A Comparative Analysis of Single-Stage Detectors from the Perspectives of Anchor-Free and Anchor-Based Approaches

Peigeng Li *

Glasgow College, University of Electronic Science and Technology of China, Chengdu, China

* Corresponding Author Email: 2720731L@student.gla.ac.uk

Abstract. Since the emergence of deep learning in object detection in 2014, there has been rapid development and widespread application of detectors across various domains such as healthcare, road traffic, and industrial inspection. The evolution of convolutional neural network (CNN)-based detectors has reached a mature stage, with single-stage detectors playing a crucial role within CNN-based frameworks. This paper provides an analysis of the developmental process and distinctive features of classic single-stage detectors categorized into two types: Anchor-free and Anchor-based. Furthermore, the paper also provides a comparison of the accuracy trends for these two types of single-stage detectors over different years. Through this analysis and comparison, potential reasons behind the advancements in single-stage detectors during different time periods can be explored, such as network architecture optimization, increase in datasets, and algorithm improvements. Additionally, this paper offers insights into current research focal points and future prospects.

Keywords: Anchor-free, anchor based, one-stage detector.

1. Introduction

Single-stage detectors are a crucial component of convolutional neural network-based detectors and play a pivotal role in real-time computer vision tasks. The primary metrics for object detection tasks encompass speed and accuracy. Unlike two-stage detectors, single-stage detectors eliminate the need for a pre-detection stage to generate bounding boxes, directly providing class, probability, and position information of detected objects. Consequently, they exhibit remarkable detection speeds suitable for real-time applications such as face recognition, traffic control, and text extraction that necessitate prompt feedback. Over the past nine years, continuous optimization has addressed numerous initial challenges encountered during the development of single-stage detectors. As a result of these optimizations, their accuracy has significantly improved and is now approaching that of two-stage detectors while maintaining their advantage in terms of detection speed. Furthermore, the development process of single-stage detectors exhibits an interesting characteristic wherein it transitions from anchor-free to anchor-based approaches before reverting back to anchor-free methods.

The objective of this review is to offer a multi-faceted perspective and enhance comprehension of single-stage detectors. This article encompasses a relatively comprehensive overview of the developmental history of single-stage detectors, qualitative analysis of the characteristics exhibited by classical detectors, and comparisons among different detector models. Through the aforementioned analysis, discernible evolutionary trends between Anchor-free and Anchor-based approaches are identified, accompanied by speculation on potential reasons for these trends. Furthermore, future research directions are examined while providing an enhanced understanding.

The remaining sections of this article are presented in a chronological order, providing comprehensive explanations of the detection principles, optimization methods, and accuracy for both Anchor-free and Anchor-based classic detectors. Subsequently, the second section compares these detectors based on datasets, evaluation metrics, and their respective accuracies.

2. One-stage detector

Single-stage object detection enables efficient inference by obtaining comprehensive information about detected objects in a single step. However, it encounters challenges in accurately detecting small and dense objects [1]. To overcome this limitation, various detectors have been developed to enhance the detection of such objects in subsequent advancements. Single-stage object detectors can be categorized into two types: Anchor-free and Anchor-based. This article categorizes several significant single-stage object detection models and compares them based on their chronological development, accuracy, and other relevant factors.

2.1. Anchor-free detector

In the process of object detection, an Anchor-free detector is employed to extract features from the network, identify feature points, and determine the precise object position through cost calculation. The utilization of this flexible and efficient Anchor-free algorithm has demonstrated time-saving benefits and reduced computational complexity. However, due to its reliance on feature points for object localization, it poses challenges in detecting overlapping or occluded objects within a region containing a single feature point. Nevertheless, recent advancements in Anchor-free algorithms have shown promising progress towards addressing this limitation.

2.1.1. YOLO

The YOLO series detectors encompass both Anchor-free and Anchor-based detectors, representing classic one-stage detection algorithms. YOLOv1, proposed in 2015, is a single-stage object detection algorithm that partitions the image into numerous regions (as depicted in Figure 1). It predicts bounding boxes and probabilities for each region [1]. This detection approach enables YOLOv1 to achieve remarkable speed, with a base network running at 45fps on Titan X GPU, while its fast version surpasses 150 fps, facilitating real-time detection [2]. The simplicity of YOLOv1's structure and adoption of a single regression method significantly enhance its detection speed. However, these high-speed characteristic compromises accuracy due to fewer steps involved. The fast version of YOLOv1 attained an mAP@0.5 accuracy of 52.7% on the PASCAL VOC2007 dataset; whereas its enhanced version achieved higher accuracy with a mAP value of 63.4% [1]. Several factors contribute to larger errors: firstly, YOLOv1 encounters difficulties in predicting objects with aspect ratios absent from the training dataset; secondly, it can only learn coarse features of objects due to limited learning capacity; thirdly, within a single grid cell, YOLOv1 can solely detect two identical objects—a limitation also applicable to Anchor-free detectors as they struggle with predicting multiple objects at a single feature point [2]. Starting from YOLOv2 onwards, the evolution trajectory of the YOLO series has shifted towards anchor-based methods until reaching YOLOX in 2021 before returning to anchor-free techniques through YOLOv6 in 2022. Both YOLOX and YOLOv6 achieved accuracies of mAP@0.5=50.1% and mAP@0.5=70%, respectively on MS COCO dataset [1].

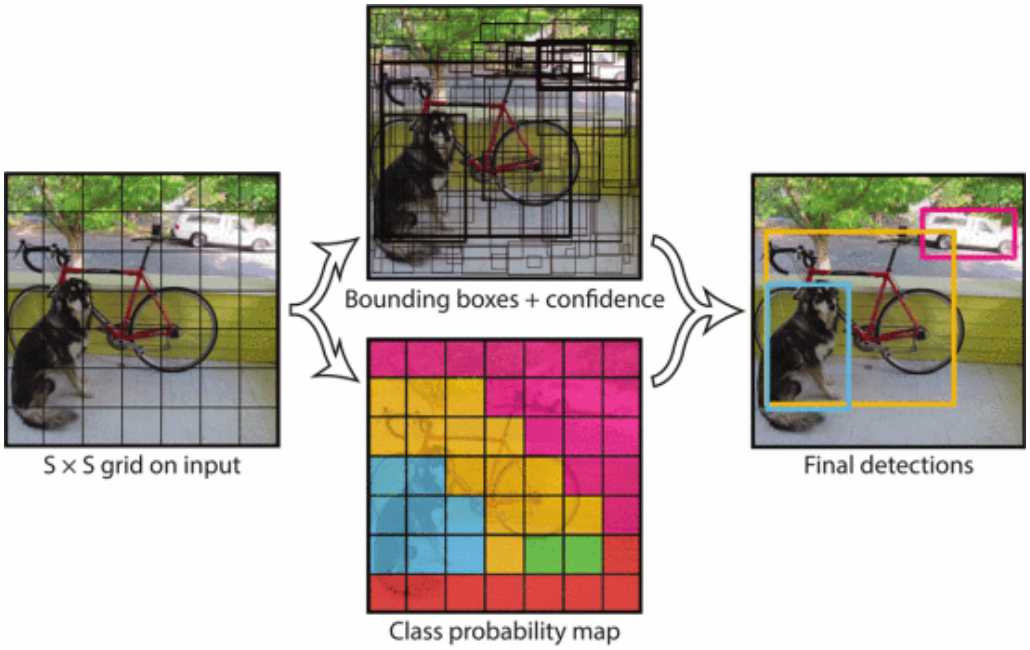

Figure 1. YOLOv1 model. This system models object detection as a regression problem, dividing the image into grids and bounding boxes [1].

2.1.2. CornerNet

Corner Net, proposed in 2018, is based on the concept of corner pooling as its core mechanism. It deviates from traditional detection boxes and instead utilizes reference points, which are a pair of diagonal points derived from the original anchor box, for object detection (Figure 2). These diagonal points are then combined and separated to form bounding boxes by leveraging additional embedded information. By reducing the multitude of diverse reference boxes caused by variations in object shape, quantity, and position, Corner Net achieves reduced energy consumption and improved suitability for real-world scenarios. Moreover, Corner Net incorporates FPN and Hourglass structures to enhance accuracy. As a result, Hourglass achieves higher precision with a mAP@.5=57.8% on the MS COCO dataset [3]. A year later, Extreme Net emerged with two major distinctions from Corner Net: it replaces predicting object detection through predicting corner points with predicting extreme points and center point; furthermore, instead of employing embedded information recombination and decoupling process like Corner Net does, Extreme Net employs brute force enumeration of extreme points to determine their grouping for generating bounding boxes.

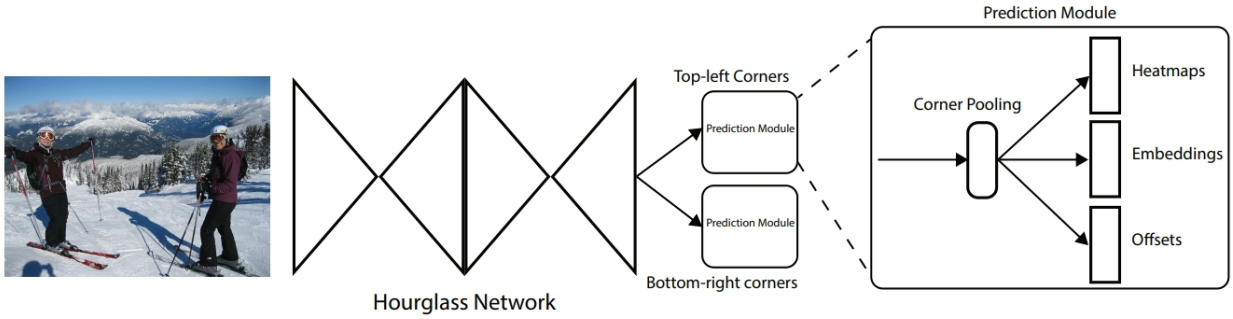

Figure 2. Two prediction modules are positioned behind the backbone network in Figure 2, with one module dedicated to locating and grouping the top left corner, while the other focuses on identifying the bottom right corner [3].

2.1.3. CenterNet

After just one year since the proposal of Corner Net in 2019, the detector Center Net was introduced. Center Net enhances the foundation laid by Corner Net through the incorporation of center keypoints, enabling simultaneous analysis of both boundary and interior information during

the detection process. Similar to Corner Net and Extreme Net, its detection procedure involves initial identification of object keypoints followed by calculation of object size (bounding box) [4]. Moreover, Center Net introduces two pooling techniques - center pooling and cascade corner pooling - to effectively extract keypoints for this detector. Serving as an end-to-end detection network, Center Net eliminates post-processing steps; during inference, it obviates operations such as Non-Maximum Suppression (NMS), requiring only necessary forward processing. Despite dispensing with post-processing steps, CenterNet achieves remarkable accuracy with a mean Average Precision (mAP) score of 61.1% on the MS COCO dataset [1, 5].

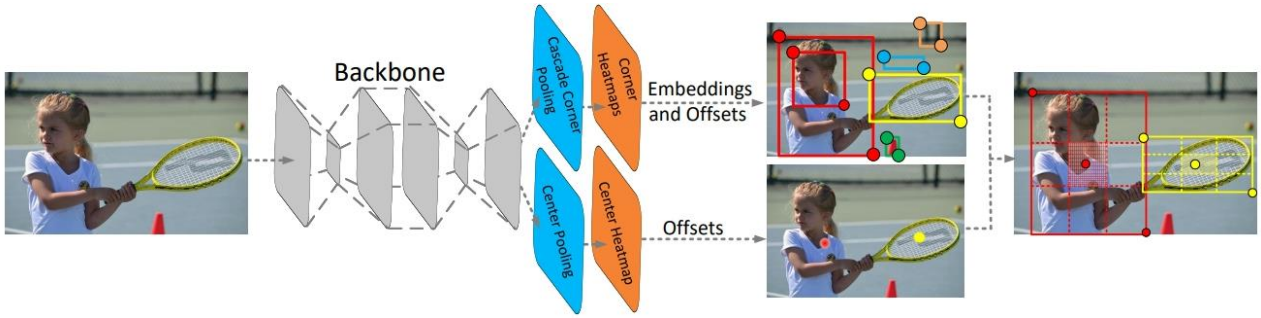

Figure 3. Center Net system architecture [5]

2.1.4. FCOS

The year 2019 witnessed the emergence of FCOS, a classic algorithm based on fully convolutional single-stage object detection. FCOS stands out as an anchor box-free and proposal-free approach that effectively avoids complex computations. Moreover, it adopts a pixel-wise prediction methodology to detect objects, akin to other dense prediction problems such as semantic segmentation. The post-processing step in FCOS is remarkably concise yet maintains high accuracy by solely employing NMS. Experimental evaluation on the MS COCO dataset demonstrates that FCOS achieves an impressive mAP of 44.7% [6]. Furthermore, in the subsequent version of FCOS (FCOSv2.0) proposed in 2020, it attains even higher accuracy with a mAP@.5 score of 50.4% on the MS COCO dataset [7].

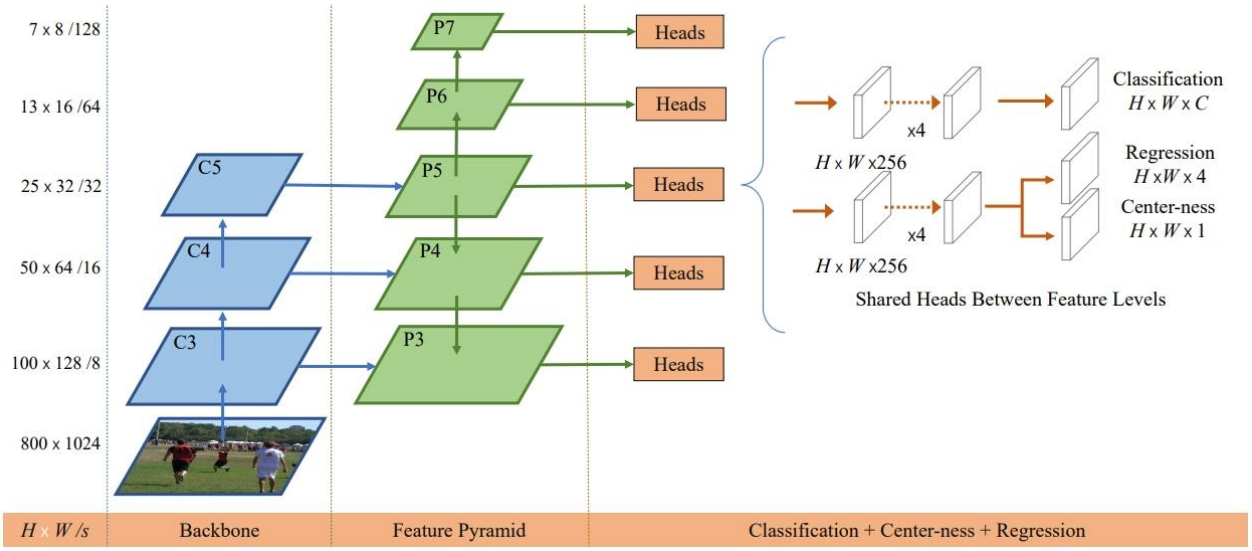

Figure 4. FCOS system architecture [8]

2.2. Anchor-based detector

In the process of object detection, Anchor-based algorithms generate prediction boxes by utilizing the center points derived from ground truth (GT) and the sizes of anchors formed through clustering. Subsequently, a neural network adjusts the position of these prediction boxes through cost calculation until accurately localizing the target object. Unlike Anchor-free methods that face limitations in

generating individual objects, Anchor-based approaches can produce dense prediction boxes. This enables Anchor-based networks to directly perform target classification and bounding box regression, significantly enhancing recall ability for small and densely packed targets. However, achieving improved detection accuracy for such targets necessitates additional effort as Anchor-based methods require meticulous parameter tuning, resulting in increased energy consumption and time.

2.2.1. SSD

The SSD, which was proposed in 2015 and introduced multi-reference and multi-resolution detection techniques, significantly enhanced the accuracy of small object detection [1]. It is noteworthy that SSD is the pioneering single-stage detector to incorporate prior boxes, akin to the anchor concept initially proposed in Faster RCNN. This innovation replaces the selective search algorithm and establishes itself as the first anchor-based detector across both single-stage and two-stage detectors. To accurately predict the offset between default boxes at various scales and aspect ratios, multiple feature layers are appended to SSD's base network (Figure 5) [8]. Ultimately, on MS COCO dataset, SSD achieves an impressive accuracy of mAP@0.5=46.5 [1].

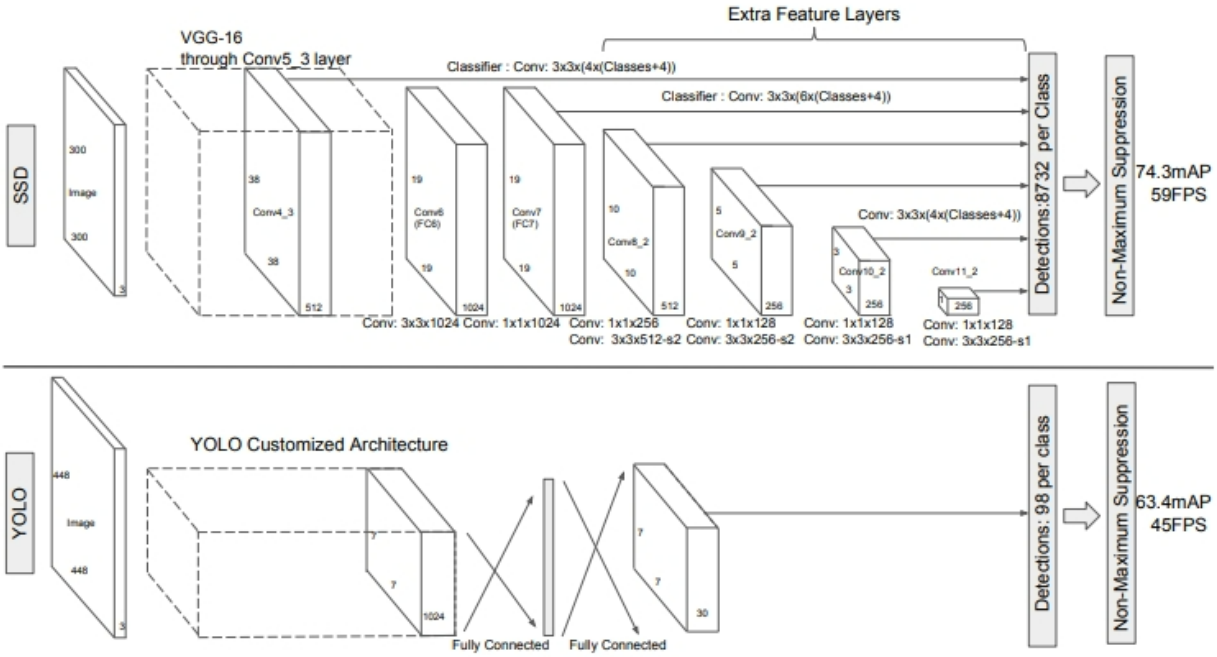

Figure 5. A comparison between SSD and YOLO, two single-shot detection models [9].

2.2.2. YOLO

The YOLOv2 introduced the application of anchor-based detectors in the YOLO series. In 2016, Joseph Redmon and Ali Farhadi proposed the YOLOv2 detector in their paper titled 'YOLO9000: Better, Faster, Stronger' [8]. YOLOv2 incorporated several enhancements over its predecessor, including improved convergence of batch normalization on all convolutional layers and reduced overfitting. It adopted a fully convolutional architecture instead of dense layers and leveraged high-resolution feature maps along with multi-scale training to capture finer-grained features. The transition to an anchor-based approach involved utilizing anchor boxes for detection. Each grid cell in YOLOv2 predicts five bounding boxes, with the system predicting coordinates and classes for each anchor box. To determine suitable priors, k-means clustering is employed on the bounding boxes. These advancements enabled YOLOv2 to achieve an average precision of 78.6% on the PASCAL VOC2007 dataset, establishing it as a fast and accurate anchor-based detector. Subsequent versions such as YOLOv3 (2018), YOLOv4 (2020), and YOLOv7 (2022) have been evaluated using the MS COCO dataset instead of PASCAL VOC2007 dataset, achieving impressive accuracies with mAP@0.5=60.6%, mAP@0.5=65. %7%, and mAP@0.5=73. %5 respectively [1,2,10].

2.2.3. RetinaNet

In 2017, Retina Net (Figure 6) was proposed to address the issue of low detection accuracy in single-stage detectors by exploring and mitigating extreme foreground-background class imbalance. The authors introduced a novel loss function called 'focal loss' by reshaping the standard cross-entropy loss, which effectively directs the detector's attention towards challenging samples and significantly improves detection accuracy. Focal loss [1,11] not only finds application in Retina Net but also demonstrates profound impact on enhancing the accuracy of anchor-free detectors. Retina Net achieved a precision of mAP@0.5=59.1% on the MS COCO dataset [1].

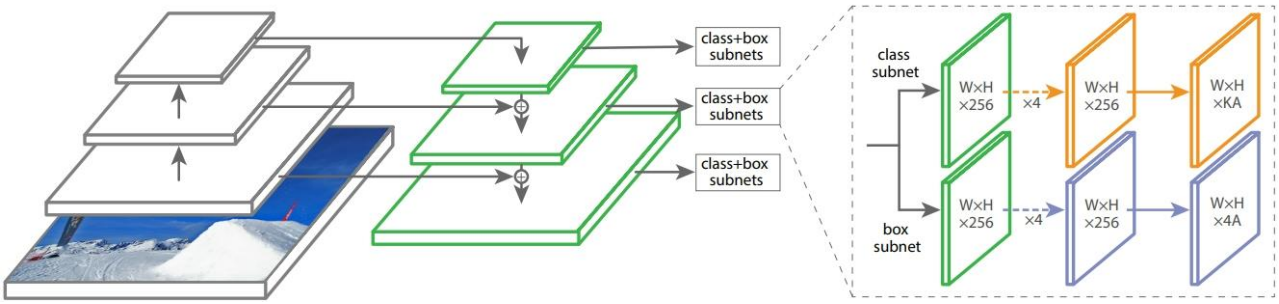

Figure 6. RetinaNet system structure[10]

3. A Comparative Analysis of Performance

3.1. Dataset

In the previous text, most of the single-stage detectors were trained, validated, and tested using the MS COCO dataset. Only YOLOv1 and YOLOv2 utilized the PASCAL VOC2007 dataset. A comparison between results obtained from these two datasets reveals that within detectors of the same era, PASCAL VOC2007 demonstrates superior accuracy compared to MS COCO.

PASCAL VOC2007 is a widely-used dataset comprising of approximately 5000 training images and nearly 5000 test images, distributed across 20 categories. However, the number of samples per category in PASCAL VOC2007 (excluding 'person') is relatively limited [11]. In contrast, MS COCO introduced its Detection dataset in 2015 and has been consistently updating it until 2020. Even the initial version from 2015 encompasses an extensive range of 80 categories, with annotations for over half a million objects and more than two hundred thousand images, each associated with five labels. The abundance of bounding boxes in MS COCO signifies the presence of numerous detectable objects within each image. Furthermore, unlike ImageNet Det dataset, MS COCO exhibits a substantial proportion of small objects which poses challenges for single-stage object detection due to its reduced accuracy in detecting densely-packed and diminutive entities [12].

The MS COCO dataset exhibits a significantly broader scope and higher level of complexity compared to the PASCAL VOC2007 dataset, rendering it exceptionally challenging for utilization. This disparity can elucidate the reason behind YOLOv3's relatively lower accuracy in contrast to YOLOv1 and YOLOv2, despite continuous optimization efforts within the YOLO series. Moreover, the formidable difficulty inherent in the MS COCO dataset also aligns it more closely with real-world application scenarios, thereby enhancing its persuasiveness.

3.2. Assessment criteria

The primary objective of this article is to employ a singular metric, mAP@0.5, as an indicator of the detector's accuracy. It signifies the average precision at an intersection over union (IoU) threshold of 50%. This metric has been extensively utilized in object detection literature [3,4,5,6,7,8,9,10,11], and nearly all single-stage detectors adopt mAP@0.5 as their performance representation. The widespread adoption of this evaluation metric facilitates comparisons among different detectors in terms of accuracy. Consequently, this article employs this metric to compare detectors across various

years and endeavors to analyze the evolutionary trajectory of single-stage object detection for deriving insights into its future trends.

3.3. Comparative analysis of data

The primary objective of this article is to employ a singular metric, mAP@0.5, as an indicator of the detector's accuracy. It signifies the average precision at an intersection over union (IoU) threshold of 50%. This metric has been extensively utilized in object detection literature [3,4,5,6,7,8,9,10,11], and nearly all single-stage detectors adopt mAP@0.5 as their performance representation. The widespread adoption of this evaluation metric facilitates comparisons among different detectors in terms of accuracy. Consequently, this article employs this metric to compare detectors across various years and endeavors to analyze the evolutionary trajectory of single-stage object detection for deriving insights into its future trends.

3.3.1. YOLO

Only for the YOLO series: YOLOv1, released in 2015, pioneered anchor-free object detection by achieving a mean average precision (mAP) of 63.4% at an intersection over union (IoU) threshold of 0.5 on the PASCAL VOC2007 dataset. In subsequent years, YOLOv2 introduced anchor boxes and transitioned to an anchor-based approach, significantly improving its performance with a mAP@0.5 of 78.6% on the same dataset in 2016. This shift from anchor-free to anchor-based detection during the early development stages of the YOLO series led to notable advancements in accuracy. Further iterations such as YOLOv3 through YOLOv4 continued utilizing anchor-based detectors and witnessed incremental improvements in mAP@0.5 from 60.6% to 65.7% on the MS COCO dataset. It is worth mentioning that although YOLOv5 also follows an anchor-based framework, it diverges from using the MS COCO dataset and will not be discussed within this context. Until 2021, YOLOX emerged as an anchor-free detector, with its anchor-free variant in the YOLO series achieving an mAP@0.5 of 50.1% on the MS COCO dataset, showcasing the potential of anchor-free development within the YOLO series. In 2022, YOLOv6 and YOLOv7 were introduced as anchor-free and anchor-based detectors respectively, both attaining a remarkable COCOmAP@0.5 of 70%. These advancements position these YOLO detectors at the forefront in terms of detection speed and accuracy. The developmental trajectory of classic YOLO detectors reveals that after their initial proposal, anchor-free methods remained dormant for nearly five years before resurfacing and now rivaling the precision achieved by anchor- based detectors; meanwhile, anchor-based methods have steadily evolved since gaining mainstream popularity.

3.3.2. All one-stage detector

For all single-stage detectors: By comparing the accuracy of each detector, it can be observed that over time, both anchor-free and anchor-based detectors have demonstrated improvements in accuracy. The development trajectory of single-stage detectors overall follows a pattern of anchor-free - anchor-based - anchor-free: In 2015, the introduction of the anchor-free detector YOLOv1 marked a significant milestone; from 2015 to 2017, the field was dominated by the emergence of anchor-based detectors such as SSD and Retina Net. However, starting from 2018, there has been an explosion in the development of anchor-free detectors including Corner Net, Center Net, Extreme Net, FCOS etc., which also witnessed a resurgence in utilizing anchor-free methods within the YOLO series. In comparison to this period of rapid progress for anchor-free methods, the advancement of anchor-based approaches has matured and slowed down. A pivotal breakthrough before the rise of anchor-free methods occurred in 2017 with FPN being proposed. FPN predicts separately for different feature layers prior to feature fusion process which enhances their ability to capture information from deep semantic features as well as shallow texture features. Particularly when only one predicted box is allowed per position in an anchor-free detector setting, FPN's structural design effectively compensates for scale variations and improves detection accuracy. Therefore, FPN plays a crucial role in facilitating transition towards more-anchor free approaches [13]. Table 1 is the comparative analysis of accuracy across various classical detectors.

Table 1. Comparative analysis of accuracy across various classical detectors

Detector	Year	Classification	Dataset	mAP
YOLOv1	2015	Anchor-free	PASCAL VOC2007	52.7%
SSD	2015	Anchor-based	MS COCO	46.5%
TOLOv2	2016	Anchor-based	PASCAL VOC2007	78.6%
RetinaNet	2017	Anchor-based	MS COCO	59.1%
CornerNet	2018	Anchor-free	MS COCO	57.8%
TOLOv3	2018	Anchor-based	MS COCO	60.6%
CenterNet	2019	Anchor-free	MS COCO	61.1%
FCOS	2019	Anchor-free	MS COCO	50.4%
YOLOv4	2020	Anchor-based	MS COCO	65.7%
YOLOv6	2022	Anchor-free	MS COCO	70.0%
YOLOv7	2022	Anchor-based	MS COCO	73.5%

4. Conclusion

The article compares the accuracy of various detectors based on the development of single-stage detectors, using Anchor-free and Anchor-based as classification standards. It also identifies a trend transitioning from Anchor-free to Anchor-based and then back to Anchor-free. Additionally, an analysis is conducted on the emergence of this regression phenomenon, revealing that FPN significantly influences the regression of Anchor-free. Despite being proposed earlier, the accuracy of Anchor-free remained low until recent years when it started regressing and approaching that of Anchor-based. Conversely, Anchor-based has reached a mature stage where further improvements in accuracy become challenging. In future developments of single-stage detectors, with the introduction of novel structures and loss functions, there will be no halt in advancing the accuracy for Anchor-free. It may even surpass that of Anchor-based. Future anchor-free detectors will explore new attempts by integrating emerging structures or loss functions, demonstrating promising prospects for development.

References

[1] J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 91, 779 - 788.

[2] *Juan Terven, Diana Cordova-Esparza. A Comprehensive Review of YOLO: From YOLOv1 and Beyond. 2023 arXiv, 2304.00501.*

[3] *Hei Law, Jia Deng. CornerNet: Detecting Objects as Paired Keypoints. 2019, arXiv, 1808.01244.*

[4] *Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, Qi Tian. (2019). CenterNet: Keypoint Triplets for Object Detection. arXiv, 1904.08189. Retireved Augst 20, 2023.*

[5] *Xingyi Zhou, Dequan Wang, Philipp Krähenbühl. Objects as Points. 2019, arXiv, 1904.07850.*

[6] *Zhi Tian, Chunhua Shen, Hao Chen, Tong He. FCOS: Fully Convolutional One-Stage Object Detection. 2019, arXiv, 1904.01355.*

[7] *Zhi Tian, Chunhua Shen, Hao Chen, Tong He. FCOS: A simple and strong anchor-free object detector. 2020 arXiv, 2006.09214.*

[8] *J Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg. SSD: Single Shot MultiBox Detector. Computer Vision 2016, 21 - 37.*

[9] *Joseph Redmon, Ali Farhadi. YOLO9000: Better, Faster, Stronger.2019 arXiv, 1612.08242.*

[10] *Tsung-Yi Lin, Priya Goyal, Ross Girshick,, Kaiming He, Piotr Dollár. Focal Loss for Dense Object Detection.2018 arXiv, 1708.02002.*

[11] *Everingham, M., Van~Gool, L., Williams, C. K. I., Winn, J., Zisserman. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision. 2010, 88, 303 - 338.*

[12] *Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár. Microsoft COCO: Common Objects in Context.2015 arXiv, 1405.0312.*

[13] *Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie. (2017). Feature Pyramid Networks for Object Detection. 2017 arXiv, 1612.03144.*