DLA: Dynamic Label Assignment for Accurate One-stage Object Detection

He Jiang
Institute of Automation, Chinese Academy of Sciences
Haidian Qu, Beijing Shi, China
School of Artificial Intelligence, University of Chinese Academy of Sciences
Haidian Qu, Beijing Shi, China
jianghe2019@ia.ac.cn

Junrui Xiao
Institute of Automation, Chinese Academy of Sciences
Haidian Qu, Beijing Shi, China
School of Artificial Intelligence, University of Chinese Academy of Sciences
Haidian Qu, Beijing Shi, China
xiaojunrui2020@ia.ac.cn

Qingyi Gu*
Institute of Automation, Chinese Academy of Sciences
Haidian Qu, Beijing Shi, China
School of Artificial Intelligence, University of Chinese Academy of Sciences
Haidian Qu, Beijing Shi, China
qingyi.gu@ia.ac.cn

ABSTRACT

One-stage object detector has been the most widely used framework in modern object detection due to its excellent performance and high efficiency. Label assignment, which is designed to discriminate positive and negative samples in training process, is closely correlated to the detection performance of one-stage detectors. Previous works commonly utilize geometric prior such as anchor box or key point to determine positive samples. Despite its simplicity, the heuristic strategy is rigid and it might limit the upper bound of detection performance. By introducing extra semantic information, prediction-aware geometric score and sample re-weighting mechanism, we propose a novel strategy called Dynamic Label Assignment in this paper. To validate the effectiveness and generalization of our method, we conduct extensive experiments on the MS COCO dataset. Without bells and whistles, our best model with ResNeXt-101 as backbone achieves state-of-the-art 46.5 AP, surpassing other strong methods such as SAPD [30] (45.4 AP), ATSS [25] (45.6 AP), and GFL [11] (46.0 AP) by a large marigin.

CCS CONCEPTS

• **Computing methodologies** → **Object detection**.

KEYWORDS

deep learning, object detection, one-stage detector, label assignment

ACM Reference Format:
He Jiang, Junrui Xiao, and Qingyi Gu. 2022. DLA: Dynamic Label Assignment for Accurate One-stage Object Detection. In *2022 11th International Conference on Software and Computer Applications (ICSCA 2022), February 24–26, 2022, Melaka, Malaysia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3524304.3524317

*Corresponding author.

ICSCA 2022, February 24–26, 2022, Melaka, Malaysia

ACM ISBN 978-1-4503-8577-0/22/02.
https://doi.org/10.1145/3524304.3524317

1 INTRODUCTION

Object detection, which aims to recognize the category and detect the location of objects, has been a fundamental but challenging task in computer vision for a long time. In recent years, with the rapid development of convolutional neural networks (CNN), numerous works have been proposed to improve the performance from several aspects [1, 6, 12, 13, 15, 18]. Generally, current detectors can be divided into two-stage [4, 18] and one-stage methods [13, 15]. Due to the simplicity and high efficiency of one-stage methods, they attract wider attention and have been the most popular frameworks in modern object detection.

In the training process, the label assignment strategy plays an important role and it is responsible for separating positive and negative samples. Currently, most one-stage detectors merely utilize geometric information such as IoU to determine the positive samples. To be specific, they tile anchor boxes at each spatial location with various scales and aspect ratios. These anchor boxes are introduced as geometric priors and they are assigned as positive or negative samples based on their IoUs with ground truth (GT). An anchor is selected as the positive sample if its IoU with any GT box exceeds a certain threshold. This simple strategy is heuristic and it might suffer from two major limits as shown in Figure 1. First, **the assignment strategy is misaligned with the NMS procedure**. Current one-stage detectors commonly utilize classification score to rank predicted boxes in NMS procedure. However, the label assignment strategy is only based on geometric information, causing a certain degree of misalignment. As a result, a box with higher classification score (anchor B) could suppress the box (anchor A) with higher IoU score. Second, **the assignment strategy is fixed during training process**. Conventional label assignment strategy could not adaptively determine the positive and negative samples as training processes. Therefore, the training samples are actually fixed for each GT, which might hurt the final performance. Although recent works [7, 10, 25, 26] try to improve the label assignment strategy, they do not solve these two issues directly.

In this paper, we propose Dynamic Label Assignment (DLA), a hyper-parameter insensitive label assignment strategy to explicitly address the problems above. To be more specific, we design a new assignment metric and introduce an auxiliary task for NMS procedure. By incorporating semantic score and proposing prediction-aware geometric score, our DLA could dynamically assign the samples

Figure 1: Illustration of the defects of the conventional label assignment in current one-stage detectors. They suffer from two major limits: (1) Misalignment between assignment strategy and NMS procedure. (2) Rigid sample assignment during training process.

for each GT during training process. We also utilize a sample re-weighting mechanism to encourage the learning from high-quality samples. Without bells and whistles, DLA is able to improve the performance of one-stage object detectors significantly and our best model achieves state-of-the-art 46.5 AP on MS COCO dataset. The contributions of this paper are summarized as follows:

- We systematically analyze the defects of the conventional label assignment in modern one-stage detectors.
- A novel assignment strategy called DLA is proposed to significantly boost the performance of existing one-stage models.
- Extensive experiments are conducted on MS COCO dataset to verify the effectiveness of our proposed method.

2 RELATED WORKS

2.1 One-stage Detector

Anchor-based method. SSD [15] is the pioneer anchor-based one-stage detector, which spreads anchor priors in multi-scale layers to directly perform classification and regression. With the advantage of SSD, substantial progress has been made to improve the performance of one-stage detectors in various ways. [2, 11, 13] propose new loss functions to address the imbalance problem of positive and negative samples. Other works [16, 22, 27] try to enhance the representative ability of features by extracting extra contextual information. Despite their simplicity, all these methods need to spread massive anchors across scales, which causes unnecessary computational burden and memory consumption. Besides, the overall performance depends heavily on the design of anchor priors.
Anchor-free method. To eliminate the efforts for hand-designed anchor priors, anchor-free [9, 20, 28, 29] one-stage methods are proposed and they use points to represent objects. To be specific, CornerNet [9] and ExtremeNet [29] utilizes corners and extreme points to perform object detection in a bottom-up way respectively. CenterNet [28] directly leverages the center points to regress bounding boxes while FCOS [20] uses all points inside a GT box to predict the distances to four boundaries. These anchor-free detectors provide new views for object detection and they surpass the anchor-based counterparts by a large margin. However, ATSS [25] demonstrates that the essential difference between anchor-based and anchor-free methods is the label assignment strategy. When a proper strategy is employed, they could achieve similar performance. Inspired by ATSS, our work focuses on the label assignment method and it's applicable to both anchor-based and anchor-free frameworks.

2.2 Label Assignment Strategy

The label assignment strategy aims to determine the positive and negative samples for detection during training process, which significantly affects the performance of the object detector. Anchor-based detectors [13, 15] assign anchors to objects (as positive) or backgrounds (as negative) based on their IoUs with GT boxes while anchor-free methods [9, 20, 28, 29] utilizes key points (such as corners or centers) to discriminate positive and negative samples. These heuristic strategies are simple and have achieved substantial success. However, they only consider the geometric prior and are lack of flexibility. To overcome the bottleneck of rigid label assignment, MetaAnchor [24] predicts the distribution of anchors by sub-network and adaptively assigns anchors. GuidedAnchor [21] uses feature maps with semantics to predict the shape of anchors. FreeAnchor [26] selects the best anchor based on the loss function in order to improve the matching quality between anchors and

targets. PAA [8] introduces the Gaussian mixture model to simulate the probability distribution of positive and negative samples. Unlike the previous methods, GFL [11] uses IoU score as the target of classification head to combine classification and regression predictions. Noisy Anchor [10] designs a soft label and re-weights the anchor to avoid the noise of the binary classification label. ATSS [25] proposes to separate positive and negative samples by calculating the statistics of IoU and remove the need for fixed thresholds. These excellent works mentioned above have inspired the current work and we make a further step to explore the effective label assignment strategy for one-stage detectors.

3 METHOD

To achieve high performance in one-stage object detection, the label assignment strategy should satisfy the following three rules. First, both semantic and geometric information should be considered when designing a assignment metric. Second, for each ground truth, its corresponding positive samples should be dynamically determined based on the metric score during the training process. Third, the score used for NMS procedure should be compatible with the assignment metric. In this section, we will introduce our Dynamic Label Assignment (DLA) in detail and demonstrate how it fulfills all the requirements mentioned above.

3.1 Assignment Metric

Our assignment metric is composed of two parts, *i.e.*, semantic score and geometric score. For simplicity, we use a linear combination to incorporate them jointly and introduce a trade-off parameter α to control the importance of each part. Mathematically, the designed assignment metric score can be formulated as

$$m = (1 - \alpha) \cdot s + \alpha \cdot g \tag{1}$$

where m indicates the assignment metric while s and g represents semantic score and geometric score, respectively. For semantic score, since there is no available prior knowledge, we directly utilize the classification output from the detector. At the beginning of the training process, the classification branch is initialized to predict low scores for all the categories. Therefore, the semantic score is not discriminative and it can't provide useful information for selecting high-quality samples. To deal with this issue, we initialize α to 1.0 and decrease it exponentially to a certain value α_0 as the training progresses. Thus, the parameter α becomes

$$\alpha = (1.0 - \alpha_0) \cdot e^{-t/\sqrt{T}} + \alpha_0 \tag{2}$$

where t indicates the current training epoch and T represents the total number of epochs. In this way, we can avoid utilizing semantic score at the beginning and gradually increase its importance in the training process. As for the geometric score, the ideal solution is calculating the IoU between the predicted bounding box and the ground truth. However, the regression output is also not reliable when training starts. Therefore, the geometric prior such as center prior or anchor prior is required for stable training. To make geometric score aware of the prediction ability of the detector, we leverage semantic score s to measure how well the detector is trained and design a mixed criteria to automatically adjust the importance of the geometric prior. As a result, the geometric score

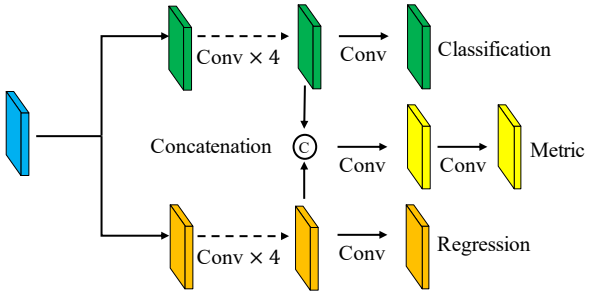

Figure 2: Illustration of the auxiliary task in the detection head.

is calculated as

$$g = (1 - s) \cdot IoU_{prior} + s \cdot IoU_{pred} \tag{3}$$

where IoU_{prior} represents the IoU between the anchor prior and the ground truth while IoU_{pred} is the IoU between the predicted box and the ground truth. By this means, the geometric score is dynamically determined for each prediction. At the beginning, the semantic score s is very low so g is almost equal to IoU_{prior}. When the detector is sufficiently trained, it can produce high semantic score and g depends more on IoU_{pred}.

3.2 Label Assignment

Our label assignment strategy is similar to ATSS [25] and it mainly contains two steps. First, for each ground truth, we select top k samples whose centers are closest to its center based on L2 distance from each pyramid level of FPN [12]. After constructing the candidate bag, we compute the assignment metric m for each sample in the bag. To eliminate the need for hand-designed hyper-parameters such as positive threshold and negtive threshold, we calculate the statistics, *i.e.*, mean m_{mean} and standard deviation m_{std} of the assignment metric. The samples whose metric scores are greater than the threshold $m_{mean} + m_{std}$ are assigned as the positive samples. Then, the other samples are treated as negative samples. Following ATSS, we also limit the positive samples' center inside the ground truth box to stabilize the training process. If a sample is assigned to multiple ground truths, it's only selected for the one with the highest metric score.

3.3 Loss Function

Unlike previous works that directly use classification score as the ranking criteria for NMS post-processing, we introduce an auxiliary task to learn the assignment metric score explicitly. As shown in Figure 2, since assignment metric is a combination of semantic score and geometric score, we utilize both classification and regression features to make predictions jointly. To be more specific, we concatenate the features from two branches together and use a 1×1 convolution to reduce the channel dimension first. Then, a 3×3 convolution is applied to the reduced feature to predict the assignment metric. In the test time, we leverage the metric score to conduct NMS. In this way, the ranking criteria of NMS is aligned with our label assignment strategy. Since we introduce an auxiliary

Table 1: Comparisons with other state-of-the-art methods on COCO $test-dev2017$.

Method	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet [13]	ResNet-101	39.1	59.1	42.3	21.9	42.7	50.2
FCOS w/imprv [20]	ResNet-101	43.0	61.7	46.3	26.0	46.8	55.0
Noisy Anchor [10]	ResNet-101	41.8	61.1	44.9	23.4	44.9	52.9
MAL [7]	ResNet-101	43.6	62.8	47.1	25.0	46.0	55.8
SAPD [30]	ResNet-101	43.5	63.6	46.5	24.9	46.8	54.6
ATSS [25]	ResNet-101	43.6	62.1	47.4	26.1	47.0	53.6
PAA [8]	ResNet-101	44.8	63.3	48.7	26.5	48.8	56.3
DLA (ours)	ResNet-101	**44.6**	63.1	48.5	26.9	48.0	55.2
SAPD [30]	ResNeXt-101-64x4d	45.4	65.6	48.9	27.3	48.7	56.8
ATSS [25]	ResNeXt-101-64x4d	45.6	64.6	49.7	28.5	48.9	55.6
PAA [8]	ResNeXt-101-64x4d	46.6	65.6	50.8	28.8	50.4	57.9
GFL [11]	ResNeXt-101-32x4d	46.0	65.1	50.1	28.2	49.6	56.0
DLA (ours)	ResNeXt-101-64x4d	**46.5**	65.2	50.8	28.9	49.8	58.4

task, our loss function consists of three parts and they are described as follows.

Classification loss. The classification loss is the same as the common practices [13, 20]. For each prediction with a classification score p_i, it has a corresponding binary label l_i. We adopt Focal Loss [13] to tackle the imbalance problem between positive and negative samples and the classification loss can be written as

$$\begin{aligned} L_{cls} &= \frac{1}{N_{pos}} \sum_i FocalLoss(p_i, l_i) \\ &= \frac{1}{N_{pos}} [\sum_{i=1}^{N_{pos}} \alpha(1-p_i)^\gamma \log p_i + \sum_{i=1}^{N_{neg}} (1-\alpha) p_i^\gamma \log(1-p_i)] \end{aligned} \tag{4}$$

where α and γ are the hyper-parameters introduced in [13].

Regression loss. As discussed in [10], the positive samples should not be treated equally during the training process. To be specific, learning from high-quality samples could benefit the detector while those with low metric scores might hurt the detection performance due to their noises. To facilitate the learning procedure, we re-weight the positive samples according to their assignment metric scores and the weight is denoted as

$$w = m^\gamma = [(1-\alpha) \cdot s + \alpha \cdot g]^\gamma \tag{5}$$

where γ is used to control the degree of re-weighting. When γ is less than 1.0, w tends to narrow the gap between different samples. Conversely, if γ is greater than 1.0, the contribution of high-quality positive samples is amplified. For regression task, each positive sample b_i is associated with a ground truth box gt_i and we utilize GIoU Loss [19] to perform optimization. By introducing the re-weighting mechanism, the regression loss can be formulated as

$$L_{reg} = \frac{1}{\sum_{i=1}^{N_{pos}} w_i} \sum_{i=1}^{N_{pos}} w_i GIoULoss(b_i, gt_i) \tag{6}$$

Metric loss. The metric loss is similar to the classification loss and it is a form of the Generalized Focal Loss [11]. Here we do not incorporate the parameter α described in [13] to balance the positive and negative losses. For this auxiliary task, each sample predicts a metric score c_i and its target m_i is a continuous value ranging from 0 to 1. We directly take the original assignment metric m as the learning target and do not apply any transformations to it. Thus, the metric loss can be calculated as

$$L_{metric} = \frac{1}{N_{pos}} [\sum_{i=1}^{N_{pos}} |c_i - m_i|^\gamma BCE(c_i, m_i) + \sum_{i=1}^{N_{neg}} c_i^\gamma BCE(c_i, 0)] \tag{7}$$

where BCE represents the binary cross-entropy loss.

Combining the aforementioned three parts, our loss function can be represented as

$$L = a \cdot L_{cls} + b \cdot L_{reg} + c \cdot L_{metric} \tag{8}$$

where a, b and c are used to balance the contribution of each part.

4 EXPERIMENTS

We conduct all the experiments on the challenging MS COCO benchmark [14]. The MS COCO dataset consists of 80 categories and is split into *train*2017, *val*2017, and *test-dev2017*. Following the common practices [13, 20, 25], we train our models on the *train*2017 split without any extra data. For ablation studies, we evaluate our method on the *val*2017 split. For comparisons with other state-of-the-art methods, we report our results on the *test-dev*2017 split whose labels are not publicly available. The detection performance is measured by the standard COCO-style Average Precision (AP).

4.1 Implementation Details

We adopt the common 'Backbone-FPN-Head' as our pipeline. The backbone is pre-trained on the ImageNet [5] and we choose ResNet [6] and ResNeXt [23] to conduct our experiments. Following ATSS [25], we only tile one anchor as the geometric prior for each position. Unless otherwise stated, we set $\alpha_0 = 0.9$, $k = 10$ and $\gamma = 1.0$. Our DLA is applicable to both anchor-based and anchor-free detectors and we report the results with anchor-free method by default. As with most one-stage detectors [13, 20, 25], the input resolution is set as 1333×800. As for the loss weights, we adopt $a = c = 1.0$ and $b = 2.0$, respectively. Our codebase is based on PyTorch [17] and MMDetection [3]. We train our models on 4 GPUs with 4 images

Table 2: Ablation study on trade-off parameter α.

α	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
dy, 0.9	**40.3**	58.5	43.7	24.2	44.2	52.2
fix, 1.0	40.0	58.1	43.3	23.8	43.9	52.3
fix, 0.9	40.2	58.5	43.5	24.1	44.2	52.1
fix, 0.8	40.1	58.5	43.5	23.5	44.1	52.1

Table 3: Ablation study on sampling number k.

k	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
5	40.2	58.2	43.5	23.6	43.9	52.3
10	40.3	58.5	43.7	24.2	44.2	52.2
15	**40.4**	58.8	43.8	23.4	44.4	52.6
20	39.9	58.2	43.2	23.4	44.3	51.8

per GPU in a mini-batch. We take ATSS as our baseline and all the training hyper-parameters are kept unchanged. The '1×' and '2×' training schedules follow the default settings in MMDetection.

4.2 Main Results

For comparisons with other state-of-the-art methods [7, 10, 13, 20], we train our model with '2×' schedule and adopt scale jitter. We report the performance on $test-dev2017$ under single-model and single-scale test. As shown in Table 1, taking ResNet-101 [6] and ResNeXt-101 [23] as backbone, DLA achieves 44.6 AP and 46.5 AP respectively, which is superior or comparable to other strong works. Notably, PAA [8] is trained under '3×' schedule and it is much longer than the common practices [11, 25, 30]. Without bells and whistles, DLA could significantly increase the upper bound of detection performance for one-stage detectors using the same network structures.

4.3 Ablation Studies

For ablation studies, we use ResNet-50 as the backbone and train the model with '1×' schedule. All the results are reported on the *val*2017 split.

Ablation study on trade-off parameter α. To evaluate the effectiveness of trade-off parameter α, we vary its value and compare the results between fixed and dynamic settings. The corresponding results are shown in Table 2. From the second row to the fourth row, we can see that geometric score plays a more important role in the assignment metric and $\alpha = 0.9$ yields the best performance. Besides, we even achieve 40.0 AP when only geometric score is considered. This is because that we incorporate the semantic score into the formulation of g. As a result, we do not need to emphasize the semantic score explicitly. By introducing dynamic mechanism, our DLA achieves slightly higher performance as shown in the first row of Table 2, which validates the effectiveness of our method.

Ablation study on sampling number k. The sampling number k controls the size of our candidate bag for computing metric statistics and selecting high-quality positive samples. To validate the robustness of our DLA, we conduct ablation study on k by varying the value from 5 to 20. As shown in Table 3, when k is between 5 and 15, the detection performance is very stable and it only fluctuates about 0.2 AP. As k increases to 20, the overall performance drops significantly and we only achieve 39.9 AP. Considering the assignment procedure, we conjecture that a large sampling number k might hurt the average quality of the candidate bag. As a result, more low-quality candidates are selected as the positive samples, which could distract the detector during training process. Therefore, the sampling number k should be limited to a proper range for stable and high performance.

Table 4: Ablation study on weight parameter γ.

γ	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0.0	40.1	58.4	43.3	23.4	44.0	52.1
0.5	40.2	58.5	43.7	23.5	44.1	52.0
1.0	**40.3**	58.5	43.7	24.2	44.2	52.2
1.5	40.0	58.1	43.5	22.8	44.1	52.4

Table 5: Generalization on different frameworks. 'AB' is short for anchor-based and 'AF' is short for anchor-free, respectively.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ATSS AB [25]	39.3	57.5	42.8	24.3	43.3	51.3
DLA AB	**40.0**	58.5	43.2	24.8	44.1	51.7
ATSS AF [25]	39.2	57.3	42.4	22.7	43.1	51.5
DLA AF	**40.3**	58.5	43.7	24.2	44.2	52.2

Ablation study on weight parameter γ. To validate the effectiveness of re-weighting samples in regression task, we conduct experiments by adopting various values of γ. As shown in Table 4, our DLA is robust to the variance of γ, which indicates that the main contribution of our method comes from the label assignment strategy rather than the weight mechanism. Notably, our DLA could achieve 41.1 AP without re-weighting samples ($\gamma = 0.0$). When weight mechanism is applied, $\gamma = 1.0$ is a proper value to control the degree of re-weighting and a large value of γ could deteriorate the detection performance.

Generalization on different frameworks. Similar to ATSS [25], our DLA is also applicable to both anchor-based and anchor-free frameworks. As shown in Table 5, our DLA consistently outperforms the ATSS by a large margin under different situations, which validates the generalization ability of our method. By introducing semantic information and prediction-aware geometric score into the label assignment strategy, the detection performance could be improved significantly. In addition, the extra computational costs are only considered during training process so the efficiency in test time is not influenced at all.

5 CONCLUSION

In this work, we systematically analyze the intrinsic defects of the conventional label assignment strategy in one-stage detection. To address the problems, we propose a new method called DLA as a substitute to the previous strategy. Specifically, we design a

new assignment metric by incorporating semantic and geometric information jointly. To overcome the instability at the beginning of training process, we dynamically adjust the importance of each part and propose prediction-aware geometric score. We also utilize sample re-weighting mechanism to enhance the learning from high-quality samples. With the aforementioned improvements, our DLA achieves state-of-the-art 46.5 AP under single-model and single-scale test, surpassing other strong methods such as ATSS and GFL.

ACKNOWLEDGMENTS

This work was supported by the Scientific Instrument Developing Project of the Chinese Academy of Sciences under Grant YJKYYQ202 00045.

REFERENCES

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. 2017. Soft-NMS - Improving Object Detection with One Line of Code. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 5562–5570. https://doi.org/10.1109/ICCV.2017.593

[2] Kean Chen, Jianguo Li, Weiyao Lin, John See, Ji Wang, Lingyu Duan, Zhibo Chen, Changwei He, and Junni Zou. 2019. Towards Accurate One-Stage Object Detection With AP-Loss. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 5119–5127. https://doi.org/10.1109/CVPR.2019.00526

[3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *CoRR* abs/1906.07155 (2019). arXiv:1906.07155 http://arxiv.org/abs/1906.07155

[4] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. 2016. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 379–387. https://proceedings.neurips.cc/paper/2016/hash/577ef1154f3240ad5b9b413aa7346a1e-Abstract.html

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[7] Wei Ke, Tianliang Zhang, Zeyi Huang, Qixiang Ye, Jianzhuang Liu, and Dong Huang. 2020. Multiple Anchor Learning for Visual Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 10203–10212. https://doi.org/10.1109/CVPR42600.2020.01022

[8] Kang Kim and Hee Seok Lee. 2020. Probabilistic Anchor Assignment with IoU Prediction for Object Detection. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXV (Lecture Notes in Computer Science, Vol. 12370)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 355–371. https://doi.org/10.1007/978-3-030-58595-2_22

[9] Hei Law and Jia Deng. 2018. CornerNet: Detecting Objects as Paired Keypoints. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV (Lecture Notes in Computer Science, Vol. 11218)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer, 765–781. https://doi.org/10.1007/978-3-030-01264-9_45

[10] Hengduo Li, Zuxuan Wu, Chen Zhu, Caiming Xiong, Richard Socher, and Larry S. Davis. 2020. Learning From Noisy Anchors for One-Stage Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 10585–10594. https://doi.org/10.1109/CVPR42600.2020.01060

[11] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. 2020. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/f0bda020d2470f2e74990a07a607ebd9-Abstract.html

[12] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. 2017. Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 936–944. https://doi.org/10.1109/CVPR.2017.106

[13] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2999–3007. https://doi.org/10.1109/ICCV.2017.324

[14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 8693)*, David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single Shot MultiBox Detector. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 9905)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer, 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

[16] Jing Nie, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. 2019. Enriched Feature Guided Refinement Network for Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 9536–9545. https://doi.org/10.1109/ICCV.2019.00963

[17] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 8024–8035. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[18] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 91–99. https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html

[19] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 658–666. https://doi.org/10.1109/CVPR.2019.00075

[20] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. 2019. FCOS: Fully Convolutional One-Stage Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 9626–9635. https://doi.org/10.1109/ICCV.2019.00972

[21] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. 2019. Region Proposal by Guided Anchoring. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2965–2974. https://doi.org/10.1109/CVPR.2019.00308

[22] Tiancai Wang, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. 2019. Learning Rich Features at High-Speed for Single-Shot Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 1971–1980. https://doi.org/10.1109/ICCV.2019.00206

[23] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 5987–5995. https://doi.org/10.1109/CVPR.2017.634

[24] Tong Yang, Xiangyu Zhang, Zeming Li, Wenqiang Zhang, and Jian Sun. 2018. MetaAnchor: Learning to Detect Objects with Customized Anchors. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 318–328. https://proceedings.neurips.cc/paper/2018/hash/69adc1e107f7f7d035d7baf04342e1ca-Abstract.html

[25] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. 2020. Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 9756–9765. https://doi.org/10.1109/CVPR42600.2020.00978

[26] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. 2019. FreeAnchor: Learning to Match Anchors for Visual Object Detection. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 147–155. https://proceedings.neurips.cc/paper/2019/hash/43ec517d68b6edd3015b3edc9a11367b-Abstract.html

[27] Zhishuai Zhang, Siyuan Qiao, Cihang Xie, Wei Shen, Bo Wang, and Alan L. Yuille. 2018. Single-Shot Object Detection With Enriched Semantics. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 5813–5821. https://doi.org/10.1109/CVPR.2018.00609

[28] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. 2019. Objects as Points. *CoRR* abs/1904.07850 (2019). arXiv:1904.07850 http://arxiv.org/abs/1904.07850

[29] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krähenbühl. 2019. Bottom-Up Object Detection by Grouping Extreme and Center Points. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 850–859. https://doi.org/10.1109/CVPR.2019.00094

[30] Chenchen Zhu, Fangyi Chen, Zhiqiang Shen, and Marios Savvides. 2020. Soft Anchor-Point Object Detection. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX (Lecture Notes in Computer Science, Vol. 12354)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 91–107. https://doi.org/10.1007/978-3-030-58545-7_6