Foreground-Background Imbalance Problem in Deep Object Detectors: A Review

Joya Chen[1], Qi Wu[2], Dong Liu[1], and Tong Xu[1]

[1]University of Science and Technology of China, Hefei 230027, China
[2]Institute of Intelligent Machines, Chinese Academy of Sciences
{chenjoya, wqdfj}@mail.ustc.edu.cn {dongeliu, tongxu}@ustc.edu.cn

Abstract

Recent years have witnessed the remarkable developments made by deep learning techniques for object detection, a fundamentally challenging problem of computer vision. Nevertheless, there are still difficulties in training accurate deep object detectors, one of which is owing to the foreground-background imbalance problem. In this paper, we survey the recent advances about the solutions to the imbalance problem. First, we analyze the characteristics of the imbalance problem in different kinds of deep detectors, including one-stage and two-stage ones. Second, we divide the existing solutions into two categories: sampling heuristics and non-sampling schemes, and review them in detail. Third, we experimentally compare the performance of some state-of-the-art solutions on the COCO benchmark. Promising directions for future work are also discussed.

1. Introduction

Object detection, which consists of recognizing the categories and identifying the locations of object instances that appeared in an image, attracts numerous research efforts for several decades. As a fundamental task in computer vision, it is the basis for solving more complex and high-level vision tasks, e.g. instance segmentation [12], image caption [1], and scene understanding [36]. Moreover, it plays a key role in a series of real-world applications, such as autonomous driving, robotic vision, and video surveillance.

In earlier years, the sliding window paradigm with handcrafted features [9, 34] was widely used for detecting objects. With the rapid development of deep learning techniques [18], deep object detectors [2, 12, 14, 16, 17, 20, 21, 22, 25, 28, 29, 30, 31, 33, 37, 39, 40] quickly come to dominate the research of object detection, and have substantially pushed the detection accuracy forward. Despite the apparent differences in various detection architectures, e.g. one-stage [25, 28] versus two-stage [31], previous works [19, 22, 25, 31, 32] reveal that the *foreground-background imbalance* problem universally exists in training object detectors, i.e. there is extreme inequality between the number of foreground examples and the number of background examples. Strong evidence [19, 22, 25, 32] has shown that the imbalance problem impedes detectors from achieving a higher detection accuracy.

In this paper, we thoroughly review the recent advances in solving the foreground-background imbalance problem. Firstly, as the imbalance problem incurs different consequences among various object detectors, we carefully analyze the characteristics of the imbalance for different object detectors, i.e., anchor-based one-stage, anchor-free one-stage, and two-stage approaches. Subsequently, we divide the solutions into two groups: sampling heuristics [3, 19, 22, 26, 31, 32] and non-sampling schemes [4, 5, 6, 27], and systematically review the existing solutions for the imbalance problem in detail. Meanwhile, a comparison of their performance is given. Finally, several promising directions are discussed to inspire future research.

1.1. Scope

As a longstanding difficulty, the class imbalance problem has been studied for a long while in machine learning research. While the foreground-background imbalance problem in deep object detectors could also be viewed as a class imbalance problem, it is attributed to the large searching space of detectors, rather than the usual causes such as data

This work was supported by the National Key Research and Development Program of China under Grant 2018YFA0701603, and by the Natural Science Foundation of China under Grants 61772483 and 61931014. *(Corresponding author: Dong Liu.)*

distribution (i.e., due to biased dataset). Therefore, we only discuss the foreground-background imbalance in object detection. Furthermore, as the state-of-the-art performance is often achieved by deep object detectors, we will ignore the imbalance solutions in classic non-deep object detectors.

1.2. Comparison with Previous Reviews

Several surveys e.g. [24] have comprehensively review the object detection tasks, datasets, metrics, and methods. However, they did not specifically discuss the imbalance problems of object detection in detail. Oksuz et al. [15] provide a review for different kinds of imbalance problems in object detection, including class imbalance, scale imbalance, spatial imbalance, and objective imbalance. They did not focus on the imbalance between foregrounds and backgrounds in deep object detectors. We pay attention to the foreground-background imbalance problem and provide a more dedicated review of the solutions to this problem.

1.3. Paper Organization

This paper is organized as follows: Section 2 introduces the research background of deep object detectors with the explanation of the foreground-background imbalance problem. Section 3 describes the solutions to the foreground-background imbalance problem in detail and compares the performance between different solutions. Section 4 concludes the paper and discusses several promising directions.

2. Research Background

Here we briefly introduce the deep object detectors and the foreground-background imbalance problem. Following the previous work of [24], we summarize various deep object detectors into *one-stage* and *two-stage* approaches, but further divide one-stage approach into *anchor-based* one-stage detectors and *anchor-free* one-stage detectors, as the foreground-background imbalance problem has different characteristics for them. For each category, we will analyze what causes the foreground-background imbalance.

2.1. Imbalance in One-Stage Object Detectors

Anchor-Based One-Stage Detectors. With the dense, predefined bounding-boxes (i.e., anchors [31]) tiled over an image, anchor-based one-stage detectors [19, 22, 25, 29, 30, 38] could directly recognize objects by refining the locations and classifying the category of these anchors. Early representatives include SSD [25] and YOLOv2 [29] that manage to predict objects on multiple feature levels, which achieve impressive speed/accuracy trade-off.

However, there is a large gap between the foreground examples and the background example (e.g. $\sim$100 vs. $\sim 100k$) during training, i.e., foreground-background imbalance. As illustrated in previous works [19, 22, 25], this imbalance would impede anchor-based one-stage detectors from becoming more accurate. RetinaNet [22], RefineDet [38], and GHM [19] explore different solutions for addressing the imbalance, yielding much better detection accuracy.

Anchor-Free One-Stage Detectors. As anchors would introduce multiple hyper-parameters to determine the shape (e.g. scales, aspect ratios), some researchers have started to explore an anchor-free paradigm. Early efforts include DenseBox [14], YOLO [28], and CornerNet [17], which rely on the central region, the fixed cell, and the key point to determine the initial location, respectively. Their successors could be divided into two categories: center-based [16, 33, 39] and point-based [37, 40] frameworks. Some two-stage approaches also draw lessons from anchor-free one-stage pipelines, e.g. GA-RPN [35].

In practice, both the key points and the central regions of objects only occupy a small part of the image, while the majority in the image is the background. Despite that anchor-free approaches discard dense anchors to cover objects by key-points/the central regions, they still suffer from the imbalance caused by the overwhelming number of background points or regions, which could be identified as a foreground-background imbalance problem. Therefore, it is not strange that most anchor-free ones apply Focal Loss [22] or its variants to address the foreground-background imbalance.

2.2. Imbalance in Two-Stage Object Detectors

To date, two-stage (region-based) object detectors lead the top accuracy on several benchmarks [8, 23], which shows superiority over one-stage ones in terms of the detection accuracy. These approaches are mainly based on the architecture of Faster R-CNN [31], which firstly generates a sparse set of candidate object proposals by a RPN [31], then determine the accurate bounding boxes and the classes by convolutional networks. A large number of R-CNN variations [2, 12, 20, 21, 26, 35] appear over the years, yielding a large improvement in detection accuracy.

Similar to one-stage detectors, the imbalance problem also exists in two-stage detectors. Firstly, the proposal stage could be viewed as an anchor-based one-stage detector for binary classification (i.e., foreground or background), thus the RPN usually suffers from an extreme imbalance, which requires to apply mini-batch sampling heuristics to alleviate the imbalance. After RPN filters vast background examples, the remaining examples still contain a large number of background examples (e.g., the foreground-to-background ratio is $\sim 1:10$). Therefore, the per-region stage is also equipped with mini-batch sampling heuristics.

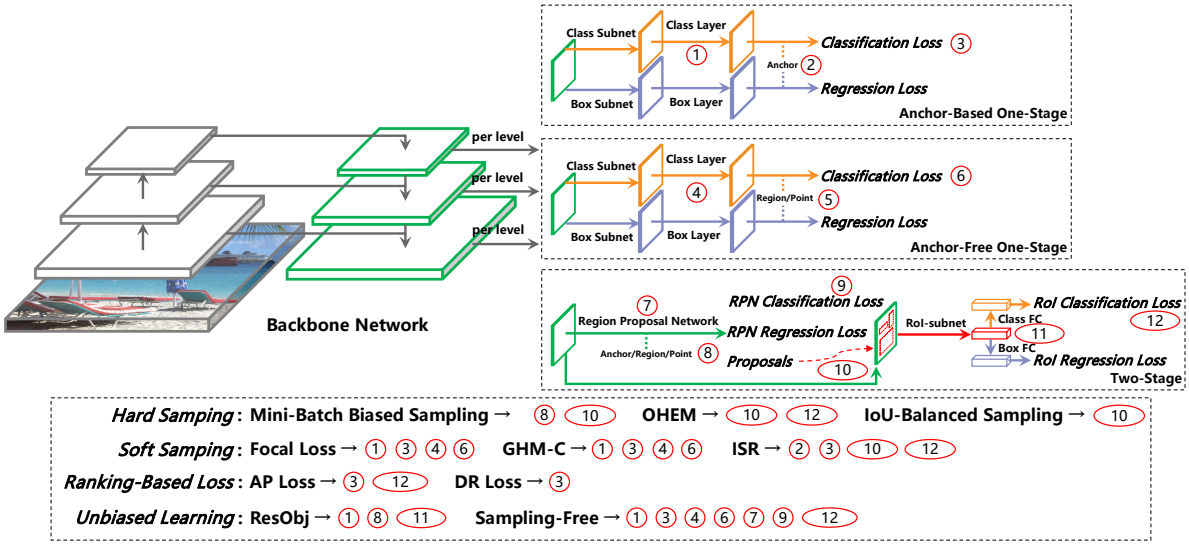

Figure 1. This figure concludes different solutions for addressing the foreground-background imbalance problem in various object detection frameworks (i.e., anchor-based one-stage, anchor-free one-stage, two-stage approaches). These solutions include mini-batch biased sampling [31], OHEM [32], IoU-balanced sampling [26], Focal Loss [22], GHM-C [19], ISA [3], ResObj [4], sampling-free [5], AP Loss [6], DR Loss [27]. We visualize their scope of the usage in the detection pipeline.

3. Solutions

We divide the solutions for addressing the foreground-background imbalance problem into two groups: (1) sampling heuristics, including hard [26, 31, 32] and soft sampling heuristics [3, 19, 22]; (2) non-sampling schemes, consisting of ranking-based loss functions [6, 27] and unbiased learning mechanisms [4, 5]. In this section, we will describe these solutions in detail.

3.1. Sampling Heuristics

In essence, sampling heuristics addresses the imbalance by changing the contribution of each example e.g. put more focuses on rare foreground examples:

$$L_i^{cls} = \Omega(G, E, P, i) \cdot CE(p_i, g_i), \tag{1}$$

where $CE()$ indicates the cross-entropy cost. L_i^{cls}, p_i and g_i denotes the classification loss, predicted probability and the ground-truth label of i-th example, respectively. We denote $\Omega()$ as the sampling heuristics, i.e., given the input ground-truths G, examples E, predictions P, and example index i, $\Omega(G, E, P, i)$ outputs the weighting factor for i-th example. For simplicity, we denote N_X as the number of X. Based on Figure 1 and Equation 1, we describe different sampling heuristics as follows.

3.1.1 Hard Sampling

Hard sampling selects a part of training examples while ignores others, i.e., $\Omega(G, E, P, i) \in \{0, 1\}$. We further denote $Pr(G, E, P, i)$ as the probability of $\Omega(G, E, P, i) = 1$.

Mini-Batch Biased Sampling. It is widely used in two-stage approaches [31, 21], which randomly selects examples by a predefined foreground-to-background ratio fg/bg and the required number of examples N_Q. For example, a default setting in Faster R-CNN [31, 21] is, randomly sampling 256 anchors, 512 RoIs with a ratio of $fg/bg = 1/1$, $fg/bg = 1/3$ in RPN, RoI-subnetwork, respectively. For example, $Pr(G, E, P, i)_{g_i>0} = (N_Q/N_{E_{g_i>0}}) \cdot [fg/(fg + bg)]$ is the sampled probability of foreground examples.

OHEM. Online hard example mining (OHEM [32]) prefers to sample harder examples than easier ones. A typical form is, $Pr(G, E, P, i) = 1$ only if $CE(p_i, g_i)$ within top 512 values, otherwise $Pr(G, E, P, i) = 0$. Compared with the mini-batch biased sampling, OHEM needs additional memory and causes more training time.

IoU-Balanced Sampling. IoU-balanced sampling [26] assumes that backgrounds with higher IoU to ground-truths tend to be the harder ones. It evenly splits the IoU interval (e.g. $[0, 0.5)$) into K bins, then randomly selects examples in these bins. Therefore, IoU-balanced sampling saves the extra loss computation required by OHEM. Note that this is only applied for backgrounds, e.g. in k-th bin, $Pr(G, E, P, i)^k_{g_i=0} = (N_{Q_{g_i=0}}/N_{E_{g_i=0}}) \cdot (1/N_{E^k_{g_i=0}})$.

3.1.2 Soft Sampling

Soft sampling scales the contribution $\Omega(G, E, P, i)$ of each example during training. Unlike hard sampling, no example is discarded in this way.

Focal Loss. The well-known Focal Loss [22] modifies the standard cross-entropy loss to dynamically down-weight the contribution of easy examples:

$$\Omega(G, E, P, i) = \alpha(1 - |g_i - p_i|)^{\gamma}, \tag{2}$$

where α and γ are the introduced hyper-parameters. Furthermore, it introduces a biased initialization method in the classification layer (See Figure 1). As reported in the original paper, the RetinaNet [22] achieves the optimal accuracy on COCO [23] when $\alpha = 0.25$ and $\gamma = 2$.

GHM-C. Gradient Harmonizing Mechanism (GHM) [19] claims that the imbalance of examples with different attributes (hard/easy and pos/neg) can be implied by the distribution of gradient norm. It also considers that the density of examples with very large gradient norm (very hard examples) as outliers. Therefore, the classification part of GHM (i.e., GHM-C) focuses on the harmony of gradient contribution, i.e., $\Omega(G, E, P, i) = N_E/GD(|g_i - p_i|)$, where $GD(g_i - p_i)$ is the gradient density of i-th example. Note that it follows the biased initialization used in Focal Loss.

ISR. Importance-based Sample Reweighting (ISR) is the classification part of PISA [3]. It proposes to weight an example according to the evaluation metric (i.e., mAP [8, 23]). In other words, a part of examples is more important than others in an mAP metric, thus ISR puts more focuses on them. To find the important examples, ISR is built on HLR (IoU-Hierarchical Local Rank), which could select examples that have a large impact on the mAP.

3.2. Non-Sampling Schemes

More recently, some studies propose several non-sampling schemes, which discard sampling heuristic to maintain the distribution of examples. We divide them into two groups: ranking-based loss functions [6, 27] and unbiased learning mechanisms [4, 5].

3.2.1 Ranking-Based Loss Functions

AP Loss. Instead of modifying the classification loss, AP Loss [6] proposes to replace the classification task with a ranking task to optimize average precision (AP). It proposes a novel error-driven learning algorithm to effectively optimize the non-differentiable AP based objective function. More specifically, some extra transformations are added to the output score of an one-stage detector to obtain the AP Loss, which includes a linear transformation that transforms the scores to pairwise differences, and a non-linear and non-differentiable "activation function" that transform the pairwise differences to primary terms of AP Loss. Then the AP Loss can be obtained by the dot product between the primary terms and the label vector.

DR Loss. Similar to AP Loss [6], distributional ranking (DR) Loss [27] manages to address the imbalance by viewing the classification task as a ranking task. It introduces the DR loss to rank the constrained distribution of foreground above that of background candidates. By reweighting the candidates to derive the distribution corresponding to the worst-case loss, the loss can focus on the decision boundary between foreground and background distribution. Besides, DR Loss ranks the expectation of distributions instead of original examples, which reduces the number of pairs in ranking and improves efficiency.

3.2.2 Unbiased Learning Mechanisms

ResObj. The complicated, heuristic sampling methods could be replaced by a learning-based algorithm. Residual Objectness (ResObj [4]) is a fully learning-based algorithm, which utilizes multiple cascaded objectness-related modules to address the imbalance. ResObj first transfers the imbalance to the objectness module, to down-weight the contributions of overwhelming backgrounds. Subsequently, by building residual connections between objectness-related modules, they reformulate the objectness estimation to a consecutive refinement procedure, thereby progressively addressing the imbalance.

Sampling-Free. Sampling-Free mechanism [5] is proposed as an alternative to sampling heuristics, which manages to maintain the training stability from initialization. It demonstrates that sampling heuristics for different types of object detectors could be discarded, while similar or better accuracy could be achieved. The authors of [5] observed that, unlike common class imbalance that is introduced by data distribution, the foreground-background imbalance should be attributed to the large searching space of detectors, which means it equally exists in training and inference with the same distribution. But sampling heuristics will change this distribution during training, thereby resulting in a mismatch between training and inference. Under their observation, the obstacle that impedes detectors without sampling from yielding high accuracy should be attributed to the instability during training. Motivated by this, they develop a sampling-free mechanism [5], consisting of three schemes: (1) the optimal bias initialization scheme enables the training to be fastly converged under the imbalance; (2) the guided loss scheme avoids the classification loss to be dominated by numerous background examples; (3) the class-adaptive threshold scheme mitigates the confidence shifting problem incurred by the imbalance.

Solutions	Abbreviation	Detector (ResNet-50-FPN)	Codebase	AP	ΔAP	Parameters	Speed
Hard Sampling	Biased Sampling [31]	Faster R-CNN [31]	maskrcnn-benchmark	36.8	-	2	-
			mmdetection	36.4	-	2	-
	OHEM [32]	Faster R-CNN	mmdetection	36.6	+0.2	2	↓
	IoU-balanced sampling [26]	Faster R-CNN	mmdetection	36.8	+0.4	3	↓
Soft Sampling	Focal Loss [22]	RetinaNet	maskrcnn-benchmark	36.3	-	2	-
			mmdetection	35.6	-	2	-
			detectron	33.9 (600×)	-	2	-
				35.7 (800×)	-	2	-
	GHM-C [19]	RetinaNet	mmdetection	35.8	+0.2	1	↓
	ISR [3]	Faster R-CNN	mmdetection	37.9	+1.5	4	↓
Ranking-Based Loss	AP Loss [6]	RetinaNet	detectron	35.0	+1.1	2	↓
	DR Loss [27]	Faster R-CNN	detectron	37.2	+1.5	4	↓
Unbiased Learning	ResObj [4]	RetinaNet	detectron	35.4	+1.3	More conv	↓
	Sampling-Free [5]	RetinaNet	maskrcnn-benchmark	36.6	+0.3	**0**	↑
		Faster R-CNN	maskrcnn-benchmark	**38.4**	**+1.6**	**0**	↓

Table 1. This table illustrates the comparison of different solutions for addressing the foreground-background imbalance problem. We report the accuracy (AP), relative accuracy improvement (ΔAP), the number of hyper-parameters (Parameters), and efficiency (Speed) to compare them. These results come from the well-known detection codebases (i.e., maskrcnn-benchmark **[10],** mmdetection **[7],** detectron **[11]). "-" indicates the baseline models. If not specified, we report the performance achieved with backbone ResNet-50-FPN [13, 21], input scale** 1333×800**, evaluated on COCO** `minival` **[23]. We observed that sampling-free [5] achieves the highest AP and ΔAP without introduced hyper-parameters. ISR [3] and DR Loss [6] also achieve impressively relative accuracy improvement.**

3.3 Comparison

To better review the solutions, we comprehensively compare their performances in Table 1. These results are either from the ablation study of the original paper or from the model description page (i.e., MODEL_ZOO.md) of the public codebases. For a fair comparison, we select RetinaNet [22] and Faster R-CNN [31] as the baseline models, which apply mini-batch biased sampling and Focal Loss to address the imbalance, respectively. We compare these solutions from four aspects: accuracy, relative accuracy improvement, required hyper-parameters and training speed (compared to without the solution).

Accuracy. As presented in Table 1, Sampling-Free and ISR achieve the highest and the second-highest accuracy, respectively. Meanwhile, they keep this advantage in the "ΔAP", which refers to the relative accuracy improvement.

Hyper-Parameters. See the "Parameters" column in Table 1. Despite ISR and DR Loss obtain the second-highest "ΔAP", they introduce the most hyper-parameters (4 hyper-parameters). In contrast, sampling-free introduces no hyper-parameter, whereas other solutions introduce at least 1 hyper-parameter.

Training Speed. As the reported training speed may be evaluated on different devices, the speed comparison may not be fair. Instead, we compare the speed relative to the baseline models. The "Speed" column in Table 1 shows that the majority of solutions are slower than baseline models, except for RetinaNet with the sampling-free mechanism.

4. Discussions

In this paper, we present a comprehensive review of the foreground-background imbalance in deep object detectors. Our review first analyzes different deep object detectors and explain the causes of the foreground-background imbalance. Subsequently, we categorize and describe the existing solutions for addressing the imbalance, including sampling heuristics and non-sampling schemes. Finally, we experimentally compare the performances of different solutions. Based on these summarizations, we discuss two important questions about the foreground-background imbalance:

First, what is the cause of the foreground-background imbalance problem? It appears to be the large searching space of deep object detectors. To date, both anchor-based and anchor-free frameworks follow a dense prediction paradigm, which produces numerous background examples during training. This imbalance may disappear if a detector with a small searching space would be developed.

Second, which solution should be used? As shown in Table 1, Sampling-Free [5] has achieved the highest relative accuracy improvement in two-stage approaches, without introduced hyper-parameters. DR Loss [27] achieves the highest relative accuracy improvement in one-stage approaches. Despite more hyper-parameters introduced in DR Loss, its success suggests that a solution should be developed according to an evaluation metric, as mAP is a ranking-related metric. If there are multiple evaluation metrics in the real-world application, we recommend the sampling-free mechanism as a baseline method, whose code can be found at https://github.com/ChenJoya/sampling-free.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.

[2] Z. Cai and N. Vasconcelos. Cascade R-CNN: delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018.

[3] Y. Cao, K. Chen, C. C. Loy, and D. Lin. Prime sample attention in object detection. *CoRR*, abs/1904.04821, 2019.

[4] J. Chen, D. Liu, B. Luo, X. Peng, T. Xu, and E. Chen. Residual objectness for imbalance reduction. *CoRR*, abs/1908.09075, 2019.

[5] J. Chen, D. Liu, B. Luo, T. Xu, S. Zhang, S. Wu, B. Luo, X. Peng, and E. Chen. Is sampling heuristics necessary in training deep object detectors? *CoRR*, abs/1909.04868, 2019.

[6] K. Chen, J. Li, W. Lin, J. See, J. Wang, L. Duan, Z. Chen, C. He, and J. Zou. Towards accurate one-stage object detection with ap-loss. In *CVPR*, pages 5119–5127, 2019.

[7] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin. Mmdetection: Open mmlab detection toolbox and benchmark. *CoRR*, abs/1906.07155, 2019.

[8] M. Everingham, L. J. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[9] P. F. Felzenszwalb, R. B. Girshick, and D. A. McAllester. Cascade object detection with deformable part models. In *CVPR*, pages 2241–2248, 2010.

[10] M. Francisco and G. Ross. maskrcnn-benchmark. github.com: facebookresearch/maskrcnn-benchmark, 2018.

[11] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He. Detectron. github.com: facebookresearch/detectron.

[12] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988, 2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[14] L. Huang, Y. Yang, Y. Deng, and Y. Yu. Densebox: Unifying landmark localization with end to end object detection. *CoRR*, abs/1509.04874, 2015.

[15] Kemal, B. C. Cam, S. Kalkan, and E. Akbas. Imbalance problems in object detection: A review. *CoRR*, abs/1909.00169, 2019.

[16] T. Kong, F. Sun, H. Liu, Y. Jiang, and J. Shi. Foveabox: Beyond anchor-based object detector. *CoRR*, abs/1904.03797, 2019.

[17] H. Law and J. Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, pages 765–781, 2018.

[18] Y. LeCun, Y. Bengio, and G. E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[19] B. Li, Y. Liu, and X. Wang. Gradient harmonized single-stage detector. In *AAAI*, pages 8577–8584, 2019.

[20] Y. Li, Y. Chen, N. Wang, and Z. Zhang. Scale-aware trident networks for object detection. In *ICCV*, pages 6054–6063, 2019.

[21] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944, 2017.

[22] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, pages 2999–3007, 2017.

[23] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014.

[24] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen. Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, pages 1573–1405, 2019.

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. SSD: single shot multibox detector. In *ECCV*, pages 21–37, 2016.

[26] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin. Libra R-CNN: towards balanced learning for object detection. In *CVPR*, pages 821–830, 2019.

[27] Q. Qian, L. Chen, H. Li, and R. Jin. DR loss: Improving object detection by distributional ranking. *CoRR*, abs/1907.10156, 2019.

[28] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[29] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. In *CVPR*, pages 6517–6525, 2017.

[30] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018.

[31] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017.

[32] A. Shrivastava, A. Gupta, and R. B. Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, pages 761–769, 2016.

[33] Z. Tian, C. Shen, H. Chen, and T. He. FCOS: fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019.

[34] P. A. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[35] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin. Region proposal by guided anchoring. In *CVPR*, pages 2965–2974, 2019.

[36] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh. Graph R-CNN for scene graph generation. In *ECCV*, pages 690–706, 2018.

[37] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin. Reppoints: Point set representation for object detection. In *ICCV*, pages 9657–9666, 2019.

[38] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Single-shot refinement neural network for object detection. In *CVPR*, pages 4203–4212, 2018.

[39] X. Zhou, D. Wang, and P. Krähenbühl. Objects as points. *CoRR*, abs/1904.07850, 2019.

[40] X. Zhou, J. Zhuo, and P. Krähenbühl. Bottom-up object detection by grouping extreme and center points. In *CVPR*, pages 850–859, 2019.