Pseudo-IoU: Improving Label Assignment in Anchor-Free Object Detection

Jiachen Li[1], Bowen Cheng[1], Rogerio Feris[2], Jinjun Xiong[3],
Thomas S. Huang[1], Wen-Mei Hwu[1,4] and Humphrey Shi[1,5,6]
[1]UIUC, [2]MIT-IBM Watson AI Lab, [3]IBM T.J. Watson Research Center,
[4]NVIDIA, [5]University of Oregon, [6]Picsart AI Research (PAIR)

Abstract

Current anchor-free object detectors are quite simple and effective yet lack accurate label assignment methods, which limits their potential in competing with classic anchor-based models that are supported by well-designed assignment methods based on the Intersection-over-Union (IoU) metric. In this paper, we present ***Pseudo-Intersection-over-Union (Pseudo-IoU)****: a simple metric that brings more standardized and accurate assignment rule into anchor-free object detection frameworks without any additional computational cost or extra parameters for training and testing, making it possible to further improve anchor-free object detection by utilizing training samples of good quality under effective assignment rules that have been previously applied in anchor-based methods. By incorporating Pseudo-IoU metric into an end-to-end single-stage anchor-free object detection framework, we observe consistent improvements in their performance on general object detection benchmarks such as PASCAL VOC and MSCOCO. Our method (single-model and single-scale) also achieves comparable performance to other recent state-of-the-art anchor-free methods without bells and whistles. Our code is based on mmdetection toolbox and will be made publicly available at https://github.com/SHI-Labs/Pseudo-IoU-for-Anchor-Free-Object-Detection.*

1. Introduction

Label assignment is very important in training accurate object detection models. In recent years, anchor-based object detection methods have been popular in the community, which starts with Faster RCNN [27]. Typically, during the training process, after feature extraction with deep convolutional neural networks (DCNN), each point on the feature maps is assigned with multiple anchors and each anchor is assigned a positive or negative label based on its overlap with ground truth provided from the dataset. Then, proportional positive and negative samples are selected for training the network and the whole process is defined as label assignment in anchor-based object detection. Since the assignment process selects training samples for regression and classification, improving assignment method could significantly enhance detection results by assigning accurate training samples. Recent work on Cascade RCNN [1] employs cascaded classifiers and regressors with improved Intersection-over-Union (IoU) thresholds to assign better positive training samples. OHEM [29] adopts online hard negative sampling to bring more hard negative samples with larger losses into training. Libra RCNN [24] uses an IoU-based sampling method to achieve a more balanced training.

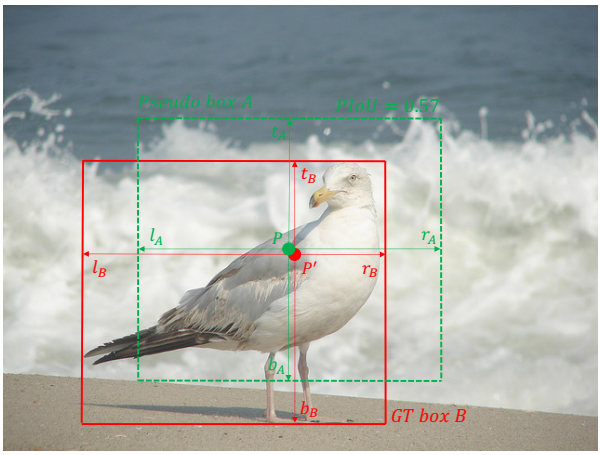

Figure 1: As shown in the image, the red box is the ground truth box and the green box is a pseudo box with the same size to the ground truth box. The green point P and red point $P^{'}$ are the same point from feature map projection. (l, r, t, b) represent distances from a point inside box to the left, right, top and bottom side of the box, respectively. Point P is at center of the pseudo box A and Point $P^{'}$ is in the ground truth box B. The Pseudo-IoU between the two boxes is 0.57.

All these prior works prove that better assignment for training samples is indispensable and effective in anchor-based object detection.

Recently, another popular branch of object detection methods are anchor-free models that do not assume predefined anchors during the whole training process, which reduces many hyper-parameters with anchors that require heuristic tuning for good performance. Anchor-free models predict bounding boxes directly from points to the left, right, top and bottom side of ground truth box like FCOS [32] and FSAF [43]. However, due to lack of accurate assignments, they both use other methods to compensate for the performance gap. For FCOS, it views all points inside shrinked ground truth as positive samples and adds a centerness branch to reweigh detection outputs that decreases some false positives. For FSAF, it employs online feature selection and a combination of anchor-free and anchor-based methods.

In this paper, we propose a Pseudo-Intersection-over-Union (Pseudo-IoU) metric that brings an accurate label assignment rule into current anchor-free object detectors, which is illustrated in Figure 1. For each feature map point inside ground truth boxes, after mapping to the original input image, we assume a corresponding pseudo box that is centered at the point and has the same size to ground truth box. Then we can easily compute IoU between the centered pseudo box and ground truth box. Since the IoU is based on a pseudo box assigned to each point, we name the metric Pseudo-Intersection-over-Union (Pseudo-IoU). After Pseudo-IoU computation, each point can be assigned a Pseudo-IoU value v like each anchor with an IoU for assignment in anchor-based methods. Now each point on the feature map pyramid has a Pseudo-IoU value $v \in [0, 1]$, they can be labeled as positive or negative training samples based on a Pseudo-IoU threshold $T \in [0, 1]$ with,

$$labels = \begin{cases} +1 & if \quad v \geq T \\ -1 & if \quad v < T \end{cases}$$

From Figure 1, it shows some points near the sides of bounding boxes are assigned as negative samples. It leads to more false positives and inaccurate bounding boxes if taking these points with low Pseudo-IoUs as positive samples. To demonstrate the effectiveness of our Pseudo-IoU metric, we build an anchor-free baseline with ResNet-101 [14] as backbone and FPN [19] as neck [3] to make dense anchor-free predictions on enhanced feature map pyramids. Furthermore, we add Pseudo-IoU metric for assignment and conduct extensive experiments on Pascal VOC and MSCOCO dataset. Specifically, the anchor-free methods with Pseudo-IoU based assignment outperforms baseline by 2.1% mean Average Precision (mAP) on Pascal VOC 2007 test set, 3.0% mAP on MSCOCO minival set and 3.1% mAP on MSCOCO test-dev set. It also reaches results comparable to recent state-of-the-art methods without bells and whistles.

To summarize, our contributions are as follows:

- We investigate and analyze the problems and bottlenecks of current anchor-free object detectors, which is lack of accurate label assignment process that is well-defined and designed in anchor-based methods.
- We propose Pseudo-Intersection-over-Union (Pseudo-IoU) metric, which takes accurate label assignment rule into anchor-free framework without any additional cost during training or testing, making it possible to improve anchor-free detection by improving assignment process.
- Our single-scale model improves the performance of baseline by a large margin and achieves performance very comparable to other recent state-of-the-art detection methods.

2. Related Work

Anchor-based Object Detectors. In the past few years, anchor-based object detectors have been the mainstream in the object detection area. It starts with Faster R-CNN [27], a supreme version of previous R-CNN [11] and Fast R-CNN [10], which first proposes the idea of using anchors as pre-defined bounding boxes for subsequent regression to ground truth. Following Faster RCNN, most two-stage object detectors inherit the anchor-based manner, like R-FCN [5], DCN [6], FPN[19], Mask RCNN [13], Cascade RCNN [1] and DCR [4]. Some methods adopt efficient training like SNIP [30] and SNIPER [31], that take clips of images for training. Some approaches improve quality of anchors like GA-RPN [33], which learn the setting of anchors for efficient training. Moreover, some popular one-stage object detectors also employ anchors for better regression, which are proposal free and directly make predictions from anchors to final bounding boxes, like SSD [22], YOLOv2 [25], RetinaNet [20], RefineDet [38] and GHM [18], introduces gradient harmonized loss function for a balanced training process.

Anchor-Free Object Detectors. Different from anchor-based object detectors, another way to predict bounding boxes is to regress directly from points, which starts from Densebox [15] and YOLO [25]. They make prediction by regressing the bounding boxes directly from central points in ground truth area to final bounding boxes. Moreover, with the power of feature pyramid networks, recent paper like FCOS [32] and FSAF [43] build a whole anchor-free detection pipeline, which reach results very comparable to one-stage object detectors like RetinaNet. Another branch is keypoint-based anchor-free object detection,

which makes bounding boxes prediction after keypoint prediction, including Cornernet [17], Centernet [39] and Extremenet [40]. Since the key point prediction is sparse and accurate, post processing like NMS can be removed in these methods.

Assignment Rules. Recently, researchers notice that label assignment is essential for a balanced and stable training process. For anchor-based object detectors, anchors are assigned with positive or negative labels according their IoU with ground truth. To improve quality of positive samples, Cascade RCNN [1] employs increased IoU threshold to select positive samples to train cascaded classifiers. OHEM [29] uses a hard false positive mining branch which select more samples with higher losses as negative training samples. Libra RCNN [24] further proposes an IoU based sampling methods and PISA [2] focuses on prime samples for training. ATSS [37] focuses on bridging gap between anchor-free and anchor-based methods with adaptively selecting positive and negative samples according to statistical characteristics of objects. SAPD [42] uses soft anchor point for label assignment and AutoAssign [41] employs automatic label assignment strategy in anchor-free detectors.

3. Our Approach

In this section, we firstly review assignment process at anchor-based detectors, then propose how we employ Pseudo-IoU metric in anchor-free detectors to make accurate assignment available during training process. Next, we show the whole pipeline of our anchor-free detector with Pseudo-IoU based assignment and introduce each component in the pipeline.

3.1. Pseudo-Intersection-over-Union

Intersection-over-Union (IoU) has been well-defined and applied in anchor-based methods. It is used for comparing similarity between two arbitrary shapes A and B, and the IoU between them is,

$$IoU(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

In anchor-based object detectors like Faster RCNN [27], anchors are pre-defined bounding boxes surrounding with the center of sliding windows, namely points on feature maps. For an anchor A with an area S_A and a ground truth B with an area S_B, their IoU will be

$$IoU(A,B) = \frac{|S_A \cap S_B|}{|S_A \cup S_B|} = \frac{|S_A \cap S_B|}{|S_A| + |S_B| - |S_A \cap S_B|}$$

Typically, an anchor has an IoU overlap more than 0.5 would be assign as a positive sample and has an IoU overlap less than 0.4 would be assign as a negative sample. This

Algorithm 1: Pseudo-Intersection-over-Union

1. Get $l_B, r_B, t_B, b_B, l_A, r_A, t_A, b_A, S_A$ and S_B:

$$l_A = r_A = (l_B + r_B)/2$$
$$t_A = b_A = (t_B + b_B)/2$$
$$S_B = (l_B + r_B) * (t_B + b_B)$$
$$S_A = (l_A + r_A) * (t_A + t_A)$$

2. Intersection box parameters:

$$l_* = min(l_A, l_B) \quad r_* = min(r_A, r_B)$$
$$t_* = min(t_A, t_B) \quad b_* = min(b_A, b_B)$$

3. Intersection box area:

$$|A \cap B| = S_* = (l_* + r_*) * (t_* + b_*)$$

4. Union box area:

$$|A \cup B| = S_A + S_B - S_*$$

5. Compute Pseudo-IoU:

$$Pseudo - IoU = \frac{|A \cap B|}{|A \cup B|} = \frac{S_*}{S_A + S_B - S_*}$$

IoU-based process is adopted by most of two-stage detectors and some one-stage detectors like SSD [22] and RetinaNet [20]. However, for recent anchor-free object detectors, there is no such counterpart during training process. The main reason is that without anchors, the regressor could not regress offsets from anchors to ground truth boxes. On the contrary, it regresses from a point inside ground truth to the four sides to make direct predictions. For example, given a ground truth box B and a point P from feature map that is inside the box, l_B, r_B, t_B and b_B represent distance from point P to the left, right, top and bottom side of box B. we can view the point P as a center point and assign a pseudo box A around it, l_A, r_A, t_A and b_A represent distance from point P to the left, bottom, right and top side of pseudo box A. Then the process to compute IoU between ground truth box B and pseudo box A is described in $Algorithm$ 1.

We name the IoU between the assigned pseudo box and ground truth box Pseudo-Intersection-over-Union, namely Pseudo-IoU, and the valve of Pseudo-IoU is $v \in [0, 1]$, making it possible for selecting training examples with a specific Pseudo-IoU threshold in anchor-free pipelines.

3.2. Anchor-Free Architecture

Most state-of-the-art anchor-free detectors follow the architecture of RetinaNet, which uses a ResNet-101 [14] as backbone and FPN [19] as neck [3], followed by detection heads with two branches: one for image classification with focal loss function, and the other one for bounding boxes

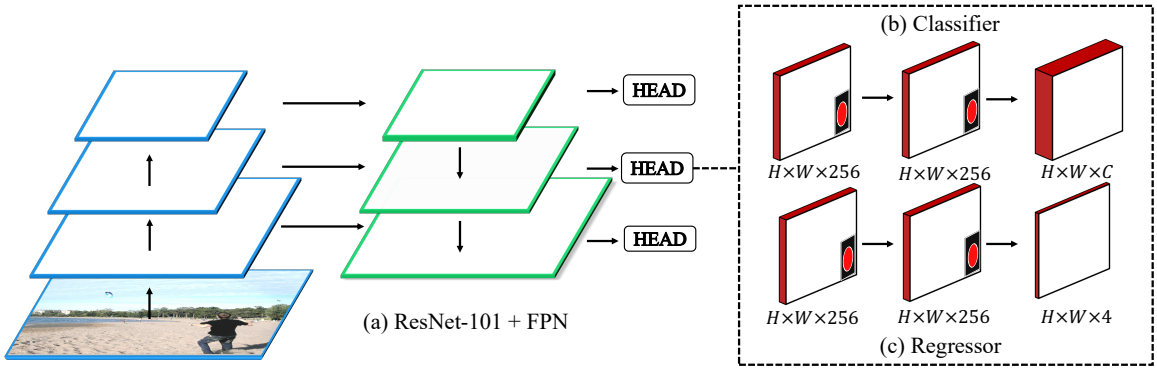

Figure 2: Anchor-free object detector pipeline with Pseudo-IoU based assignment.

regression with smooth L1 loss function. Our anchor-free detection architecture with Pseudo-IoU based assigner is illustrated in Figure 2.

Detection Backbone and Neck Following RetinaNet, we adopt ResNet-101 as our backbone to extract features and use FPN to augment feature maps through top-down pathway and lateral connections, which bulids a rich multi-scale feature pyramids with a single resolution input image and has been proved a strong detection component, as we illustrated in Figure 2 Part(a). Similar to RetinaNet, in the downsampling period, we use 5 stages in ResNet that generates feature maps C_1, C_2, C_3, C_4, C_5 and the downsampling rate is 2^l for C_l feature map. In the upsampling stage, we remove P_2 for less computational burden and use feature maps P_3, P_4, P_5, P_6 and P_7. P_3, P_4 and P_5 are generated from C_3, C_4 and C_5 through upsampling and lateral connections. P_6 and P_7 are computed by stride-2 convolution over P_5 for detection on large targets. All feature maps from P_3 to P_7 have an output channel $C = 256$, which are sent to detection heads for subsequent training process.

Pseudo-IoU based Assignment After feature extraction by ResNet-101 and FPN, we can assign points in feature maps from P_3 to P_7 with positive or negative labels. For a point at feature map P_l, its location can be projected back to original input image with an upsampling rate 2^l. Suppose that its location is (x, y) on the input image and if it falls into a ground truth box with label (x_c, y_c, w, h), (x_c, y_c) is the center of the box and w, h represent its width and height respectively. Then, we can compute l, r, t, b, which are distance from (x, y) to the left, right, top and bottom side of box (x_c, y_c, w, h) by

$$l = x - x_c + \frac{w}{2}, \quad r = x_c + \frac{w}{2} - x$$

$$t = y - y_c + \frac{h}{2}, \quad b = y_c + \frac{h}{2} - y$$

according to the Pseudo-IoU computation algorithm in section 3.1, we can easily have

$$l_B = l, r_B = r, t_B = t, b_B = b$$

then compute Pseudo-IoU following the flow diagram presented in *Algorithm* 1. In anchor-based methods, it usually takes anchors with IoU more than a specific threshold as positive samples, we follow the design and set our Pseudo-IoU threshold T, too. Therefore, all points with Pseudo-IoU larger than T are assigned as positive samples and others are assigned as negative samples. Since focal loss performs good and stable on balanced training and FCOS[32] further proves that one main advantage of anchor-free detectors is all-in samples for training, we use all labeled samples for subsequent training. In Figure 2, for example, on all $H \times W \times 256$ feature maps, there is a black rectangle ground truth area and the red area inside contains all positive samples contribute to the training procedure.

Detection Heads After all feature maps are extracted from FPN and all points on the feature maps are labeled with positiveness or negativeness, the training process moves forward to detection heads part. The detection head is a FCN [23] that attached to each output feature maps from FPN and it contains two subsets: a classifier and a regressor. For the classifier subnet illustrated in Figure 2 part(b), it follows four stacked 3×3 convolution layers with 256 filters and finally attached a 3×3 convolution layer with C filters. C is the number of classes for final classification on each point. For the regressor subnet, in Figure 2 part(c), it also follows four stacked 3×3 convolution layers with 256 filters and finally attached a 3×3 convolution layer with 4 filters per spatial location. At each location, these 4 outputs represent a predicting

bounding box vector (l, r, t, b). Moreover, the two subsets do not share weights on the four consecutive stacked 3×3 convolution layers.

Loss Function The loss function in our anchor-free architecture is

$$L(A_{(x,y)}, B_{(x,y)}) = \frac{1}{N_{pos}} \sum_{(x,y)} L_{fl}(C_{A_{x,y}}, C_{B_{x,y}})$$

$$+ \frac{\lambda}{N_{pos}} \sum_{(x,y)} L_{iou}(R_{A_{x,y}}, R_{B_{x,y}})$$

where L_{fl} is focal loss for classification and L_{iou} is IoU loss [36] for regression. $A_{(x,y)}$ is the set of all selected training samples and $B_{(x,y)}$ is the set of all ground truth boxes. $C_{A_{x,y}}$ and $R_{A_{x,y}}$ represent the output of classifier and regressor. $C_{B_{x,y}}$ and $R_{B_{x,y}}$ represent structured labels of ground truth. N_{pos} is the number of selected positive samples and λ denotes the balance weight for IoU loss.

Inference Process The inference process is straightforward and clear for anchor-free detectors. The input image goes through backbone ResNet-101 and FPN to generate feature maps from P_3 to P_7, then the regressor and classifier make predictions on all points and assemble them together for post processing. Generally, the post processing includes a confidence thresholding and a non maximum suppression (NMS) for final predictions.

4. Experiments

We present all of our experiments mainly on Pascal VOC [8] and MSCOCO [21] benchmark. For Pascal VOC benchmark, we follow common $07 + 12$ practice that uses Pascal VOC 2007 trainval split and 2012 trainval split for training and 2007 test split for testing. For MSCOCO benchmark, we use trainval35k for training, common minival 5k images for validating and test on test-dev set. We first show the training and testing setting of the two benchmarks. Then, we present the effectiveness of Pseudo-IoU in ablation study and compare our model with other state-of-the-art detectors on MSCOCO dataset. All the codes are based on mmdetection [3] toolbox.

4.1. Training and Testing

Pascal VOC We use an ImageNet [7] pretrained ResNet-101[14] as backbone with fixed Batch Normalization [16] layers and build the whole network illustrated in Figure 2 with number of classes $C = 21$ and Group Normalization [34] for stable training. Our network is trained with stochastic gradient descent (SGD) for 12 epochs with a warmup[12] training for 500 iterations. The initial learning rate is 0.01 and reduced by a factor of 10 at epoch 8 and 10, respectively. Weight decay and momentum are set to be 0.0001 and 0.9. The input images size are resized to have their shorter side being 600 and their longer side less or equal to 1000. We only use image flipping as the only data augmentation method since most experiments on Pascal VOC are ablation studies. The model is trained on $4 \times$ 1080 Ti GPU with a total batch size at 16. The loss function is an addition of focal loss and IoU loss, the weight balance parameter of IoU loss is $\lambda = 1$. In the testing process, we only use same single resolution to the training process for inference. For post-processing, the scoring threshold and NMS's threshold are set to be 0.05 and 0.5, respectively.

MSCOCO We build the whole network with a illustrated in Figure 2 with the number of classes $C = 81$ and Group Normalization [34] for stable training. Our network employs ResNet-101 as backbone with feature pyramid networks and is trained with stochastic gradient descent (SGD) for 24 epochs with a warmup training for 1500 iterations. The initial learning rate is 0.01 and reduced by a factor of 10 at epoch 16 and 22, respectively. Weight decay and momentum are set to be 0.0001 and 0.9. The input images are resized to have their shorter side being 800 and their longer side less or equal to 1333 for single-scale training. The model is trained on $4 \times$ V100 GPU with a total batch size at 16. The loss function setting remains unchanged. During the testing process, we only use same single resolution to the training settings for inference with the same post-processing used at Pascal VOC's models.

4.2. Ablation Study

Pseudo-IoU is Essential In Table 1, we carefully present all ablation studies experiments with the effectiveness of Pseudo-IoU on Pascal VOC 2007 test set. The baseline model is in Figure 3 without Pseudo-IoU based assignment, which takes all points inside ground truth as positive samples. Table 1.(a) shows that the baseline model with a 0.4 Pseudo-IoU threshold could get a 2.1% improvement on mAP, which proves the importance of accurate assignment during training process. It also shows that a higher Pseudo-IoU threshold would lead to a performance drop due to insufficient selected positive training samples. In Table 1.(b) and (c), it indicates that anchor-free model with Pseudo-IoU based assignment could bring a consistent around 2% mAP improvement under different training epochs and batchsizes, which shows that the performance gap due to imbalanced assignment of training samples could not be compensated with extended training and larger batchsize.

Assignment with Other Metrics In current state-of-the-art anchor-free object detectors, FCOS [32] and FSAF [43] also achieve detection results comparable to anchor-based

Model	Thrs	mAP
Baseline	0.0	76.9
Pseudo-IoU	0.1	76.8
-	0.2	77.5
-	0.3	78.2
-	**0.4**	**79.0**
-	0.5	78.9
-	0.6	45.6
-	0.7	22.1

(a) Diff Pseudo-IoU Thrs

Model	Epochs	mAP
Baseline	8	76.9
Pseudo-IoU	8	79.0
Baseline	12	78.4
Pseudo-IoU	**12**	**80.4**
Baseline	16	78.3
Pseudo-IoU	16	80.2
Baseline	24	77.8
Pseudo-IoU	24	80.1

(b) Diff Training Epos

Model	Batchsize	mAP
Baseline	4	73.6
Pseudo-IoU	4	75.2
Baseline	6	76.6
Pseudo-IoU	6	77.9
Baseline	8	77.0
Pseudo-IoU	8	78.8
Baseline	16	76.9
Pseudo-IoU	16	**79.0**

(c) Diff Batchsizes

Table 1: Ablation studies results evaluated on Pascal VOC 2007 test set. **Bold** indicates best single-model results and corresponding parameters.

Centerness	mAP	S-box	mAP	P-IoU	mAP
0.2	72.9	0.3	24.5	0.3	78.2
0.25	77.4	0.4	63.8	**0.4**	**79.0**
0.3	78.5	0.5	77.1	0.5	78.9
0.35	**78.7**	**0.6**	**78.3**	0.6	45.6
0.4	77.6	0.7	77.3	0.7	22.1
0.45	77.6	0.8	76.8	0.8	13.5

Table 2: Ablation studies with different assignment rules and all results evaluated on Pascal VOC 2007 test set. **Bold** indicates best single-model results and corresponding parameters.

methods. In FCOS, it adds a centerness branch to reweigh output results. In FSAF, it adopts a central part in ground truth box as positive samples. These two recent work motivate us to bring a Pseudo-IoU based assignment method into anchor-free pipeline. Moreover, there are some other assignment options may be as effective as Pseudo-IoU. We propose other two options adapted from FSAF and FCOS that may be alternatives to our Pseudo-IoU metric. The first one is called scaled-box, which uses all points inside scaled ground truth boxes as positive samples. Given a point (x, y) that falls into a ground box (x_c, y_c, w, h) and a scaled parameter s, the point is positive if it fits that

$$|x - x_c| < \frac{sw}{2} \quad |y - y_c| < \frac{sh}{2}$$

The other one is called centerness, under the same setting to scaled-box, we can compute l, r, t, b according to algorithm presented in section 3.2, then centerness threshold is

$$c = \sqrt{\frac{\min(l, r)}{\max(l, r)} \times \frac{\min(t, b)}{\max(t, b)}}$$

which follows the expression of centerness in FCOS. However, here we use centerness for selecting positive samples rather than reweigh output detection. We compare the detection results on Pascal VOC 2007 test set under these two new metrics and Pseudo-IoU in Table 2. It shows that anchor-free models can achieve different results under different assignment methods. Overall, using Pseudo-IoU could achieve a little better results (79.0% mAP) than centerness (78.7% mAP) and scaled-box (78.3% mAP) metric. More importantly, use Pseudo-IoU based metric may employ previous research on IoU like Cascade RCNN [1] or Libra RCNN [24] into anchor-free manners, which bring more possibilities to anchor-free object detection.

4.3. Comparison with State-of-the-art Detectors

MSCOCO minival set Firstly, we validate our model on MSCOCO minival set. The baseline model employs ResNet-50-FPN backbone with fixed batch normalization layers and an anchor-free detection head that takes all points from feature pyramids inside ground truth as positive samples. The input image is resized to their shorter side being 800 and their longer side less or equal to 1333 for single-scale training. Image flipping is the only strategy for data augmentation. During testing, we only use the same scale setting to the training process and all results are listed in Table 3. Without accurate assignment, there is a performance gap between our anchor-free baseline and RetinaNet. After using Pseudo-IoU based assigment, it outperforms RetinaNet about 1.1% mAP performance and brings 2.9% mAP improvement on baseline model. Moreover, employing GIoU and centerness branch could further improve the performance, which implies that our Pseudo-IoU is compatible with other object detection approaches.

MSCOCO test-dev set We also report our results on

Method	Pseudo-IoU	GIoU	Centerness	AP	AP^{50}	AP^{75}	AP^S	AP^M	AP^L
RetinaNet[20]	-	-	-	35.7	55.0	38.5	18.9	38.9	46.3
AF-Baseline	-	-	-	33.9	54.5	35.3	19.1	38.7	44.1
Pseudo-IoU	✓	-	-	36.8	56.4	39.2	20.4	41.0	48.7
Pseudo-IoU	✓	✓	-	37.1	55.9	39.7	20.4	41.1	49.1
Pseudo-IoU	✓	✓	✓	37.4	56.5	40.1	20.5	41.2	49.4

Table 3: Anchor-free model with Pseudo-IoU based assignment on MSCOCO minival 5k set with ResNet-50-FPN backbone. AF-baseline stands for anchor-free baseline. Pseudo-IoU could bring 2.9% mAP improvement on baseline model. Using GIoU[28] for regression loss and adding a centerness branch like FCOS[32] could both bring further improvements.

Method	Backbone	AP	AP^{50}	AP^{75}	AP^S	AP^M	AP^L
Multi-Stage methods							
Faster RCNN [27]	R-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
R-FCN [5]	Aligned-Inception-R	35.5	55.6	-	17.8	38.4	49.3
Deformable R-FCN [6]	Aligned-Inception-R	37.5	58.0	-	19.4	40.1	52.5
Mask RCNN [13]	R-101-FPN	38.2	60.3	41.7	20.1	41.1	50.2
Libra RCNN [24]	R-101-FPN	41.1	62.1	44.7	23.4	43.7	52.5
Cascade RCNN [1]	R-101-FPN	42.8	62.1	46.3	23.7	45.5	55.2
One-Stage methods							
YOLOv3 [26]	Darknet-53	33.0	57.9	34.4	18.3	25.4	41.9
SSD513 [22]	R-101	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [9]	R-101	33.2	53.3	35.2	13.0	35.4	51.1
RefineDet [38]	R-101	36.4	57.5	39.5	16.6	39.9	51.4
RetinaNet [20]	R-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
GHM [18]	R-101-FPN	39.9	60.8	42.5	20.3	43.6	54.1
Anchors-Free methods							
FSAF [43]	R-101-FPN	40.9	61.5	44.0	24.0	44.2	51.3
FCOS [32]	R-101-FPN	41.5	60.7	45.0	24.4	44.8	51.6
CornetNet [17]	Hourglass-104	40.5	56.5	43.1	19.4	42.7	53.9
CenterNet [39]	Hourglass-104	42.1	61.1	45.9	24.1	45.5	52.8
Our methods							
AF-Baseline	R-101-FPN	38.3	59.2	40.1	22.6	42.1	46.9
Pseudo-IoU	R-101-FPN	41.5	61.0	44.5	24.1	44.6	51.9
Pseudo-IoU	R-101-FPN-DCN	43.4	63.2	46.8	25.8	47.8	56.7
Pseudo-IoU	R-32x8d-FPN-DCN	44.0	63.8	47.3	28.0	47.5	58.1
Pseudo-IoU	R-64x4d-FPN-DCN	44.5	64.4	48.1	26.5	48.8	58.4

Table 4: Detection results of our best single model with Pseudo-IoU based label assignment vs. state-of-the-art one-stage, multi-stage and anchor-free detectors on the MSCOCO test-dev set. R stands for ResNet here.

test-dev set. We train all model with a ResNet-101-FPN backbone and an anchor-free detection head illutrated in Figure 2. The input image is resized to keep the longer side to 1333 and short side randomly selected from 640 to 800 without image flipping or other data augmentation approach. All detection results are listed in Table 4. To make fair comparisons, we only select results with single-model and single-scale inference from their original publications, some results based on multi-scale testing or better backbone like ResNeXt [35] are not included.

4.4. Visualization

We visualize some detection results of both anchor-free baseline and the baseline trained with sampling in Figure 3. It is noticeable that the baseline model has many false positives and inaccurate bounding box predictions. On the contrary, the model adopts sampling based on PIoU produces much less false positives and more accurate localization,

Figure 3: Visualization of some detection results that are best viewed when zoomed in. The first and third rows of images are detection results of anchor-free baseline; the second and fourth rows of images are detection results of anchor-free baseline with sampling based on PIoU metric at 0.5 threshold. As shown from the detection results, our method produces much less false positives and more accurate localization.

which improves performance by a large margin.

5. Conclusion

In this paper, we first point out that current anchor-free object detectors lack accurate and standardized assignment process that limits its potential to compete with many state-of-the-art anchor-based methods. To solve this problem, we propose a simple Pseudo-Intersection-over-Union (Pseudo-IoU) metric that brings more accurate and standardized assignment rule into anchor-free framework by computing IoU between the pseudo box centered on points from feature pyramids and ground truth box. Then, we adopt it as Pseudo-IoU value of each point inside ground truth area. Furthermore, we conduct extensive experiments on PASCAL VOC and MSCOCO dataset, it shows that anchor-free models with Pseudo-IoU based assignment could bring a consistent improvement without bells and whistles. It decreases many false positive detections and leads to more accurate localization of bounding boxes. Moreover, in the future, it is possible that some recent research on improving assignment in anchor-based methods can be transferred to anchor-free models based on the proposed Pseudo-IoU assignment metric.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018.

[2] Yuhang Cao, Kai Chen, Chen Change Loy, and Dahua Lin. Prime sample attention in object detection. *arXiv preprint arXiv:1904.04821*, 2019.

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[4] Bowen Cheng, Yunchao Wei, Honghui Shi, Rogerio Feris, Jinjun Xiong, and Thomas Huang. Revisiting rcnn: On awakening the classification power of faster rcnn. In *ECCV*, pages 453–468, 2018.

[5] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, pages 379–387, 2016.

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[9] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.

[10] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

[12] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[15] Lichao Huang, Yi Yang, Yafeng Deng, and Yinan Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015.

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[17] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, pages 734–750, 2018.

[18] Buyu Li, Yu Liu, and Xiaogang Wang. Gradient harmonized single-stage detector. volume 33, pages 8577–8584, 2019.

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

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

[22] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[24] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *CVPR*, pages 821–830, 2019.

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[26] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[28] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pages 658–666, 2019.

[29] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, pages 761–769, 2016.

[30] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection snip. In *CVPR*, pages 3578–3587, 2018.

[31] Bharat Singh, Mahyar Najibi, and Larry S Davis. Sniper: Efficient multi-scale training. In *NIPS*, pages 9310–9320, 2018.

[32] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. *arXiv preprint arXiv:1904.01355*, 2019.

[33] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *CVPR*, pages 2965–2974, 2019.

[34] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, pages 3–19, 2018.

[35] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

[36] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *ACM Multimedia*, pages 516–520. ACM, 2016.

[37] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, 2019.

[38] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Single-shot refinement neural network for object detection. In *CVPR*, pages 4203–4212, 2018.

[39] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[40] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *CVPR*, pages 850–859, 2019.

[41] Benjin Zhu, Jianfeng Wang, Zhengkai Jiang, Fuhang Zong, Songtao Liu, Zeming Li, and Jian Sun. Autoassign: Differentiable label assignment for dense object detection. *arXiv preprint arXiv:2007.03496*, 2020.

[42] Chenchen Zhu, Fangyi Chen, Zhiqiang Shen, and Marios Savvides. Soft anchor-point object detection. *arXiv preprint arXiv:1911.12448*, 2019.

[43] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *CVPR*, June 2019.