d), the final confidence of each boundary is assigned using the higher score within S_I and S'_I instead of totally using S'_I. This assignment strategy results from the following case, *e.g.* the recombined low-rank box contains boundaries far away from the ground-truth. Then, the confidences s'_i of four boundaries after recombination are much lower than their original one s_i. The severely drifted confidence scores lead to unstable gradient back-propagation in the training stage.

Thus, for reliable network training, each boundary is optimized under the supervision of the IoU loss estimated based on the ground-truth and the optimal box with its corresponding better boundary score. Especially, our final regression loss consists of two parts:

$$\begin{aligned} L_{IoU}^{D\&R} = & \frac{1}{N_{pos}} \sum_{I} (\mathbb{1}_{\{S'_I > S_I\}} L_{IoU}(B'_I, T_I) \\ & + \mathbb{1}_{\{S_I \geqslant S'_I\}} L_{IoU}(B_I, T_I)), \end{aligned} \tag{2}$$

where $\mathbb{1}_{\{S_I \geqslant S'_I\}}$ is an indicator function, being 1 if the original score is greater than the recombined one, vice versa for $\mathbb{1}_{\{S'_I > S_I\}}$. The gradient of each boundary is selected to update network according to the higher IoU score between the original box and the recombined box. Compared to the original IoU loss Equation (1) where gradients are back-propagated in local receptive fields, Equation (2) updates the network in context without extra parameterized computations. As box in B'_I is composed by boundaries from different boxes, features are updated in an instance-wise fashion. Note that there are no further parameters added in D&R module. In short, we only change the way how gradient be updated.

3.2 Semantic Consistency Module

Since the performance of our D&R module to some extent depends on the box predictions of dense pixels inside an instance, an adaptive filtering method is required to help the network learning focus on positive pixels while rule out negative pixels. Namely, the labeling space of pixels inside an instance is expected to be consistent with their semantics. Different from previous works [13, 27, 28] which pre-define pixels around the center of the bounding box of an instance as the positive, our network evolves to learn the accurate labeling space without extra spatial assumptions in the training stage.

The formula of semantic consistency is expressed as:

$$\begin{cases} \overline{C_{I\downarrow}} \bigcap \overline{R_{I\downarrow}} \leftarrow \text{negative}, \\ \overline{C_{I\uparrow}} \bigcup \overline{R_{I\uparrow}} \leftarrow \text{positive}, \\ c_i = \max_{j=0}^{g}(c_j) \in C_I, \end{cases} \tag{3}$$

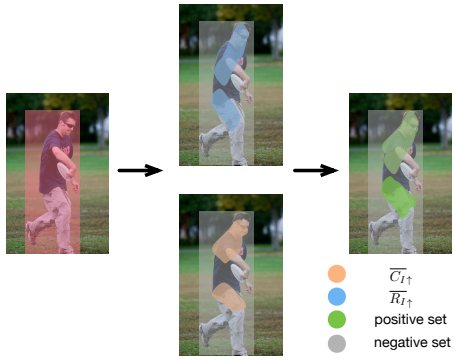


Fig. 5: **Visualized example of semantic consistency module.** The intersection regions of positive regression and positive classification sets are regarded as consistent targets.

where R_I is the set of IoU scores between the ground-truth and the predicted boxes of pixels inside the instance I, $\overline{R_I}$ is the mean IoU score of the set R_I, $\overline{R_{I\downarrow}}$ denotes pixels which have lower IoU confidence than the mean IoU $\overline{R_I}$. Inversely, $\overline{R_{I\uparrow}}$ denotes pixels which have higher IoU confidence than $\overline{R_I}$. The element $c_i \in C_I$ is the maximal classification score among all categories of the i-th pixel, and g denotes the number of categories. Similarly, $\overline{C_{I\downarrow}}$ denotes pixels which have lower classification scores than the mean score of C_I. Labels of categories are agnostic in this approach so that the predictions of incorrect categories will not be rejected during training. Finally, as shown in Fig. 5, the intersection pixels in $\overline{R_{I\downarrow}}$ and $\overline{C_{I\downarrow}}$ are assigned negative, while the union pixels in $\overline{R_{I\uparrow}}$ and $\overline{C_{I\uparrow}}$ are assigned positive. Meanwhile, if pixels are covered by multiple instances, they prefer to represent the smallest instance.

More to the point, the filtering method determined by Equation (3) is able to adaptively control the ratio of positive and negative pixels of instances with different sizes during the training stage, which have a significantly effect on the detection capability of the network. In the experiments, we investigate the performance of different fixed ratio, and then find that the adaptive selection by mean threshold performs best.

After the labels of pixels are determined autonomously according to the semantic consistency, the inner significance of each positive pixel is considered in the learning process of our network, similarly to the centerness score in FCOS [27]. Thus, our network is able to emphasize on more important part of an instance and is learnt more effectively. Especially, the inner significance of each pixel is defined as the IoU between the predicted box and the ground-truth. Then, an extra branch of estimating the semantic consistency of each pixel is added to the network supervised by the inner significance. The loss for semantic consistency is expressed as in Equation (4), where r_i is the output of semantic

consistency branch. $IoU(p_i, p_I^*)$ denotes the inner significance of each pixel.

$$L_{con} = \frac{1}{N_{pos}} \sum_{I} \sum_{i \in \overline{C_{I\uparrow}} \bigcup \overline{R_{I\uparrow}}} CE(r_i, IoU(p_i, p_I^*)). \tag{4}$$

Generally, the overall training loss is defined as:

$$L = L_{cls} + L_{reg}^{D\&R} + L_{con}, \tag{5}$$

where L_{cls} is the focal loss as in [16].

4 Experiments

4.1 Experimental Setting

Dataset. Our method is comprehensively evaluated on a challenging COCO detection benchmark [17]. Following the common practice of previous works [14,16,27], the COCO *trainval35k* split (115K images) and the *minival* split (5K images) are used for training and validation respectively in our ablation studies. The overall performance of our detector is reported on the *test-dev* split and is evaluated by the server.

Network Architecture. As shown in Fig. 3, Feature Pyramid Network (FPN) [15] is exploited as the fundamental detection network in our approach. The pyramid is constructed with the levels $P_l, l = 3, 4, ..., 7$ in this work. Note that each pyramid level has the same number of channels (C), where $C = 256$. At the level P_l, the resolution of features is down-sampled by 2^l compared to the input size. Please refer to [15] for more details. Note that four heads are attached to each layer of FPN. Apart from the regression and classification heads, a head for semantic consistency estimation is provided, consisting of a normal convolutional layer. The regression targets of different layers are assigned in the same way as in [27].

Training Details. Unless specified, all ablation studies take ResNet-50 as the backbone network. To be specific, the stochastic gradient descent (SGD) optimizer is applied and our network is trained for 12 epochs over 4 GPUs with a minibatch of 16 images (4 images per GPU). Weight decay and momentum are set as 0.0001 and 0.9 respectively. The learning rate starts at 0.01 and reduces by the factor of 10 at the epoch of 8 and 11 respectively. Note that the ImageNet pre-trained model is applied for the network initialization. For newly added layers, we follow the same initialization method as in RetinaNet [16]. The input images are resized to the scale of 1333×800 as the common convention. For comparison with state-of-the-art detectors, we follow the setting in [27] that the shorter side of images in the range from 640 to 800 are randomly scaled and the training epochs are doubled to 24 with the same reduction at epoch 16 and 22.

Inference Details. At post-processing stage, the input size of images are the same as the one in training. The predictions with classification scores $s > 0.05$ are selected for evaluation. With the same backbone settings, the inference speed of DDBNet is same as the detector in FCOS [27].

Table 1: **Comparison with state-of-the-art two stage and one stage Detectors** (*single-model and single-scale results*). DDBNet outperforms the anchor-based detector [16] by 2.9% AP with the same backbone. Compared with anchor-free models, DDBNet is in on-par with these state-of-the-art detectors. † means the NMS threshold is 0.6 and others are 0.5.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods:							
Faster R-CNN w/ FPN [15]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN w/ TDM [25]	Inception-ResNet-v2-TDM [26]	36.8	57.7	39.2	16.2	39.8	52.1
Faster R-CNN by G-RMI [10]	Inception-ResNet-v2	34.7	55.5	36.7	13.5	38.1	52.0
RPDet [30]	ResNet-101-DCN	42.8	65.0	46.3	24.9	46.2	54.7
Cascade R-CNN [1]	ResNet-101	42.8	62.1	46.3	23.7	45.5	55.2
One-stage methods:							
YOLOv2 [21]	DarkNet-19 [21]	21.6	44.0	19.2	5.0	22.4	35.5
SSD [19]	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
DSSD [5]	ResNet-101	33.2	53.3	35.2	13.0	35.4	51.1
FSAF [33]	ResNet-101	40.9	61.5	44.0	24.0	44.2	51.3
RetinaNet [16]	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	53.9
CornerNet [14]	Hourglass-104	40.5	56.5	43.1	19.4	42.7	53.9
ExtremeNet [32]	Hourglass-104	40.1	55.3	43.2	20.3	43.2	53.1
FCOS† [27]	ResNet-101-FPN	41.5	60.7	45.0	24.4	44.8	51.6
FCOS† [27]	ResNeXt-64x4d-101-FPN	43.2	62.8	46.6	26.5	46.2	53.3
FCOS† w/improvements [27]	ResNeXt-64x4d-101-FPN	44.7	64.1	48.4	27.6	47.5	55.6
DDBNet (Ours)	ResNet-101-FPN	42.0	61.0	45.1	24.2	45.0	53.3
DDBNet (Ours)	ResNeXt-64x4d-101-FPN	43.9	63.1	46.7	26.3	46.5	55.1
DDBNet (Ours)§	ResNeXt-64x4d-101-FPN	**45.5**	**64.5**	**48.5**	**27.8**	**47.7**	**57.1**

§ GIoU [24] and Normalization methods of 'improvements' proposed in FCOS [27] are applied, ctr.sampling in 'improvements' [27] are not compatible with our setting and we do not use.

4.2 Overall Performance

We compare our model denoted as DDBNet with other state-of-the-art object detectors on the *test-dev* split of COCO benchmark, as listed in Table 1. Compared to the anchor-based detectors, our DDBNet shows its competitive detection capabilities. Especially, it outperforms RetinaNet [16] by 2.9% AP. When it comes to the anchor-free detectors, especially detectors such as FCOS [27] and CornerNet [14] benefiting from the point-based representations, our DDBNet achieves performances gains of 0.5% AP and 1.5% AP respectively. Based on the ResNeXt-64x4d-101-FPN backbone [29], DDBNet works better than [27] with a 0.7% AP gain. Especially for large objects, our DDBNet gets 55.1% AP, better than 53.3 % reported in FCOS [27]. We also apply part of 'improvement' methods proposed in FCOS to DDBNet and gets 0.8% better than the FCOS with all 'improvements' applied. To sum up, compared to detectors exploiting point-based representations, our DDBNet can similarly benefit from the mid-level boundary representations without heavy computation burdens. Furthermore, DDBNet is compared to several two stage models. It overpasses [15] by a large margin.

Table 2: **Ablative experiments for DDBNet on the COCO *minival* split.** We evaluate the improvements brought by the Box Decomposition and Recombination(D&R) module and the semantic consistency module.

Modules			AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	D&R	Consistency						
✓			33.6	53.1	35.0	18.9	38.2	43.7
✓	✓		34.8	54.0	36.4	19.7	39.0	44.9
✓		✓	37.2	55.4	39.5	21.0	41.7	48.6
✓	✓	✓	**38.0**	**56.5**	**40.8**	**21.6**	**42.4**	**50.4**

4.3 Ablation Study

In this section, we explore the effectiveness of our method, including two main modules of box D&R module and semantic consistency module. Additionally, we conduct in-depth analysis of the performance metrics of our method.

4.3.1 Comparison with Baseline Detector

It should be noted that FCOS detector [27] without the centerness branch in both training and inference stages is taken as our baseline. Here we conduct in-depth analysis of the performance metrics of our method.

Box D&R module. As shown in Table 2, by incorporating the D&R module into the baseline detector, a 1.2% AP gain is obtained, which proves that our D&R module can boost the overall performance of the detector. Especially for the AP_{75}, a 1.4% improvement is achieved, which means that D&R performs better on localization even in a strict IOU threshold. Furthermore, D&R module achieves a better performance on large instances according to the large gain on AP_L. With explicit boundary analysis, large instances are often surrounded by numbers of predicted boxes. As a result, it gets easier to find the well-aligned boundaries, then the boxes re-organization can be more effective. Compared to the baseline results in metrics including AP_{50}, AP_S and AP_M, D&R obtains stable performance gains respectively, which shows the stability of our proposed module. By breaking the atomic boxes into boundaries, D&R module makes each boundary find the better optimization direction. The optimization of boundary is not limited by the box its in, instead of depending on a sorted of related boxes. Generally, by adjusting the boundary optimization, the detection network is learnt better.

Semantic Consistency module. The semantic consistent module described in Section 3.2 presents an adaptive filtering method. It forces our detection network into autonomously focusing on positive pixels whose semantics are consistent with the target instance. As shown in Table 2, the semantic consistency module contributes to a significant performance gain of 3.6% AP compared to the baseline detector. This variant surpasses the baseline by large margins in all metrics. Due to that the coarse bounding boxes would contain backgrounds and distractors inevitably, the network is learnt with less confusion about the targets

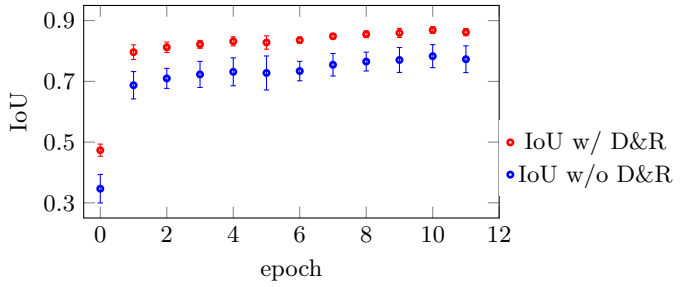


Fig. 6: **Average IoU scores for all predicted boxes during the training.** The red points denote the IoU scores with D&R module while the blue points are the IoU scores without optimization. Vertical lines indicate the variance of IoU scores.

when equips our adaptive filtering module. More ablation analysis on semantic consistency module is provided in Section 4.3.2.

Cooperation makes better. In our final model denoted as DDBNet, the semantic consistency module first filters out a labeling space of pixels inside each instance that is strongly relative to the geometric and semantic characteristics of the instance. The box predictions of the filtered positive pixels are further optimized by the D&R module, leading to more accurate detection results. Consequently, DDBNet achieves 38% AP, better than all the variants in Table 2. Our method boosts detection performance over the baseline by 2.7%, 4.2%, and 6.7% respectively on AP_S, AP_M, AP_L.

4.3.2 Analysis on D&R Module.

Statistical comparison with conventional IoU Loss. As we mentioned in Section 3.1, IoU loss with D&R updates the gradient according to the optimal boundary scores. To confirm the stability of D&R module, we plot the average IoU scores and variances of boxes before and after D&R respectively. We can see that with D&R module, the average values of IoU scores are higher than the means of origin IoU scores by a large margin around 10% in the whole training schedule, as in Fig. 6. At the start of training, the mean of optimal boxes gets 0.47 which is better than 0.34 of origin boxes. As training goes on, both average scores of origin and optimal boxes increase and remain at 0.77 and 0.86 at the end. Variances of IoU scores with D&R are much lower than the origin IoU scores, which indicates D&R module improves the overall quality of boxes and provides better guidance for training.

Visualization on D&R module. We provide some qualitative results of box predictions before and after incorporating the D&R module into the baseline detector, as shown in Fig. 7. For clear visualization, we plot origin boxes and boxes after recombination individually. Predictions are presented in green and the lighter colors indicate higher IoU scores. With D&R module, boundaries are recombined together to obtain a tighter box of each instance. The distribution of

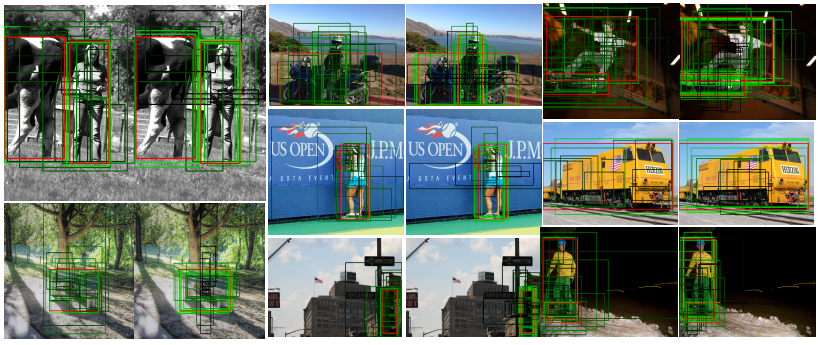


Fig. 7: **Illustration of improved box predictions provided by our DDBNet.** We visualize the boxes before the decomposition (left images of the pairs) and the boxes after the recombination (right images of the pairs). **Red**: ground-truth boxes. **Green**: the predictions, where the lighter colors indicate higher IoU scores. **Black**: the boxes with low score, which will be masked according to the regression loss. Boxes ranked by D&R module are much better organized than the origin boxes and the localizations are much correlated to the instances. All the results are from DDBNet with ResNet-50 as backbone on *trainval*35*k* split.

Table 3: **Comparison among different positive assignment strategies.** 'None' means no sampling method is applied. 'PN' denotes as the definition in [13], which means center regions are positive and others are negative. 'PNI' is the sampling used in [28, 33], ignore regions are added between positive and negative. Note that the consistency term is not included in this table.

Settings	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
None	33.6	53.1	35.0	18.9	38.2	43.7
PN	34.2	53.2	36.3	20.8	38.9	44.2
PNI	33.7	53.0	35.5	17.9	38.3	44.1
Ours	**35.3**	**55.4**	**37.1**	**20.9**	**39.6**	**45.9**

boxes after D&R module is fitter than the origin boxes which is robust than the conventional regression. As we mentioned in Section 3.1, there exists recombined low-rank boxes with boundary scores lower than the origin. These boundaries are masked according to the Equation (2).

4.3.3 Analysis on Semantic Consistency

Dynamic or predefined positive assignment. To further show the superiority of dynamic positive assignment in semantic consistency module, we investigate other variants using different predefined strategies mentioned in previous works. FoveaBox [13] (denoted as 'PN') applies center sampling in their experiments to improve the detection performance. This center sampling method defines the central area of a target box based on a constant ratio as positive

Table 4: **Comparison among different ratio settings.** where c is the sampling ration for each instance.

ratios	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
$c = 0.4$	34.6	54.2	36.6	19.1	38.5	45.2
$c = 0.5$	34.1	53.5	35.9	19.2	38.4	44.2
$c = 0.6$	34.7	54.2	36.5	19.0	38.7	45.5
$c = 0.7$	35.1	54.6	**37.1**	19.3	39.1	45.7
mean	**35.3**	**55.4**	**37.1**	**20.9**	**39.6**	**45.9**

while the others as negative. 'PNI' is taken used in [28, 33] which exploits positive, ignore and negative regions for supervised network training. According to the result in Table 3, 'PN' (second line) gets slight improvement compared to the baseline where no sampling method is adopted. So restricting the searching space to the central area makes sense in certain cases and indeed helps improve object detection. But the 'PNI' gets a lower performance, especially on AP_S. Namely, adding an ignore region between the ring of negative areas and the central positive areas does not further improve the performance and gets a large drop on the detection of small objects. The limited number of candidates of small objects and the lower ratio of positive candidates in 'PNI' result in the poor detection capability. Contrastively, our proposed filtering method does not need to pre-define the spatial constraint while show best performances in all metrics.

Adaptive or constant ratio. As mentioned in section 3.2, we investigate the constant ratio to replace the adaptive selection by mean. Four variants are obtained where the constant ratio is set from 0.4 to 0.7. For instance I with M candidates, top $\lfloor c \times M \rfloor$ candidates are considered as positive, and others are negative, where c is the constant sampling ratio applied to all instances. As shown in Table 4, these results indicate that the adaptive way in our method is better than the fixed way to select positives from candidates.

5 Conclusion

We propose an anchor-free detector DDBNet, which firstly proposes the concept of breaking boxes into boundaries for detection. The box decomposition and recombination optimizes the model training by uniting atomic pixels and updating in a bottom-up manner. We also re-evaluate the semantic inconsistency during training, and provide an adaptive perspective to solve this problem universally with no predefined assumption. Finally, DDBNet achieves a state-of-the-art performance with inappreciable computation overhead for object detection.

References

1. Cai, Z., Vasconcelos, N.: Cascade R-CNN: Delving into high quality object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6154–6162 (2018) 1, 10
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: Imagenet: A large-scale hierarchical image database. In: IEEE conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009) 1
3. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: CenterNet: Object detection with keypoint triplets. In: IEEE International Conference on Computer Vision (ICCV). pp. 6569–6578 (2019) 1, 4
4. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015) 1
5. Fu, C.Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: DSSD: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659 (2017) 1, 10
6. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3354–3361 (2012) 1
7. Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision (ICCV). pp. 1440–1448 (2015) 1
8. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 580–587 (2014) 1
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: IEEE International Conference on Computer Vision (ICCV). pp. 2961–2969 (2017) 1
10. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al.: Speed/accuracy trade-offs for modern convolutional object detectors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7310–7311 (2017) 10
11. Huang, L., Yang, Y., Deng, Y., Yu, Y.: DenseBox: Unifying landmark localization with end to end object detection. arXiv preprint arXiv:1509.04874 (2015) 4
12. Jie, Z., Liang, X., Feng, J., Lu, W.F., Tay, E.H.F., Yan, S.: Scale-aware pixelwise object proposal networks. IEEE Transactions on Image Processing (TIP) **25**(10), 4525–4539 (2016) 1
13. Kong, T., Sun, F., Liu, H., Jiang, Y., Shi, J.: FoveaBox: Beyond anchor-based object detector. arXiv preprint arXiv:1904.03797 (2019) 4, 5, 7, 13
14. Law, H., Deng, J.: CornerNet: Detecting objects as paired keypoints. In: European Conference on Computer Vision (ECCV). pp. 734–750 (2018) 1, 4, 9, 10
15. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2117–2125 (2017) 4, 9, 10
16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (2017) 3, 9, 10
17. 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: European Conference on Computer Vision (ECCV). pp. 740–755 (2014) 1, 9
18. 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: European Conference on Computer Vision (ECCV). pp. 740–755 (2014) 19, 20

19. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. In: European Conference on Computer Vision (ECCV). pp. 21–37 (2016) 1, 3, 10
20. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 779–788 (2016) 1
21. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7263–7271 (2017) 1, 3, 10
22. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018) 1
23. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (6), 1137–1149 (2017) 1, 3
24. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union (June 2019) 10
25. Shrivastava, A., Sukthankar, R., Malik, J., Gupta, A.: Beyond skip connections: Top-down modulation for object detection. arXiv preprint arXiv:1612.06851 (2016) 10
26. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI Conference on Artificial Intelligence (AAAI) (2017) 10
27. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: Fully convolutional one-stage object detection. In: IEEE International Conference on Computer Vision (ICCV). pp. 9627–9636 (2019) 1, 3, 4, 5, 7, 8, 9, 10, 11, 19, 20
28. Wang, J., Chen, K., Yang, S., Loy, C.C., Lin, D.: Region proposal by guided anchoring. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2965–2974 (2019) 3, 4, 7, 13, 14
29. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1492–1500 (2017) 10
30. Yang, Z., Liu, S., Hu, H., Wang, L., Lin, S.: RepPoints: Point set representation for object detection. In: IEEE International Conference on Computer Vision (ICCV). pp. 9657–9666 (2019) 1, 4, 10
31. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: Unitbox: An advanced object detection network. In: Proceedings of the 24th ACM international conference on Multimedia (MM). pp. 516–520 (2016) 5, 6
32. Zhou, X., Zhuo, J., Krahenbuhl, P.: Bottom-up object detection by grouping extreme and center points. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 850–859 (2019) 1, 4, 10
33. Zhu, C., He, Y., Savvides, M.: Feature selective anchor-free module for single-shot object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3, 4, 10, 13, 14

This is our supplementary material which includes more experiments on semantic consistency and performance analysis to show the effectiveness of our work.

A Analysis on Semantic Consistency Module

We visualize the dynamic consistency at different epoches to see how semantic consistency affects on the learning targets. As shown in Fig. 8, the sampled points in each epoch with both high classification scores among categories and high IoU scores are highlighted, named high consistency samples. The low consistency samples are appear in dark colors. Part of sample points at initial stage is not locate at the instance, as the model is not robust at the beginning. With the semantic consistency module, the learned positive samples are progressively distributed at the semantic area of the instance. As the training going on, high consistency samples become robust and appear in lighter colors. We also evaluate to see how inconsistency problem be solved by our method. Some qualitative results are presented in fig. 9, the typical inconsistency in which center-like annotations cannot handle presented in Fig. 1 are improved to a large extent. By utilizing the segmentation annotations, we found that the proportion of samples locate on background reduced around 15% (from 51.7% at initial to 36.1% when training finished) with the semantic consistency module.

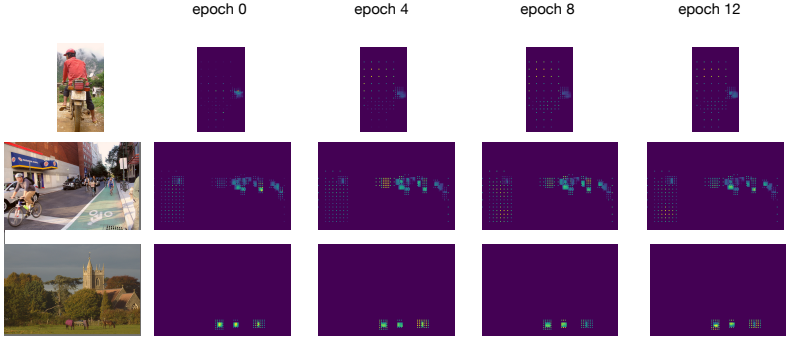
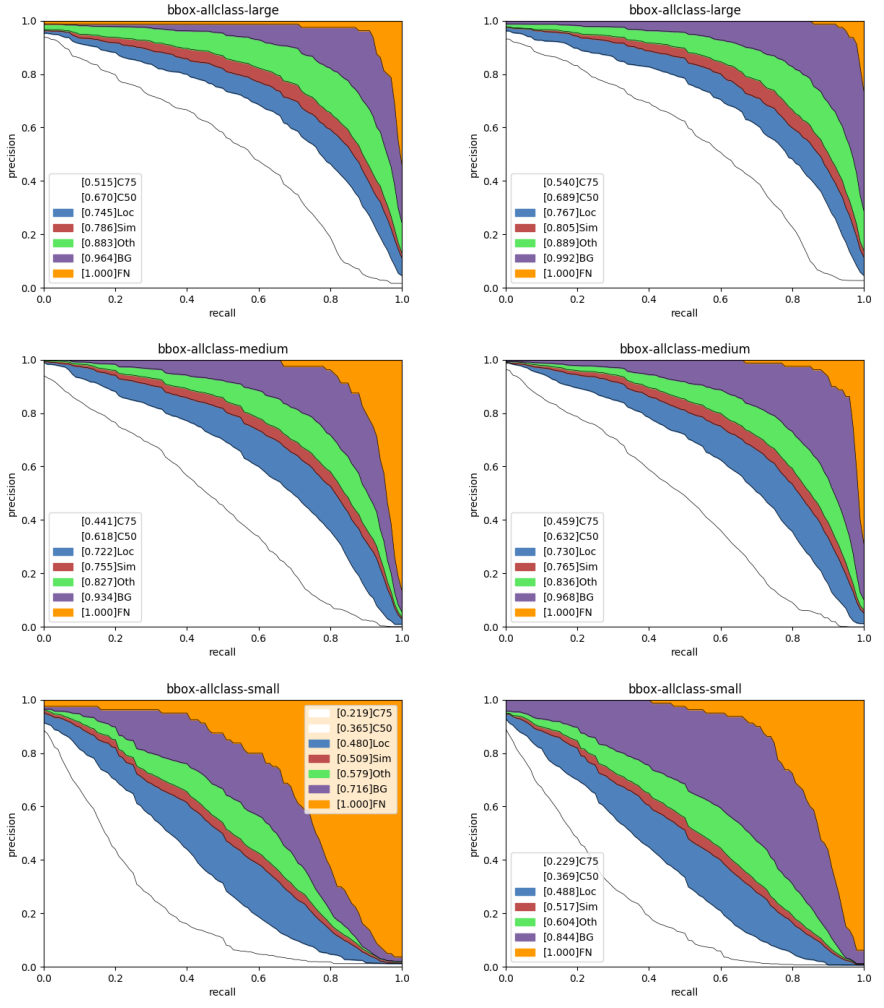


Fig. 8: Visualized examples of semantic consistency module. The left image of each row is the training data select from COCO *trainval35k*. Rest images on the right are the heatmaps of sampled points with semantic consistency module at different training epoch. Note that, entries of heatmaps represent the product results of IoU scores and classification scores. Sampled points with high IoU scores and high classification scores are highlighted in the heatmaps. Sampled points with low IoU scores or low classification scores are in dark colors. *Better viewed in colors and zoom in.*

Fig. 9: Visualization on center inconsistency examples. Sample points with high IoU scores and high classification scores are highlighted. The corresponding areas preferred by the semantic consistency module are marked as red boxes on the images. Images are select from COCO *trainval*35*k* and evaluated with the trained model with ResNet-50 as the backbone.

B Precision-Recall curves

The precision-recall(PR) curves of FCOS [27] and DDBNet under different evaluation settings provided by [18] on the *minival* split are shown in Fig. 10. PR curves were plotted for small-, medium- and large-scale objects in two models. The area in orange indicates the false negative(FN) portion of the evaluated dataset, which can be considered as the PR with all errors removed. The purple area presents the falsely detected objects. We can see that the area of orange in DDBNet is much lower than the one in FCOS [27], which means DDBNet is much robust after all background and class confusions removed.



(e) (f)

Fig. 10: **Precision Recall Curves.** Precision-recall(PR) curves of FCOS [27] and DDBNet under different evaluation settings provided by [18] on the *minival* split with ResNet-50 as backbone. (a)(c)(e): Evaluation results in FCOS. (b)(d)(f): Evaluation results in DDBNet. DDBNet gets better performance under the strict evaluation settings. Especially, we find out that DDBNet works much robust after all background and class confusions removed.