OTA: Optimal Transport Assignment for Object Detection

Zheng Ge[1,2], Songtao Liu[2]*, Zeming Li[2], Osamu Yoshie[1], Jian Sun[2]
[1]Waseda University, [2]Megvii Technology
jokerzz@fuji.waseda.jp;liusongtao@megvii.com;lizeming@megvii.com;
yoshie@waseda.jp;sunjian@megvii.com

Abstract

Recent advances in label assignment in object detection mainly seek to independently define positive/negative training samples for each ground-truth (gt) object. In this paper, we innovatively revisit the label assignment from a global perspective and propose to formulate the assigning procedure as an Optimal Transport (OT) problem – a well-studied topic in Optimization Theory. Concretely, we define the unit transportation cost between each demander (anchor) and supplier (gt) pair as the weighted summation of their classification and regression losses. After formulation, finding the best assignment solution is converted to solve the optimal transport plan at minimal transportation costs, which can be solved via Sinkhorn-Knopp Iteration. On COCO, a single FCOS-ResNet-50 detector equipped with Optimal Transport Assignment (OTA) can reach 40.7% mAP under 1× scheduler, outperforming all other existing assigning methods. Extensive experiments conducted on COCO and CrowdHuman further validate the effectiveness of our proposed OTA, especially its superiority in crowd scenarios. The code is available at https://github.com/Megvii-BaseDetection/OTA.

1. Introduction

Current CNN-based object detectors [27, 30, 21, 47, 33, 29, 36] perform a dense prediction manner by predicting the classification (*cls*) labels and regression (*reg*) offsets for a set of pre-defined anchors[1]. To train the detector, defining *cls* and *reg* targets for each anchor is a necessary procedure, which is called *label assignment* in object detection.

Classical label assigning strategies commonly adopt predefined rules to match the ground-truth (*gt*) object or background for each anchor. For example, RetinaNet [21] adopts Intersection-over-Union (IoU) as its thresholding criterion for *pos*/*neg* anchors division. Anchor-free detectors like FCOS [38] treat the anchors within the center/bbox region of any *gt* object as the corresponding positives. Such static strategies ignore a fact that for objects with different sizes, shapes or occlusion condition, the appropriate positive/negative (*pos*/*neg*) division boundaries may vary.

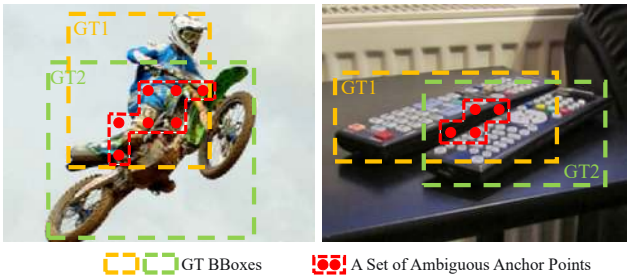

Figure 1. An illustration of ambiguous anchor points in object detection. Red dots show some of the ambiguous anchors in two sample images. Currently, the assignment of these ambiguous anchors is heavily based on hand-crafted rules.

Motivated by this, many dynamic assignment strategies have been proposed. ATSS [47] proposes to set the division boundary for each *gt* based on statistical characteristics. Other recent advances [48, 19, 51, 16] suggest that the predicted confidence scores of each anchor could be a proper indicator to design dynamic assigning strategies, *i.e.*, high confidence anchors can be easily learned by the networks and thus be assigned to the related *gt*, while anchors with uncertain predictions should be considered as negatives. Those strategies enable the detector to dynamically choose positive anchors for each individual *gt* object and achieve state-of-the-art performance.

However, independently assigning *pos*/*neg* samples for each *gt* without context could be sub-optimal, just like the lack of context may lead to improper prediction. When dealing with ambiguous anchors (*i.e.*, anchors that are qualified as positive samples for multiple *gt*s simultaneously as seen in Fig. 1.), existing assignment strategies are heavily based on hand-crafted rules (*e.g.*, Min Area [38], Max IoU [16, 21, 47].). We argue that assigning ambiguous an-

*Corresponding author

[1]For anchor-free detectors like FCOS [38], the feature points can be viewed as shrunk anchor boxes. Hence in this paper, we collectively refer to anchor box and anchor point as "anchor".

chors to any *gt* (or *background*) may introduce harmful gradients *w.r.t.* other *gt*s. Hence the assignment for ambiguous anchors is non-trivial and requires further information beyond the local view. Thus a better assigning strategy should get rid of the convention of pursuing optimal assignment for each *gt* independently and turn to the ideology of global optimum, in other words, finding the global high confidence assignment for all *gt*s in an image.

DeTR [3] is the first work that attempts to consider *label assignment* from global view. It replaces the detection head with transformer layers [39] and considers one-to-one assignment using the Hungarian algorithm that matches only one query for each *gt* with global minimum loss. However, for the CNN based detectors, as the networks often produce correlated scores to the neighboring regions around the object, each *gt* is assigned to many anchors (*i.e.*, one-to-many), which also benefits to training efficiency. In this one-to-many manner, it remains intact to assign labels with a global view.

To achieve the global optimal assigning result under the one-to-many situation, we propose to formulate *label assignment* as an Optimal Transport (OT) problem – a special form of Linear Programming (LP) in Optimization Theory. Specifically, we define each *gt* as a supplier who supplies a certain number of labels, and define each anchor as a demander who needs one unit label. If an anchor receives sufficient amount of positive label from a certain *gt*, this anchor becomes one positive anchor for that *gt* . In this context, the number of positive labels each *gt* supplies can be interpreted as "how many positive anchors that *gt* needs for better convergence during the training process". The unit transportation cost between each anchor-*gt* pair is defined as the weighted summation of their pair-wise *cls* and *reg* losses. Furthermore, as being negative should also be considered for each anchor, we introduce another supplier – *background* who supplies negative labels to make up the rest of labels in need. The cost between *background* and a certain anchor is defined as their pair-wise classification loss only. After formulation, finding the best assignment solution is converted to solve the optimal transport plan, which can be quickly and efficiently solved by the off-the-shelf Sinkhorn-Knopp Iteration [5]. We name such an assigning strategy as Optimal Transport Assignment (OTA).

Comprehensive experiments are carried out on MS COCO [22] benchmark, and significant improvements from OTA demonstrate its advantage. OTA also achieves the SOTA performance among one-stage detectors on a crowded pedestrian detection dataset named CrowdHuman [35], showing OTA's generalization ability on different detection benchmarks.

2. Related Work

2.1. Fixed Label Assignment

Determining which *gt* (or *background*) should each anchor been assigned to is a necessary procedure before training object detectors. Anchor-based detectors usually adopt IoU at a certain threshold as the assigning criterion. For example, RPN in Faster R-CNN [33] uses 0.7 and 0.3 as the positive and negative thresholds, respectively. When training the R-CNN module, the IoU threshold for *pos*/*neg* division is changed to 0.5. IoU based *label assignment* is proved effective and soon been adopted by many Faster R-CNN's variants like [2, 12, 20, 42, 26, 49, 37], as well as many one-stage detectors like [31, 32, 25, 27, 23, 21].

Recently, anchor-free detectors have drawn much attention because of their concision and high computational efficiency. Without anchor box, FCOS [38], Foveabox [17] and their precursors [30, 14, 46] directly assign anchor points around the center of objects as positive samples, showing promising detection performance. Another stream of anchor-free detectors [18, 8, 50, 45, 4] view each object as a single or a set of key-points. They share distinct characteristics from other detectors, hence will not be further discussed in our paper.

Although detectors mentioned above are different in many aspects, as for *label assignment*, they all adopt a single fixed assigning criterion (*e.g.*, a fixed region of the center area or IoU threshold) for objects of various sizes, shapes, and categories, etc, which may lead to sub-optimal assigning results.

2.2. Dynamic Label Assignment

Many recent works try to make the label assigning procedure more adaptive, aiming to further improve the detection performance. Instead of using pre-defined anchors, GuidedAnchoring [40] generates anchors based on an anchor-free mechanism to better fit the distribution of various objects. MetaAnchor [44] proposes an anchor generation function to learn dynamic anchors from the arbitrary customized prior boxes. NoisyAnchors [19] proposes soft-label and anchor re-weighting mechanisms based on classification and localization losses. FreeAnchor [48] constructs top-k anchor candidates for each *gt* based on IoU and then proposes a detection-customized likelihood to perform *pos*/*neg* division within each candidate set. ATSS [47] proposes an adaptive sample selection strategy that adopts *mean*+*std* of IoU values from a set of closest anchors for each *gt* as a *pos*/*neg* threshold. PAA [16] assumes that the distribution of joint loss for positive and negative samples follows the Gaussian distribution. Hence it uses GMM to fit the distribution of positive and negative samples, and then use the center of positive sample distribution as the *pos*/*neg* division boundary. AutoAssign [51] tackles *label*

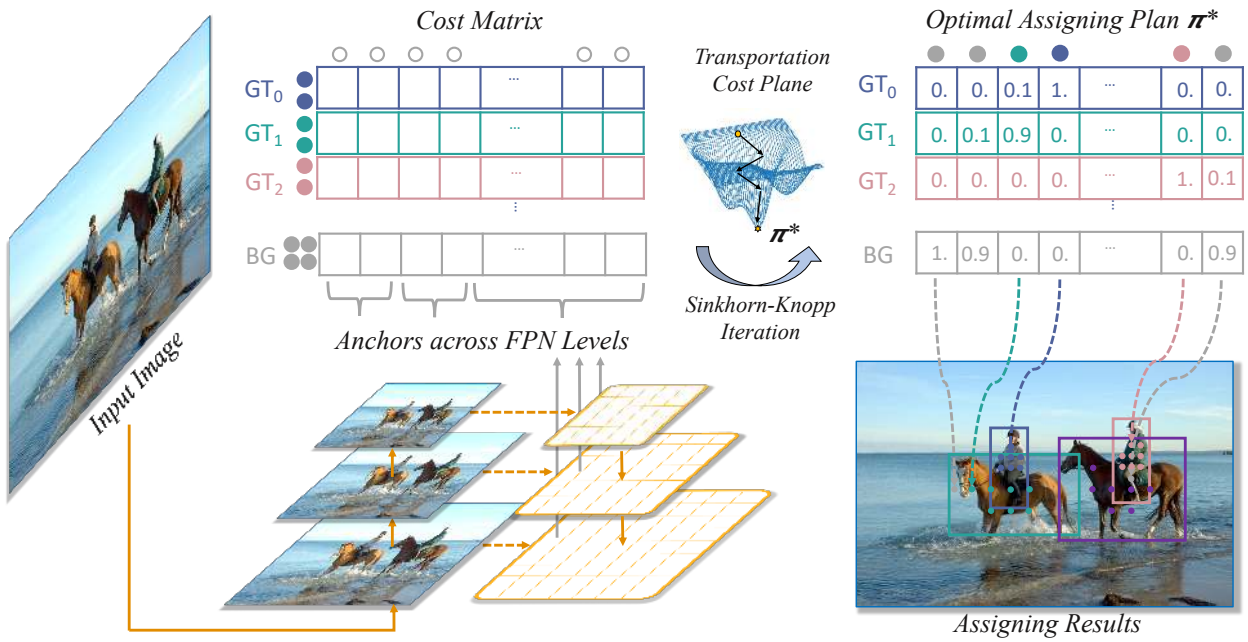

Figure 2. An illustration of Optimal Transport Assignment. *Cost Matrix* is composed of the pair-wise *cls* and *reg* losses between each anchor-*gt* pair. The goal of finding the best label assigning is converted to solve the best transporting plan which transports the labels from suppliers (*i.e.* GT and BG) to demanders (*i.e.* anchors) at a minimal transportation cost via Sinkhorn-Knopp Iteration.

assignment in a fully data-driven manner by automatically determine the positives/negatives in both spatial and scale dimensions.

These methods explore the optimal assigning strategy for individual objects, while failing to consider context information from a global perspective. DeTR [3] examines the idea of global optimal matching. But the Hungarian algorithm they adopted can only work in a one-to-one assignment manner. So far, for the CNN based detectors in one-to-many scenarios, a global optimal assigning strategy remains uncharted.

3. Method

In this section, we first revisit the definition of the Optimal Transport problem and then demonstrate how we formulate the *label assignment* in object detection into an OT problem. We also introduce two advanced designs which we suggest adopting to make the best use of OTA.

3.1. Optimal Transport

The Optimal Transport (OT) describes the following problem: supposing there are m suppliers and n demanders in a certain area. The i-th supplier holds s_i units of goods while the j-th demander needs d_j units of goods. Transporting cost for each unit of good from supplier i to demander j is denoted by c_{ij}. The goal of OT problem is to find a transportation plan $\pi^* = \{\pi_{i,j} | i = 1, 2, ...m, j = 1, 2, ...n\}$, according to which all goods from suppliers can be transported to demanders at a minimal transportation cost:

$$\begin{aligned} \min_{\pi} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\pi_{ij}. \\ \text{s.t.} \quad & \sum_{i=1}^{m} \pi_{ij} = d_j, \quad \sum_{j=1}^{n} \pi_{ij} = s_i, \\ & \sum_{i=1}^{m} s_i = \sum_{j=1}^{n} d_j, \\ & \pi_{ij} \geq 0, \quad i = 1, 2, ...m, j = 1, 2, ...n. \end{aligned} \tag{1}$$

This is a linear program which can be solved in polynomial time. In our case, however, the resulting linear program is large, involving the square of feature dimensions with anchors in all scales. We thus address this issue by a fast iterative solution, named Sinkhorn-Knopp [5] (described in Appendix.)

3.2. OT for Label Assignment

In the context of object detection, supposing there are m *gt* targets and n anchors (across all FPN [20] levels) for an input image I, we view each *gt* as a supplier who holds k units of positive labels (*i.e.*, $s_i = k, i = 1, 2, ..., m$), and each anchor as a demander who needs one unit of label (*i.e.*, $d_j = 1, j = 1, 2, ..., n$). The cost c^{fg} for transporting one

unit of positive label from gt_i to anchor a_j is defined as the weighted summation of their cls and reg losses:

$$c_{ij}^{fg} = L_{cls}(P_j^{cls}(\theta), G_i^{cls}) + \alpha L_{reg}(P_j^{box}(\theta), G_i^{box}), \quad (2)$$

where θ stands for model's parameters. P_j^{cls} and P_j^{box} denote predicted cls score and bounding box for a_j. G_i^{cls} and G_i^{box} denote ground truth class and bounding box for $gt\ i$. L_{cls} and L_{reg} stand for cross entropy loss and IoU Loss [46]. One can also replace these two losses with Focal Loss [21] and GIoU [34]/SmoothL1 Loss [11]. α is the balanced coefficient.

Besides positive assigning, a large set of anchors are treated as negative samples during training. As the optimal transportation involves all anchors, we introduce another supplier – *background*, who only provides negative labels. In a standard OT problem, the total supply must be equal to the total demand. We thus set the number of negative labels that *background* can supply as $n - m \times k$. The cost for transporting one unit of negative label from *background* to a_j is defined as:

$$c_j^{bg} = L_{cls}(P_j^{cls}(\theta), \varnothing), \quad (3)$$

where $\varnothing$ means the $background$ class. Concatenating this $c^{bg} \in \mathbb{R}^{1\times n}$ to the last row of $c^{fg} \in \mathbb{R}^{m\times n}$, we can get the complete form of the cost matrix $c \in \mathbb{R}^{(m+1)\times n}$. The supplying vector s should be correspondingly updated as:

$$s_i = \begin{cases} k, & if \quad i \le m \\ n - m \times k, & if \quad i = m+1. \end{cases} \quad (4)$$

As we already have the cost matrix c, supplying vector $s \in \mathbb{R}^{m+1}$ and demanding vector $d \in \mathbb{R}^n$, the optimal transportation plan $\pi^* \in \mathbb{R}^{(m+1)\times n}$ can be obtained by solving this OT problem via the off-the-shelf Sinkhorn-Knopp Iteration [5]. After getting π^*, one can decode the corresponding label assigning solution by assigning each anchor to the supplier who transports the largest amount of labels to them. The subsequent processes (*e.g.,* calculating losses based on assigning result, back-propagation) are exactly the same as in FCOS [38] and ATSS [47]. Noted that the optimization process of OT problem only contains some matrix multiplications which can be accelerated by GPU devices, hence OTA **only increases the total training time by less than 20%** and is totally cost-free in testing phase.

3.3. Advanced Designs

Center Prior. Previous works [47, 16, 48] only select positive anchors from the center region of objects with limited areas, called *Center Prior*. This is because they suffer from either a large number of ambiguous anchors or poor statistics in the subsequent process. Instead of relying on statistical characteristics, our OTA is based on global optimization methodology and thus is naturally resistant to these two issues. Theoretically, OTA can assign any anchor within the region of gts' boxes as a positive sample. However, for general detection datasets like COCO, we find the *Center Prior* still benefit the training of OTA. Forcing detectors focus on potential positive areas (*i.e.,* center areas) can help stabilize the training process, especially in the early stage of training, which will lead to a better final performance. Hence, we impose a *Center Prior* to the cost matrix. For each gt, we select r^2 closest anchors from each FPN level according to the center distance between anchors and gts [2]. As for anchors not in the r^2 closest list, their corresponding entries in the cost matrix c will be subject to an additional constant cost to reduce the possibility they are assigned as positive samples during the training stage. In Sec. 4, we will demonstrate that although OTA adopts a certain degree of *Center Prior* like other works [38, 47, 48] do, OTA consistently outperforms counterparts by a large margin when r is set to a large value (*i.e.*, large number of

Algorithm 1 Optimal Transport Assignment (OTA)

Input:

I is an input image
A is a set of anchors
G is the gt annotations for objects in image I
γ is the regularization intensity in Sinkhorn-Knopp Iter.
T is the number of iterations in Sinkhorn-Knopp Iter.
α is the balanced coefficient in Eq. 2

Output:

π^* is the optimal assigning plan

$m \leftarrow |G|, n \leftarrow |A|$
$P^{\text{cls}}, P^{\text{box}} \leftarrow$ Forward(I,A)
$s_i (i = 1, 2, ..., m) \leftarrow$ Dynamic k Estimation
$s_{m+1} \leftarrow n - \sum_{i=1}^{m} s_i$
$d_j (j = 1, 2, ..., n) \leftarrow$ OnesInit
pairwise cls cost: c_{cls}^{ij} = FocalLoss($P_j^{\text{cls}}, G_i^{\text{cls}}$)
pairwise reg cost: c_{reg}^{ij} = IoULoss(P_j^{box},G_i^{box})
pairwise Center Prior cost: $c_{ij}^{\text{cp}} \leftarrow (A_j, G_i^{\text{box}})$
$bg\ cls$ cost: $c_{\text{cls}}^{\text{bg}}$ = FocalLoss($P_j^{\text{cls}}, \varnothing$)
fg cost: $c^{\text{fg}} = c_{\text{cls}} + \alpha c_{\text{reg}} + c_{\text{cp}}$
compute final cost matrix c via concatenating $c_{\text{cls}}^{\text{bg}}$ to the last row of c^{fg}
$v^0, u^0 \leftarrow$ OnesInit
for t=0 **to** T **do**:
    $u^{t+1}, v^{t+1} \leftarrow$ SinkhornIter(c, u^t, v^t, s, d)
compute optimal assigning plan π^* according to Eq. **??**
return π^*

[2]For anchor-based methods, the distances are measured between the geometric center of anchors and gts

potential positive anchors as well as more ambiguous anchors).

Dynamic k Estimation. Intuitively, the appropriate number of positive anchors for each gt (*i.e.*, s_i in Sec. 3.1) should be different and based on many factors like objects' sizes, scales, and occlusion conditions, etc. As it is hard to directly model a mapping function from these factors to the positive anchor's number, we propose a simple but effective method to roughly estimate the appropriate number of positive anchors for each gt based on the IoU values between predicted bounding boxes and gts. Specifically, for each gt, we select the top q predictions according to IoU values. These IoU values are summed up to represent this gt's estimated number of positive anchors. We name this method as Dynamic k Estimation. Such an estimation method is based on the following intuition: The appropriate number of positive anchors for a certain gt should be positively correlated with the number of anchors that well-regress this gt. In Sec. 4, we present a detailed comparison between the fixed k and Dynamic k Estimation strategies.

A toy visualization of OTA is shown in Fig. 2. We also describe the OTA's completed procedure including *Center Prior* and Dynamic k Estimation in Algorithm 1.

4. Experiments

In this section, we conduct extensive experiments on MS COCO 2017 [22] which contains about $118k$, $5k$ and $20k$ images for *train*, *val*, and *test-dev* sets, respectively. For ablation studies, we train detectors on *train* set and report the performance on *val* set. Comparisons with other methods are conducted on *test-dev* set. We also compare OTA with other methods on CrowdHuman [35] validation set to demonstrate the superiority of OTA in crowd scenarios.

4.1. Implementation Details

If not specified, we use ResNet-50 [13] pre-trained on ImageNet [6] with FPN [20] as our default backbone. Most of experiments are trained with $90k$ iterations which is denoted as "$1\times$". The initial learning rate is 0.01 and is decayed by a factor of 10 after $60k$ and $80k$ iterations. Mini-batch size is set to 16. Following the common practice, the model is trained with SGD [1] on 8 GPUs.

OTA can be adopted in both anchor-based and anchor-free detectors, the following experiments are mainly conducted on FCOS [38] because of its simplicity. We adopt Focal Loss and IoU Loss as L_{cls} and L_{reg} that make up the cost matrix. α in Eq. 2 is set to 1.5. For back-propagation, the regression loss is replaced by GIoU Loss and is re-weighted by a factor of 2. IoU Branch is first introduced in YOLOv1 [30] and proved effective in modern one-stage object detectors by PAA [16]. We also adopt IoU Branch as a default component in our experiments. The top q in Sec. 3.3 is directly set to 20, as we find this set of parameter values can consistently yield stable results in various situations.

Method	Aux. Branch	Center	Dyn. k	AP	AP_{50}	AP_{75}
FCOS	-	✓		38.3	57.1	41.3
	CenterNess	✓		38.9	57.5	42.0
	IoU			38.8	57.7	41.8
	IoU	✓		39.5	57.6	42.9
OTA (FCOS)	-	✓		39.2	58.3	42.2
	IoU			39.6	58.1	42.5
	IoU	✓		40.3	**58.6**	43.7
	IoU	✓	✓	**40.7**	58.4	**44.3**
OTA (RetinaNet)	IoU	✓	✓	**40.7**	**58.6**	44.1

Table 1. Ablation studies on each components in OTA. "Center" stands for *Center Prior* and *Center Sampling* for OTA and FCOS, respectively. Dyn.k is the abbreviation of our proposed Dynamic k Estimation strategy.

4.2. Ablation Studies and Analysis

Effects of Individual Components. We verify the effectiveness of each component in our proposed methods. For fair comparisons, all detectors' regression losses are multiplied by 2, which is known as a useful trick to boost the AP at high IoU thresholds [28]. As seen in Table 1, when no auxiliary branch is adopted, OTA outperforms FCOS by 0.9% AP (39.2% *v.s.*38.3%). This gap almost remains the same after adding IoU branch to both of them (39.5% *v.s.* 40.3% and 38.8% *v.s.* 39.6% with or without center prior, respectively). Finally, dynamic k pushes AP to a new state-of-the-art 40.7%. In the whole paper, we emphasize that **OTA can be applied to both anchor-based and anchor-free detectors**. Hence we also adopt OTA on RetinaNet [21] with only one square anchor per-location across feature maps. As shown in Table 1, the AP values of OTA-FCOS and OTA-RetinaNet are exactly the same, demonstrating OTA's applicability on both anchor-based and anchor-free detectors.

Effects of r. The values of radius r for *Center Prior* serve to control the number of candidate anchors for each gt. If adopting a small r, only anchors near objects' centers could be assigned as positives, helping the optimization process focus on regions that are more likely to be informative. As r increases, the number of candidates also quadratically increases, leading to potential instability in the optimization process. For example, when r is set to 3, 5 or 7, their corresponding numbers of candidate anchors are 45, 125 and 245[3], respectively. We study behaviors of ATSS [47],

[3]Total number of potential positive anchors equals to (r^2∗FPN Levels).

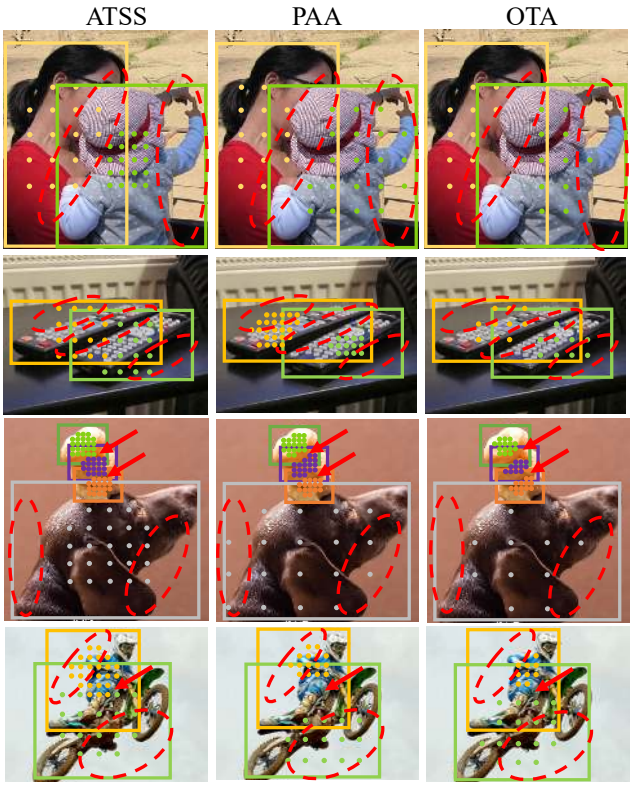

Figure 3. Visualizations of assigning results. For PAA, the dots stand for geometric centers of positive anchor boxes. For ATSS and OTA, the dots stand for positive anchor points. Rectangles represent the gt bounding boxes. To clearly illustrate the differences between different assigning strategies, we set r to 5 for all methods. Only FPN layers with the largest number of positive anchors are shown for better visualization.

PAA [16], and OTA under different values of r in Table 2. OTA achieves the best performance (40.7% AP) when r is set to 5. When r is set to 3 as ATSS and PAA do, OTA also achieves 40.6% AP, indicating that most potential positive anchors are near the center of objects on COCO. While r is set to 7, the performance only slightly drops 0.3%, showing that OTA is insensitive to the hyper-parameter r.

Ambiguous Anchors Handling. Most existing dynamic label assigning methods [47, 16, 48] only conduct a small candidate set for each gt, because a large number of candidates brings trouble – when occlusion happens or several objects are close enough, an anchor may simultaneously be a qualified candidate for multiple gts. We define such anchors as *ambiguous anchors*. Previous methods mainly handle this ambiguity by introducing hand-crafted rules *e.g., Min Area [38], Max IoU [47, 16, 21] and Min Loss*[4]. To illustrate OTA's superiority on ambiguous handling, We count the number of ambiguous anchors in ATSS, PAA and OTA, and evaluate their corresponding performance under different r in Table 2. Noted that the optimal assigning plan in OTA is continuous, hence we define anchor a_j as an ambiguous anchor if $\max \pi_j^* < 0.9$. Table 2 shows that for ATSS, the number of ambiguous anchors greatly increases as r varies from 3 to 7. Its performance correspondingly drops from 39.4% to 37.2%. For PAA, the number of ambiguous anchors is less sensitive to r, but its performance still drops 0.8%, indicating that *Max IoU* adopted by PAA is not an ideal prior to ambiguous anchors. In OTA, when multiple gts tend to transport positive labels to the same anchor, the OT algorithm will automatically resolve their conflicts based on the principle of minimum global costs. Hence the number of ambiguous anchor for OTA remains low and barely increases as r increases from 3 to 7. The corresponding performance is also stable.

[4]Assigning ambiguous anchor to the gt with the minimal loss.

Method	ATSS [47]			PAA [16]			OTA		
r	3	5	7	3	5	7	3	5	7
$N_{amb.}$	2.1	15.9	36.3	0.5	0.8	1.2	0.2	0.2	0.3
AP	39.4	38.0	37.2	40.3	40.1	39.5	40.6	**40.7**	40.4
AP_{50}	57.5	56.7	55.8	**58.9**	58.4	57.5	58.7	58.4	58.3
AP_{75}	42.7	40.4	39.8	43.4	43.4	42.4	44.1	**44.3**	43.6

Table 2. Performances of different label assigning strategies under different number of anchor candidates. $N_{amb.}$ denotes the average number of ambiguous anchors per-image calculated on COCO *train* set.

Further, we *manually* assign the ambiguous anchors based on hand-crafted rules before performing OTA. In this case, OTA is only in charge of *pos/neg* samples division. Table 3 shows that such a combination of hand-crafted rules and OTA decreases the AP by 0.7% and 0.4%, respectively. Finally, we visualize some assigning results in Fig. 3. Red arrows and dashed ovals highlight the ambiguous regions (*i.e.,* overlaps between different fgs or junctions between fgs and bg). Suffering from the lack of context and global information, ATSS and PAA perform poorly in such regions, leading to sub-optimal detection performances. Conversely, OTA assigns much less positive anchors in such regions, which we believe is a desired behavior.

Method	AP	AP_{50}	AP_{75}
Min Area [38] *f.b.* OTA	40.0	57.8	43.6
Max IoU [47] *f.b.* OTA	40.3	58.1	43.7
Min Loss *f.b.* OTA	40.3	57.9	43.6
OTA	**40.7**	**58.4**	**44.3**

Table 3. Performance comparisons on ambiguity handling between OTA and other human-designed strategies on the COCO *val* set.. *f.b.* denotes "followed by".

Effects of k. Before performing Sinkhorn-Knopp Iteration, we need to define how many positive labels can each

gt supply. This value also represents how many anchors every gt needs for better convergence. A naive way is setting k to a constant value for all gts. We try different values of k from 1 to 20. As seen in Table 4, among all different values, k=10 and k=12 achieve the best performances. As k increases from 10 to 20, the possibility that an anchor is suitable as a positive sample for two close targets at the same time also increases, but there is no obvious performance drop (0.2%) according to Table 4 which proves OTA's superiority in handling potential ambiguity. When k=1, OTA becomes a one-to-one assigning strategy, the same as in DeTR. The poor performance tells us that achieving competitive performance via one-to-one assignment under the $1\times$ scheduler remains challenging, unless an auxiliary one-to-many supervision is added [41].

k	AP	AP_{50}	AP_{75}	APs	APm	APl
1	36.5	55.4	38.8	21.4	39.7	46.2
5	39.5	58.1	42.7	23.1	43.0	50.6
8	39.8	58.4	42.9	22.7	43.6	51.5
10	40.3	**58.6**	43.7	23.4	44.2	52.1
12	40.3	**58.6**	43.6	23.2	44.2	51.9
15	40.2	58.4	43.6	23.2	44.1	51.9
20	40.1	58.2	43.6	**23.5**	44.0	52.8
Dyn. k	**40.7**	58.4	**44.3**	23.2	**45.0**	**53.6**

Table 4. Analysis of different values of k and Dynamic k Estimation strategy on the COCO *val* set.

Fixing k strategy assumes every gt has the same number of appropriate positive anchors. However, we believe that this number for each gt should vary and may be affected by many factors like objects' sizes, spatial attitudes, and occlusion conditions, etc. Hence we adopt the Dynamic k Estimation proposed in Sec 3.3 and compare its performance to the fixed k strategy. Results in Table 4 shows that dynamic k surpasses the best performance of fixed k by 0.4% AP, validating our point and the effectiveness of Dynamic k Estimation strategy.

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

We compare our final models with other state-of-the-art one-stage detectors on MS COCO *test-dev*. Following previous works [21, 38], we randomly scale the shorter side of images in the range from 640 to 800. Besides, we double the total number of iterations to $180K$ with the learning rate change points scaled proportionally. Other settings are consistent with [21, 38].

As shown in Table 5, our method with ResNet-101-FPN achieves 45.3% AP, outperforms all other methods with the same backbone including ATSS (43.6% AP), AutoAssign (44.5% AP) and PAA (44.6% AP). Noted that for PAA, we remove the *score voting* procedure for fair comparisons between different label assigning strategies. With ResNeXt-64x4d-101-FPN [43], the performance of OTA can be further improved to 47.0% AP. To demonstrate the compatibility of our method with other advanced technologies in object detection, we adopt Deformable Convolutional Networks (DCN) [54] to ResNeXt backbones as well as the last convolution layer in the detection head. This improves our model's performance from 47.0% AP to 49.2% AP. Finally, with the multi-scale testing technique, our best model achieves 51.5% AP.

4.4. Experiments on CrowdHuman

Object detection in crowded scenarios has raised more and more attention [24, 15, 9, 10]. Compared to dataset designed for general object detection like COCO, ambiguity happens more frequently in crowded dataset. Hence to demonstrate OTA's advantage on handling ambiguous anchors, it is necessary to conduct experiments on a crowded dataset – Crowdhuman [35]. CrowdHuman contains 15000, 4370, and 5000 images in training, validation, and test set, respectively, with the average number of persons in an image 22.6. For all experiments, we train the detectors for 30 epochs (*i.e.*, 2.5x) for better convergence. NMS threshold is set to 0.6. We adopt ResNet-50 [13] as the default backbone in our experiments. Other settings are the same as our experiments on COCO. For evaluation, we follow the standard Caltech [7] evaluation metric – MR, which stands for the Log-Average Missing Rate over false positives per image (FPPI) ranging in $[10^{-2}, 10^{0}]$. AP and Recall are also reported for reference. All evaluation results are reported on the CrowdHuman *val* subset.

As shown in Table 6, RetinaNet and FCOS only achieve 58.8% and 55.0% MR respectively, which are far worse than two stage detectors like Faster R-CNN (with FPN), revealing the dilemma of one-stage detectors in crowd scenarios. Starting from FreeAnchor, the performances of one-stage detectors gradually get improved by the dynamic label assigning strategies. ATSS achieves 49.5% MR, which is very close to the performance of Faster R-CNN (48.7% AP). Recent proposed LLA [10] leverages loss-aware label assignment, which is similar to OTA and achieves 47.9% MR. However, our OTA takes a step forward by introducing global information into the label assignment, boosting MR to 46.6%. The AP and Recall of OTA also surpass other existing one-stage detectors by a clear margin.

Although PAA achieves competitive performance with OTA on COCO, it performs struggling on CrowdHuman. We conjecture that PAA needs clear *pos*/*neg* decision boundaries to help GMM learn better clusters. But in crowded scenarios, such clear boundaries may not exist because potential negative samples usually cover a sufficient amount of foreground areas, resulting in PAA's poor performance. Also, PAA performs per-*gt*'s clustering, which heavily increases the training time on crowded datasets like CrowdHuman. Compared to PAA, OTA still shows promis-

Method	Iteration	Backbone	AP	AP_{50}	AP_{75}	APs	APm	APl
RetinaNet [21]	135k	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
FCOS [38]	180k	ResNet-101	41.5	60.7	45.0	24.4	44.8	51.6
NoisyAnchor [19]	180k	ResNet-101	41.8	61.1	44.9	23.4	44.9	52.9
FreeAnchor [48]	180k	ResNet-101	43.1	62.2	46.4	24.5	46.1	54.8
SAPD [52]	180k	ResNet-101	43.5	63.6	46.5	24.9	46.8	54.6
MAL [44]	180k	ResNet-101	43.6	61.8	47.1	25.0	46.9	55.8
ATSS [47]	180k	ResNet-101	43.6	62.1	47.4	26.1	47.0	53.6
AutoAssign [51]	180k	ResNet-101	44.5	**64.3**	48.4	25.9	47.4	55.0
PAA [16]	180k	ResNet-101	44.6	63.3	48.4	26.4	48.5	56.0
OTA (Ours)	180k	ResNet-101	**45.3**	63.5	**49.3**	**26.9**	**48.8**	**56.1**
FoveaBox [17]	180k	ResNeXt-101	42.1	61.9	45.2	24.9	46.8	55.6
FSAF [53]	180k	ResNeXt-64x4d-101	42.9	63.8	46.3	26.6	46.2	52.7
FCOS [38]	180k	ResNeXt-64x4d-101	43.2	62.8	46.6	26.5	46.2	53.3
NoisyAnchor [19]	180k	ResNeXt-101	44.1	63.8	47.5	26.0	47.4	55.0
FreeAnchor [48]	180k	ResNeXt-64x4d-101	44.9	64.3	48.5	26.8	48.3	55.9
SAPD [52]	180k	ResNeXt-64x4d-101	45.4	65.6	48.9	27.3	48.7	56.8
ATSS [47]	180k	ResNeXt-64x4d-101	45.6	64.6	49.7	28.5	48.9	55.6
MAL [44]	180k	ResNeXt101	45.9	65.4	49.7	27.8	49.1	57.8
AutoAssign [51]	180k	ResNeXt-64x4d-101	46.5	**66.5**	50.7	28.3	49.7	56.6
PAA [16]	180k	ResNeXt-64x4d-101	46.6	65.6	50.7	28.7	**50.5**	**58.1**
OTA (Ours)	180k	ResNeXt-64x4d-101	**47.0**	65.8	**51.1**	**29.2**	50.4	57.9
SAPD [52]	180k	ResNeXt-64x4d-101-DCN	47.4	67.4	51.1	28.1	50.3	61.5
ATSS [47]	180k	ResNeXt-64x4d-101-DCN	47.7	66.5	51.9	29.7	50.8	59.4
AutoAssign [51]	180k	ResNeXt-64x4d-101-DCN	48.3	67.4	52.7	29.2	51.0	60.3
PAA [16]	180k	ResNeXt-64x4d-101-DCN	48.6	67.5	52.7	29.9	52.2	61.5
OTA (Ours)	180k	ResNeXt-64x4d-101-DCN	**49.2**	**67.6**	**53.5**	**30.0**	**52.5**	**62.3**
ATSS [47]*	180k	ResNeXt-64x4d-101-DCN	50.7	**68.9**	56.3	33.2	52.9	62.2
PAA [16]*	180k	ResNeXt-64x4d-101-DCN	51.3	68.8	56.6	**34.3**	53.5	63.6
OTA (Ours)*	180k	ResNeXt-64x4d-101-DCN	**51.5**	68.6	**57.1**	34.1	**53.7**	**64.1**

Table 5. Performance comparison with state-of-the-art one-stage detectors on MS COCO 2017 *test-dev* set. * indicates the specific form of multi-scale testing that adopted in ATSS [47].

Method	MR	AP	Recall
Faster R-CNN *with* FPN [20]	48.7	86.1	90.4
RetinaNet [21]	58.8	81.0	88.2
FCOS [38]	55.0	86.4	94.1
FreeAnchor [48]	51.3	83.9	89.8
ATSS [47]	49.5	87.4	94.2
PAA [16]	52.2	86.0	92.0
LLA [10]	47.9	88.0	94.0
OTA (Ours)	**46.6**	**88.4**	**95.1**

Table 6. Performance comparison on the CrowdHuman validation set. All experiments are conducted under 2.5x scheduler.

ing results, which demonstrates OTA's superiority on various detection benchmarks.

5. Conclusion

In this paper, we propose Optimal Transport Assignment (OTA) – an optimization theory based label assigning strategy. OTA formulates the label assigning procedure in object detection into an Optimal Transport problem, which aims to transport labels from ground-truth objects and backgrounds to anchors at minimal transporting costs. To determine the number of positive labels needed by each gt, we further propose a simple estimation strategy based on the IoU values between predicted bounding boxes and each gt. As shown in experiments, OTA achieves the new SOTA performance on MS COCO. Because OTA can well-handle the assignment of ambiguous anchors, it also outperforms all other one-stage detectors on CrowdHuman dataset by a large margin, demonstrating its strong generalization ability.

Acknowledgements

This research was partially supported by National Key R&D Program of China (No. 2017YFA0700800), and Beijing Academy of Artificial Intelligence (BAAI).

References

[1] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. 5

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

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020. 2, 3

[4] Yihong Chen, Zheng Zhang, Yue Cao, Liwei Wang, Stephen Lin, and Han Hu. Reppoints v2: Verification meets regression for object detection. *arXiv preprint arXiv:2007.08508*, 2020. 2

[5] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013. 2, 3, 4

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[7] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–311. IEEE, 2009. 7

[8] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578, 2019. 2

[9] Zheng Ge, Zequn Jie, Xin Huang, Rong Xu, and Osamu Yoshie. Ps-rcnn: Detecting secondary human instances in a crowd via primary object suppression. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 7

[10] Zheng Ge, Jianfeng Wang, Xin Huang, Songtao Liu, and Osamu Yoshie. Lla: Loss-aware label assignment for dense pedestrian detection. *arXiv preprint arXiv:2101.04307*, 2021. 7, 8

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 4

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 7

[14] 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. 2

[15] Xin Huang, Zheng Ge, Zequn Jie, and Osamu Yoshie. Nms by representative region: Towards crowded pedestrian detection by proposal pairing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10750–10759, 2020. 7

[16] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. *arXiv preprint arXiv:2007.08103*, 2020. 1, 2, 4, 5, 6, 8

[17] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020. 2, 8

[18] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018. 2

[19] Hengduo Li, Zuxuan Wu, Chen Zhu, Caiming Xiong, Richard Socher, and Larry S Davis. Learning from noisy anchors for one-stage object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10588–10597, 2020. 1, 2, 8

[20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2, 3, 5, 8

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1, 2, 4, 5, 6, 7, 8

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 5

[23] Songtao Liu, Di Huang, et al. Receptive field block net for accurate and fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 385–400, 2018. 2

[24] Songtao Liu, Di Huang, and Yunhong Wang. Adaptive nms: Refining pedestrian detection in a crowd. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6459–6468, 2019. 7

[25] Songtao Liu, Di Huang, and Yunhong Wang. Learning spatial fusion for single-shot object detection. *arXiv preprint arXiv:1911.09516*, 2019. 2

[26] Songtao Liu, Di Huang, and Yunhong Wang. Pay attention to them: deep reinforcement learning-based cascade object detection. *IEEE transactions on neural networks and learning systems*, 31(7):2544–2556, 2019. 2

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 1, 2

[28] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 821–830, 2019. 5

[29] Han Qiu, Yuchen Ma, Zeming Li, Songtao Liu, and Jian Sun. Borderdet: Border feature for dense object detection. In

European Conference on Computer Vision, pages 549–564. Springer, 2020. 1

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1, 2, 5

[31] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2

[32] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 2

[34] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019. 4

[35] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 2, 5, 7

[36] Lin Song, Yanwei Li, Zhengkai Jiang, Zeming Li, Hongbin Sun, Jian Sun, and Nanning Zheng. Fine-grained dynamic head for object detection. *arXiv preprint arXiv:2012.03519*, 2020. 1

[37] Lin Song, Yanwei Li, Zhengkai Jiang, Zeming Li, Xiangyu Zhang, Hongbin Sun, Jian Sun, and Nanning Zheng. Rethinking learnable tree filter for generic feature transform. *arXiv preprint arXiv:2012.03482*, 2020. 2

[38] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 9627–9636, 2019. 1, 2, 4, 5, 6, 7, 8

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[40] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2965–2974, 2019. 2

[41] Jianfeng Wang, Lin Song, Zeming Li, Hongbin Sun, Jian Sun, and Nanning Zheng. End-to-end object detection with fully convolutional network. *arXiv preprint arXiv:2012.03544*, 2020. 7

[42] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *European Conference on Computer Vision*, pages 456–472. Springer, 2020. 2

[43] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 7

[44] Tong Yang, Xiangyu Zhang, Zeming Li, Wenqiang Zhang, and Jian Sun. Metaanchor: Learning to detect objects with customized anchors. In *Advances in Neural Information Processing Systems*, pages 320–330, 2018. 2, 8

[45] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9657–9666, 2019. 2

[46] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 516–520, 2016. 2, 4

[47] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9759–9768, 2020. 1, 2, 4, 5, 6, 8

[48] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. In *Advances in Neural Information Processing Systems*, pages 147–155, 2019. 1, 2, 4, 6, 8

[49] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13766–13775, 2020. 2

[50] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2

[51] 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. 1, 2, 8

[52] Chenchen Zhu, Fangyi Chen, Zhiqiang Shen, and Marios Savvides. Soft anchor-point object detection. *arXiv preprint arXiv:1911.12448*, 2019. 8

[53] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 840–849, 2019. 8

[54] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 7