Proceedings of CONF-MLA Workshop: Mastering the Art of GANs: Unleashing Creativity with Generative Adversarial Networks
DOI: 10.54254/2755-2721/80/2024CH0060

A Review of YOLO-Based Target Detection Methods

Yunfan Lyu

School of Computer and Information Technology, Shanxi University, Shanxi, China

202101004127@email.sxu.edu.cn

Abstract. Target detection is now a popular topic in computer vision. With the development and iteration of technology, deep learning is constantly emerging. The integration of deep learning in the target detection task has led to rapid improvement in accuracy and speed, among which the You Only Look Once (YOLO) series of methods have the most rapid and varied improvements and upgrades, which have been widely used in the fields of navigation, video surveillance, face detection, text detection, and aerospace, etc. This paper initially provides an overview of the research context, importance, and challenges associated with this domain, compares and analyzes the network structure and implementation of the single-phase target detection method represented by the YOLO series with the two-phase and other improved algorithms, and then introduces the research progress of the target detection algorithms of deep learning, the characteristics of the commonly used datasets, and the key parameters of the evaluation of performance indicators, and then presents a compilation of experimental outcomes associated with several widely recognized algorithms applied to prominent datasets. Subsequently, it enumerates the experimental findings of diverse algorithms on these established datasets. Ultimately, this paper anticipates future research trajectories and developmental trends pertaining to target detection algorithms.

Keywords: Target detection, Deep learning, You Only Look Once, Computer vision.

1. Introduction

The fundamental objective of target detection is to ascertain the classification of the identified target within the picture, and at the same time, a rectangular bounding box needs to be used to found the target's basic situation, and to give the corresponding confidence level [1].

Target detection started in the 1990s, and early methods were mainly based on knowledge and feature engineering, such as rule design using attributes such as shape and colour, as well as extraction of complex image features such as Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG) combined with sliding window and Support Vector Machines (SVM) classifiers. Traditional target detection methods suffer from various shortcomings, including not robust enough in complex scenes and variable lighting; poor generalisation ability to cope with new or unseen data; low computational efficiency, which is particularly limited in real-time applications; sensitivity to scale changes, which needs to be partially addressed by image pyramid or multi-scale analysis, which increases the computational burden; complex feature engineering, which requires expertise and is time-consuming; under-utilisation of contextual information, which limits scene comprehension; lack of robustness, which handles occlusion and distortion poorly; slow detection, especially with sliding window methods; low localisation accuracy, which makes it difficult to provide an accurate bounding

DOI: 10.54254/2755-2721/80/2024CH0060

box; and difficulty in updating the model, the once deployed, adapting to new targets and scenarios requires redesigning features and tuning parameters. These limitations have promoted the convergence of object detection and deep learning, especially the introduction of Convolutional Neural Networks (CNN). CNN increased the accuracy, speed and generalization of detection markedly. Into the 21st century, component-based models such as Deformable Part Models (DPM) began to address the detection of complex targets.

The era of deep learning was opened in 2012. In order to solve the lack of accuracy of traditional methods, Girshick came up with the two-stage detection method Regions with Convolutional Neural Networks (R-CNN) [2]; followed by Fast R-CNN and Faster R-CNN in 2015 to improve the speed and accuracy respectively through ROI Pooling and RPN improved the speed and accuracy. In 2016 YOLO was born. YOLO can be considered to have pioneered the single-stage target detection method, which is a single-stage method that directly predicts the target bounding box and class through a single neural network. The proposal of YOLO has increased the possibility of real-time detection. It's end-to-end detection dramatically increases the speed of detection. The reason why YOLO is used in a variety of fields is mainly because YOLO infer fast. But its considering detection accuracy is poor. For instance, The accuracy of YOLO is 6.6 lower than that of Faster R-CNN, but the YOLO inference time is 300 times faster than Faster R-CNN [3]. Until now, the version of YOLO has been updated and improved in order to improve the performance and adapt to the conditions of certain situations. Meanwhile, the introduction of Transformer architectures, such as Detection Transformer (2020), which solves the problem of global context understanding through the self-attention mechanism. It has pushed the development of target detection technology.

In this paper, the research progress and current status of deep learning-based target detection algorithms are described, followed by a detailed description of the development and characteristics of typical single-stage target detection methods. Then are descriptions of common datasets and evaluation metrics for target detection. This is followed by a comparison of the experimental results of different algorithms on mainstream datasets, and an outlook on possible future directions in the field of target detection.

2. YOLO series

The single-stage target detection method simplifies the target detection task into a single step by predicting the categories and bounding boxes directly from the input image, and thus has low computational complexity and memory occupation, which is particularly suitable for areas with high real-time requirements such as unmanned driving and video surveillance; however, it also has limitations, such as extracting the information directly from the original image is prone to be interfered by factors such as the target size, deformation, etc., resulting in relatively low detection accuracy, especially when dealing with small-sized targets or occluded scenes. The whole training process involves a multi-task loss function, including classification loss and regression loss. The classification loss is used to ensure that the network can accurately differentiate between target categories, while the regression loss is used to accurately predict the location and bounding box of the target, so as to achieve precise target localisation. Although they may not be as good as two-stage detection methods when dealing with small and dense targets, they are known for their speed and ability to achieve real-time detection. The YOLO series is a representative of single-stage target detection methods.

2.1. Application of YOLO

YOLO series can be used in many domains. In video surveillance systems, YOLO can identify targets in real time, such as traffic light, people and vehicles; in the field of autonomous driving, which is of inestimable value, YOLO helps vehicles to understand their surroundings in real time, and detects vehicles, pedestrians, traffic signs, etc.; in industrial automation, YOLO improves the production efficiency, and assists robots to identify and locate parts; in web content auditing, YOLO quickly identifies inappropriate content and maintains the online environment; in medical image analysis, YOLO assists doctors in locating and analysing diseased tissues, and has been used in cancer detection,

DOI: 10.54254/2755-2721/80/2024CH0060

remote online surgeries, resulting in improved diagnostic accuracy and a more efficient treatment process. YOLO also plays an important role in areas such as drone surveillance, augmented reality (AR), retail analytics, sports analytics, and traffic flow analysis. Figure 1 presents a bibliometric network visualization of all publications identified in Scopus that include the term "YOLO" in their titles, with a specific focus on keywords related to object detection.

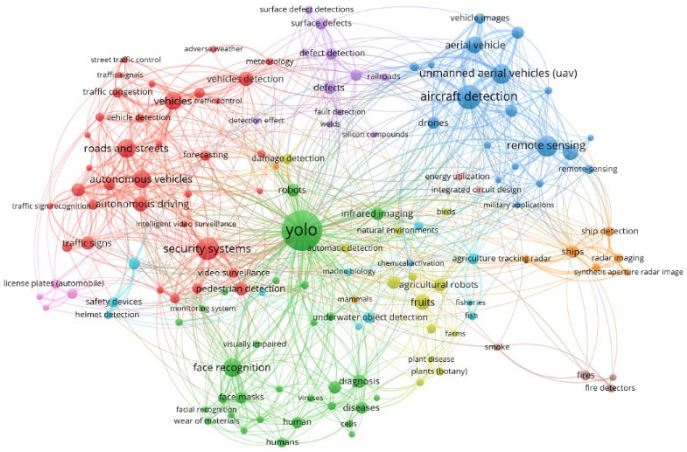

Figure 1. A bibliometric network visualization of the primary applications of YOLO has been conducted [4].

2.2. YOLOv1

Joseph Redmon et al. proposed YOLO in 2015 which is the first model that treats the target detection problem as a regression problem. YOLO directly use the input image to predict bounding boxes and category probabilities. Unsurprisingly from the name, the method only needs to be looked at once to complete the detection task. The architecture of YOLO is inspired by the GoogLeNet architecture in terms of efficient network design, multi-scale feature fusion, and modularity, using a structure similar to GoogLeNet but without the Inception module. The network structure of YOLOv1 is relatively simple, with an architecture of the DarkNet-24 network and does not employ deep networks or residual connectivity. In the prediction process, YOLOv1 divides the input image into a 7×7 grid. The strategy is that if the center of the target is on that grid, YOLO predicts the grid, with a maximum of 49 targets identified. Therefore it is not suitable for predicting dense objects. YOLOv1 has a significant speed advantage, with the base version of YOLO running at 45 frames per second without batch processing on Titan X GPUs, and the fast version running at over 150fps.This means that it can process streaming video in real time with less than 25 milliseconds of latency. However, its drawback is its poor detection of small objects. Figure 2 shows its architecture.

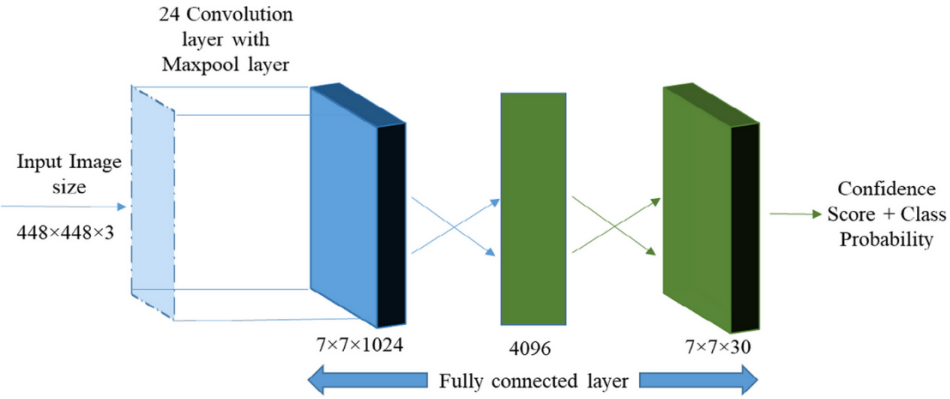

Figure 2. YOLOv1 architecture [5].

DOI: 10.54254/2755-2721/80/2024CH0060

2.3. YOLOv2

YOLOv1 used fully connected layers to directly predict the bounding box, which lost more spatial information and triggered inaccurate localisation. Joseph Redmon et al. came up with another improved version of YOLOv1, YOLOv2 in 2016.

YOLOv2 makes several improvements to the original model. YOLOv2 draws on the Visual Geometry Group network and introduces a new network structure, DarkNet-19, which contains 19 convolutional layers. Batch normalisation is incorporated to enhance both the detection speed and precision. YOLOv2 adopts the concept from Faster R-CNN by introducing anchor boxes for bounding box prediction, thus substituting the direct bounding box estimation method used in YOLOv1. But unlike Faster R-CNN, YOLOv2 computes better anchor templates in the form of K-Means clustering—a multi-scale training mechanism, and uses higher-resolution input images. YOLOv2 has an innovative improvement [6], i.e., it can be trained using multiple datasets jointly optimised for the ImageNet classification dataset and the MS YOLOv2 is faster than the previous version, while maintaining high accuracy and recognising 9000 different objects, hence the name YOLO9000.

2.4. YOLOv3

In 2018, YOLOv3 was born. YOLOv3 features the use of Darknet-53 as the base network, which borrows from short for Residual Network (ResNet)—a deeper fully convolutional network containing residual connectivity for stronger feature extraction. YOLOv3 utilises the ResNet's residual structure for preventing gradient explosion and replaces the maximum pooling operation in YOLOv1 and YOLOv2 with a convolution with a step size of 2 [7].

YOLOv3 implements multi-scale prediction, utilising different feature hierarchies for detection, enhancing the detection performance for targets of various sizes. YOLOv3 retains the speed advantage of the YOLO series while improving detection accuracy, especially for small objects.

2.5. YOLOv4

In April 2020, Bochkovskiy and colleagues built upon YOLOv3 and introduced the enhanced version, YOLOv4 [8]. The main features include:

- YOLOv4 employs CSPDarknet53 as its backbone network, which is a combination of deeply separable convolution and residual modules designed to enhance the accuracy and computational efficiency of the network.
- YOLOv4 features Mosaic data augmentation, a technique that creates new training data by randomly combining several image segments. This approach enhances the model's generalization capabilities and robustness.
- YOLOv4 uses the Complete Intersection over Union (CIoU) loss function in place of the original bounding box regression loss.
- YOLOv4 performs prediction at multiple scales, which can better handle targets of different sizes;.
- YOLOv4 utilizes the Hard Negative Mining technique, which enhances the model's detection performance by incorporating more difficult negative samples during the training process.
- YOLOv4 implements Non-Maximum Suppression to eliminate redundant detection results, while also introducing Gaussian NMS to further augment the accuracy and robustness of the process.

2.6. YOLOv5

YOLOv5 is a major release in the YOLO series, which inherits the strengths of previous releases and improves and optimises them in several aspects. The fundamental architecture of YOLOv5 exhibits similarities to that of YOLOv4. However, YOLOv5 transfers the development environment of YOLO series to Pytorch for the first time, and the biggest difference is that the YOLOv5 models, designated as N, S, M, L, and X, are developed in accordance with a scaling framework that categorizes various channels, progressing from smaller to larger model configurations., which is applicable to different computational resource constraints.

DOI: 10.54254/2755-2721/80/2024CH0060

YOLOv5 demonstrates a rapid detection capability, achieving an inference time of 0.007 seconds per image, which corresponds to a processing rate of 140 frames per second.

2.7. YOLO Series Update

YOLOX is an extended version of the YOLO series, which will be launched by Cavity Technologies in 2021. The objective of YOLOX is to enhance both the effectiveness and precision of object detection, while preserving the inherent benefits associated with the YOLO series, such as rapid processing and straightforward implementation. In terms of model architecture, YOLOX introduces a Decoupled Head design, which separates the classification and regression tasks for processing, which helps improve the accuracy of detection. Meanwhile, in contrast to conventional anchor-based detection methodologies, YOLOX employs an anchor-free framework, which simplifies the model and reduces the number of hyperparameters. YOLOX relies on YOLOv5, which inherits and improves the Mosaic data enhancement technology to bolster the model's capacity for detecting targets across multiple scales. By transitioning to the enhanced YOLOv5 architecture, which features an advanced Cross Stage Partial Network (CSPNet) backbone and supplementary PAN headers, YOLOX-L attains an Average Precision (AP) of 50.0% AP on Common Objects in Context (COCO) 640×640, which is 1.8% AP higher than the comparable YOLOv5-L [9].

YOLOv6 (2022) was proposed by the technical team of Meituan and optimized the network structure, training strategy and reasoning speed. More efficient Repetitive VGG Block and EfficientConcat were introduced, and optimization of multi-scale feature fusion was conducted.

YOLOv7 is further optimized on the basis of YOLOv6, and proposes a new Efficient Transformer structure, E-Transformer, and a new label allocation strategy. The system achieves at least 30 frames per second (FPS) and demonstrates an average precision (AP) of 56.8% when utilizing the V100 GPU [10].

3. Target detection data sets and evaluation indicators

The current popular datasets for general purpose target detection tasks are PASCAL VOC, MS COCO, ImageNet, etc. The PASCAL VOC dataset was first released in 2005, and has been updated with several versions until 2012, and it is mainly applied in the tasks of picture classification and image recognition. MS COCO is a common target database based on the daily complex scenes. It contains a total of more than 300,000 fully segmented images. The ImageNet dataset has more than 12,000,000 images and 22,000 categories, and about 1,030,000 images are labelled with target objects and categories, containing 200 object categories. The ImageNet dataset is the image recognition largest visualisation database.

The most basic definitions utilized in various annotated datasets within the object detection and scientific communities are delineated as follows:

True Positive (TP): probability of correctly identify real samples.

False Positive (FP): mistakenly identifying of a object that do not exist or the incorrect localization of an existing target.

False negative (FN): real boundary box not detected.

True Negative (TN) outcomes hold limited significance, as it is expected that an unlimited number of bounding boxes should not be tested within any single image [11].

Below are some commonly used formulas for target detection evaluation metrics:

Precision:

$$precision = \frac{TP}{TP+FP} \tag{1}$$

Precision is calculated as the proportion of correctly identified positive instances out of all instances that the model has predicted as positive. The checking rate is the ability of the model to identify only relevant objects.

Recall:

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

DOI: 10.54254/2755-2721/80/2024CH0060

Recall refers to the model's capacity to identify all pertinent instances, specifically all true positive bounding boxes. It is defined as the ratio of the total number of samples that are genuinely classified as positive to the number of those samples that the model accurately predicts as positive. This metric assesses the model's effectiveness in detecting all positively classified samples, indicating the proportion of true positive instances that are correctly recognized by the model.

F1 score:

$$2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

The F1 Score is a measure that calculates the harmonic average of Precision and Recall, functioning as a metric to assess the performance of a binary classification model, especially valuable when dealing with skewed data distributions. The F1 score serves as a very comprehensive metric

Average Precision (AP):

AP is commonly used as a metric to measure detection accuracy. The calculation of AP is complex and is obtained by calculating the precision rates at different recall rates and averaging these precision rates. Typically, the calculation of AP involves the following steps:

Calculate the IoU for each prediction frame and match it to each true frame; determine whether each prediction frame is a TP or FP based on the IoU threshold (e.g., 0.5); sort the prediction frames from highest to lowest confidence; calculate the precision rate at different recall rates; and calculate the AP using an interpolation method (e.g., 11-point interpolation or all-point interpolation).

Mean Average Precision (mAP):

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{4}$$

The mAP is usually used to measure the accuracy of the model in recognising objects from different categories. It considers two main factors: whether the position of the prediction box is accurate and whether the predicted category labels are accurate. mA is particularly important in target detection tasks because it reflects the model's overall performance on different categories, especially in the case of category imbalance. Generally speaking, the higher the mAP, the better the performance of the model. In many target detection competitions and datasets (e.g., COCO, PASCAL VOC, etc.), mAP is the main metric for evaluating and comparing the performance of different models.

4. Comparison of mAP (%) of mainstream target detection methods

The comparison of the dominant commercially available single and two-stage target detection methods on three different datasets is shown in Table 1.

Table 1. Comparison of mAP (%) of mainstream target detection methods.

Models/data sets	PASCAL VOC 2007(%)	PASCAL VOC 2012(%)	MS COCO(%)
R-CNN	58.5	6	/
Fast R-CNN	68	14	19.7
Faster R-CNN	78	18	34.9
YOLO	63.4	57.9	/
YOLOv2	78.6	73.5	21.6
YOLOv3	74.5	/	34.4
YOLOv4	87.5	/	43.5
RetinaNet	/	/	34.4
SSD	79.8	78.5	28.8

According to the content of data in the table, it can be found that the precision of the model's detection capabilities generally improves with the iteration of YOLO version, from 63.4% in the initial YOLO in PASCAL VOC 2007 to 87.5% in YOLOv4, which achieves a performance improvement of 24.1%, and from 21.6% in YOLOv2 to 43.5% in YOLOv4 on MS COCO, which reflects that the development of

DOI: 10.54254/2755-2721/80/2024CH0060

the YOLO family of models is indeed effective. The two-stage target detection demonstrates superior detection accuracy but generally the detection timeliness of the two-stage target detection algorithms is poor, and it may occasionally fail to satisfy the requirements for real-time target detection.

5. Conclusion

This article takes a detailed look at the advancement of deep learning techniques in the field of object detection, from the initial exploration to the current mature applications, showing the rapid progress of this technology. The article explains in depth the various versions of the YOLO series of algorithms, from the innovative proposal of YOLOv1 to the continuous optimization and improvement of YOLOv2, YOLOv3 and the subsequent versions, with each stage marking a major breakthrough in target detection technology. This paper summarises in detail the basic evaluation metrics of target detection, such as precision, recall, real-time and model complexity, and compares the performance of mainstream deep learning methods on different datasets (e.g. COCO, PASCAL VOC, etc.).

Deep learning techniques have been utilized in the domain of target detection across various sectors, and its technology has been evolving through people's relentless research and innovation. Both single-stage target detection methods and two-stage target detection methods are pursuing higher accuracy, faster real-time, better usability, and stronger robustness. Although the YOLO family of methods has achieved significant success in these areas, there remains potential for improvement. For instance, finer features can be extracted by exploring more advanced neural network architectures to enhance the detection capabilities for small targets and complex scenes; meanwhile, the adoption of model lightweight and hardware acceleration techniques can further improve real-time performance to fulfill the requirements of more real-time application scenarios.

These potential improvements will help the YOLO series of methods to achieve a new leap in target detection efficiency and speed, leading to more significant achievements in areas, examples include unmanned aerial vehicles, security monitoring systems, and the analysis of medical imaging. Looking ahead, with the ongoing advancement of deep learning technology, the new target detection methods will show even better performance, not only playing a greater role in existing application scenarios, but also expanding to more emerging fields, such as robot vision, augmented reality and intelligent transport systems, bringing more innovation and convenience to human society.

References

[1] Xie, F., Zhu, D. (2022). A review of deep learning target detection methods. *Computer System Appl*ications, 31(02).1-12. DOI:10.15888/j.cnki.csa.008303.

[2] Zhao, Y., Rao, Y., Dong, S., & Zhang, J. (2020). Survey on deep learning object detection. Journal of Image and Graphics, 25(04).629-654.

[3] Diwan, T., Anirudh, G., & Tembhurne, J. V. (2023). Object detection using YOLO: challenges, architectural successors, datasets and applications. Multimed Tools Appl,82(6).9243-9275. https://doi.org/10.1007/s11042-022-13644-y.

[4] Terven, J., Córdova-Esparza, D., Romero-González, J. (2023). A Comprehensive Review of YOLO Architectures in Computer Vision: from YOLOv1 to YOLOv8 and YOLO-NAS. Mach. Learn. Knowl. Extr.5(4).1680-1716. https://doi.org/10.3390/make5040083

[5] Vijayakumar, A., Vairavasundaram, S. (2024). YOLO-based Object Detection Models: a Review and its Applications. Multimed Tools Appl. https://doi.org/10.1007/s11042-024-18872-y.

[6] Luo, H., Chen, H. (2020). A review of deep learning based target detection research. Journal of Electronics, 48(6). 1230-1239. https://doi.org/10.3969/j.issn.0372-2112.2020.06.026

[7] Mi, Z., Lian, Z. (2024). A research review of YOLO methods for general-purpose target detection. Computer Engineering and Applications,1-19. http://kns.cnki.net/kcms/detail/11.2127.tp.20240705.1328.006.html.

[8] Bochkovskiy, A. et al. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv preprint arXiv: 2004.10934.

DOI: 10.54254/2755-2721/80/2024CH0060

[9] Ge, Zheng., et al. (2021). YOLOX: Exceeding YOLO Series in 2021. arXiv preprint arXiv:2107.08430.
[10] Wang, C., Bochkovskiy, A., & Liao, H. 2023. Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition(Vancouver) pp 7464-7475.
[11] Padilla, R., Netto, S. L., Da Silva, E. A. B. 2020. Int. Conf. on Systems, Signals and Image Processing(Brazil Niteroi) pp 237-242.