International Journal of Advanced Engineering and Technology Research
ISSN: 3106-8677 (Print), ISSN: 3106-8685 (Online) | Vol. 1, No. 3, 2026

Overview of the YOLO Object Detection Algorithm

Hongyu Zhang

School of Computer Science and Technology, Taiyuan Normal University, Jinzhong, 030619, Shanxi, China

Abstract: The YOLO (You Only Look Once) object detection algorithm, proposed in 2015, has now evolved to YOLOv12, boasting improved detection speed and accuracy, and is currently a hot research topic. This article focuses on introducing the basic network structure of the YOLO series algorithms, summarizing the innovations, advantages, and limitations of the YOLOv1~YOLOv12 algorithms, reviewing the applications of YOLO detection algorithms and their improved versions in industrial, agricultural, and security fields, and based on this, looking ahead to the possible future development trends of the YOLO algorithm.

Keywords: YOLO, Target detection, Convolutional Neural Networks.

1. Introduction

Object detection, a core task in computer vision, aims to accurately locate and identify specific objects from images or video sequences. Since Joseph Redmon's team first proposed the YOLO (You Only Look Once) algorithm in 2015, this framework has revolutionized the traditional detection process with its end-to-end regression paradigm. After multiple iterations, the series has now reached its twelfth generation (YOLOv12), achieving significant breakthroughs in both inference speed and recognition accuracy. Compared to two-stage methods based on region proposal (such as the R-CNN series) or single-shot detectors (SSD), the YOLO architecture exhibits a superior balance between real-time performance and accuracy. Leveraging this advantage, this technology has been widely deployed and applied in diverse scenarios such as smart agriculture, industrial automation, and public safety monitoring, becoming one of the most competitive mainstream detection solutions currently available.

2. YOLO Object Detection Algorithm

2.1. Basic Ideas of YOLO

The YOLO (You Only Look Once) architecture establishes a one-stage detection mechanism based on a regression paradigm. This algorithm abandons the traditional multi-step process, reconstructing the object recognition task into an end-to-end deep learning regression problem, utilizing convolutional neural networks to simultaneously complete bounding box localization regression and class probability prediction. As the foundational work of this series, YOLOv1 built the core framework; subsequent iterations have optimized and expanded upon this foundation, aiming to continuously overcome performance bottlenecks in feature extraction, localization accuracy, and inference speed.

The core mechanism of the YOLO algorithm lies in dividing the input image into an S×S grid structure. Each grid cell independently undertakes the detection task, specifically responsible for predicting the center point of an object falling within its region. The logic for determining the center point of an object runs through the entire training and inference process: during model training, the responsible grid is assigned based on the coordinates of manually labeled ground truth bounding boxes; while during inference and testing, the center position is calculated entirely by network regression. Once the center point of an object is determined to be within a specific grid, that cell activates its detection function, outputting the corresponding class probability and bounding box parameters.

Specifically, each grid cell needs to predict B candidate bounding boxes and their confidence scores. Each bounding box is defined by five key parameters: (x, y, w, h, confidence). Here, (x, y) represents the normalized offset of the bounding box's center point relative to the top-left corner vertex of the current grid; (w, h) represent the ratios of the bounding box's width and height relative to the overall image size. The values of these four geometric parameters are all constrained within the range [0, 1]. As for the confidence score, it quantifies the intersection-over-union (IoU) ratio between the predicted and ground truth bounding boxes. If there are no actual target objects within the grid cell, the intersection in the confidence calculation is zero, resulting in a final confidence score of zero. The final prediction result of YOLOv1 is:

$$S \times S \times (B \times 5 + C) \tag{1}$$

In the formula, S×S represents the grid size of the image, B represents the number of candidate bounding boxes generated by each grid, and C corresponds to the total number of target categories to be detected. Given that a single grid cell may output multiple predicted boxes, to address the problem of repeated detection of the same target and to select the best localization result, the algorithm introduces non-maximum suppression (NMS) as a post-processing step. The core logic of this algorithm lies in eliminating redundancy: first, the predicted box with the highest confidence score is locked as the baseline, and then its intersection-over-union (IoU) with the remaining candidate boxes is calculated. If the IoU of a box with the baseline box exceeds a preset threshold, it indicates severe overlap, and the box is judged as redundant and eliminated. This iterative selection process continues until all candidate boxes have been traversed, thus ensuring that the final output detection result is accurate and unique.

2.2. Basic YOLO Network Structure

The YOLO series of algorithms can be divided into three general modules: the backbone network, the neck feature fusion network, and the detection head. The backbone network, as the cornerstone of the YOLO architecture, is

essentially a high-efficiency feature extraction engine. Its core task is to transform the original input image into a feature representation rich in semantic information. Through cascaded convolutions, downsampling, and residual connections, it performs layered processing on the image, gradually building a feature pyramid (such as P3, P4, and P5) from low-level edge textures to high-level abstract semantics. This significantly reduces spatial resolution to expand the receptive field, while laying a solid data foundation for the subsequent multi-scale fusion of the neck network and the accurate localization and classification of the detection head. The neck network, as the key hub connecting the backbone network and the head, has the core function of performing efficient multi-scale feature fusion and enhancement. Given that deep features extracted by the backbone network possess strong semantic information but have low spatial resolution, while shallow features retain rich texture details but lack abstract semantics, the neck network constructs a feature pyramid structure (such as FPN, PANet, and their improved variants). This allows for the injection of high-level semantics into lower layers via a top-down path, while simultaneously feeding back precise localization information from lower layers to higher layers via a bottom-up path, thus achieving a complementarity between semantics and detail. Furthermore, modern YOLO versions of the neck network integrate Spatial Pyramid Pooling (SPP/SPPF), attention mechanisms, and efficient feature extraction modules such as C2f and ELAN, aiming to expand the receptive field and optimize gradient propagation. Ultimately, this generates feature representations with both high discriminative power and high accuracy, significantly improving the model's robustness and accuracy in detecting objects of different scales, especially small targets. The detection head, as the final decision and output module of the YOLO algorithm architecture, acts as a "decoder," translating the abstract high-dimensional features extracted by the neck network into specific detection results. It directly performs dense sampling on each grid cell by deploying independent prediction branches (usually containing decoupled classification and regression sub-networks) in parallel on multi-scale feature maps, predicting the geometric parameters of the bounding box (such as center coordinates, width and height), target confidence, and probability distribution of each category; then it combines post-processing algorithms such as non-maximum suppression (NMS) to remove redundant boxes, and finally transforms deep semantic information into accurate object category and location coordinates, realizing end-to-end mapping from feature representation to visualized detection results.

3. Overview of YOLO Versions

3.1. YOLOv1

In 2016, Redmon et al. first proposed YOLOv1, the initial object detection model in the YOLO (You Only Look Once) series [1]. This model transforms the traditional object detection task into a unified regression problem, takes the entire image as network input, and only needs one forward propagation to output the location information of the target bounding box and the corresponding category information at the same time. After YOLOv1 was proposed, the real-time performance of object detection was improved. However, due to its own network structure and grid division method, the model still has shortcomings in terms of localization accuracy. The accuracy of bounding box prediction is low, the number of grid divisions is small, and small-scale targets are easily ignored after entering the detection process, resulting in poor final output performance of small target detection.

3.2. YOLOv2

In response to the shortcomings of YOLOv1 in terms of localization accuracy and small target detection, Redmon et al. proposed the YOLOv2 (also known as YOLO9000) model in 2017 [2]. After YOLOv2 introduced the anchor box mechanism, the bounding box prediction method was changed from direct regression to offset prediction based on prior boxes, which improved the localization accuracy. The model uses a deeper Darknet-19 network as the backbone, and with batch normalization, multi-scale training and other operations, the model's generalization ability and detection stability are enhanced. While maintaining a fast detection speed, the detection accuracy is also improved. The model still has shortcomings. For example, a single classifier can only detect a fixed number of objects and cannot handle a variable number of detection targets. Training and adjusting the model also requires more computing resources and consumes more time.

3.3. YOLOv3

YOLOv3 [3] clearly divides the network structure of the YOLO series into three parts: backbone network, neck network, and detection head. This model uses Darknet-53 as the backbone network and FPN as the neck network. In the loss function design stage, cross-entropy loss is used to complete the target classification. The model relies on deeper feature extraction, integrates shallow detail information and deep semantic information, improves the recognition ability of multi-label targets, and completes multi-scale target prediction. Compared with the previous version, YOLOv3 has added a more complex network structure and multi-scale detection strategy for the first time. The detection accuracy has increased, but the computational cost has also increased and the inference speed has decreased. When the input size is 608×608, its real-time detection performance cannot meet the needs of high real-time applications.

3.4. YOLOv4

The YOLOv4 [4] model was proposed by Bochkovskiy et al. in 2020. This model retains the one-stage object detection framework and performs systematic optimization of the network structure and training strategy. In the backbone network, YOLOv4 adopts the CSP-Darknet-53 network structure and introduces the Cross Stage Partial (CSP) mechanism to reduce the number of parameters and computational complexity, while enhancing the feature expression capability. In terms of neck network design, YOLOv4 combines Spatial Pyramid Pooling (SPP) and Path Aggregation Network (PANet) structures to enhance the multi-scale feature fusion effect and improve the model's detection performance for small targets and complex scenes. In the training stage, YOLOv4 adds Mosaic data augmentation and CIoU loss function, combined with various training techniques, to accelerate the model convergence speed and improve detection accuracy. YOLOv4 balances detection accuracy and real-time performance, and its engineering application value is improved. However, its training process is relatively complex, and the hyperparameter setting process requires high adjustment.

Although a multi-scale detection mechanism is introduced, missed detections may still occur in dense small target scenarios.

3.5. YOLOv5

YOLOv5 [5] uses the CSP-DarkNet-53 network structure used in YOLOv4, adds a Focus module, and makes differentiated designs for the CSP structure in the backbone network and neck network to improve feature extraction and fusion capabilities. At the same time, this model changed the development framework, replacing the original Darknet with PyTorch, which optimized scalability and ease of use, and improved inference speed. YOLOv5 introduces a variety of model structures of different scales to adapt to different application scenarios and match different computing resource conditions. The shortcoming is that YOLOv5 still uses fixed-size anchor boxes, which cannot be flexibly adapted when the target scale changes significantly. In high-resolution images, the performance of small target detection will be limited.

3.6. YOLOv6

YOLOv6 [6] abandoned the anchor box mechanism and used data augmentation technology in the input stage to adjust all images to a uniform size. In the feature extraction stage, a reparameterizable and efficient EfficientRep backbone network was designed using the RepVGGstyle architecture. To improve the feature fusion effect, the Rep-PAN structure was innovatively applied in the Neck layer. In addition, the model introduced the SIoU loss function, which significantly improved the accuracy of bounding box regression through angle penalty and geometric constraint mechanisms.

3.7. YOLOv7

In the data augmentation stage, YOLOv7 [7] uses novel data augmentation methods such as Mixup, Copy-Paste, and Paste-In to enrich the content of training samples. In the design of the feature extraction network, the BConv layer is used to enhance the interaction of features between channels, and an extended efficient layer aggregation network (E-ELAN) is constructed to realize cross-level feature reuse. In order to further improve the multi-scale feature extraction capability, a multi-path convolution (MPConv) module is specially designed.

3.8. YOLOv8

YOLOv8 [8] is based on the YOLOv5 architecture and further improves upon it by introducing an enhanced feature extraction mechanism and adding an anchorless detection design. One of the core improvements of YOLOv8 is replacing the CSP module in YOLOv5 with the C2f module, which is a cross-layer feature fusion module. This structure improves the model's adaptability to multi-scale targets by deeply fusing high-level semantic features and low-level detail information. YOLOv8 uses a decoupled head, which can independently handle classification and regression tasks, improving the overall model accuracy. It also uses advanced activation functions to optimize gradient flow and feature representation. This version can be adapted to multiple tasks such as object detection, instance segmentation, and image classification. The deployment, expansion, and modification process is simple and can be directly installed via pip package. However, YOLOv8 still has some shortcomings. Maintaining high-performance detection consumes more computing resources. In scenarios with high complexity or many occlusions, the detection accuracy decreases. The training process requires a large amount of high-quality labeled data.

3.9. YOLOv9

YOLOv9 [9] innovatively proposed a programmable gradient information (PGI) architecture to address the gradient propagation distortion problem caused by the layer-by-layer decay of feature information during deep neural network training. This architecture designs multi-scale auxiliary supervision branches and embeds gradient ensemble modules within the multi-level feature pyramids of the backbone network to fuse gradient feedback information from different prediction heads.

3.10.YOLOv10

YOLOv10 [10] was proposed by Wang et al. in May 2024. It is based on YOLOv8 and has been systematically optimized around the balance between efficiency and accuracy. The core innovation of this model is to achieve complete end-to-end target detection. It designs a dual label allocation strategy and a consistent matching metric mechanism to replace the non-maximum suppression post-processing steps that the previous YOLO series relied on, which simplifies the inference process. YOLOv10 is the first real-time end-to-end detection model in the series, and the overall detection efficiency has been further improved. However, there is still room for improvement in the recognition ability in dense small target scenes.

3.11.YOLOv11

YOLOv11 [11] has made targeted optimizations at the model architecture level. The original C2f module has been replaced by the C3k2 (Cross Stage Partial with kernel size 2) module. The module uses two sets of parallel small convolutional kernels to replace a single large convolutional kernel in an equivalent form, which can compress the overall parameter scale of the model while maintaining the original receptive field size. In environments with high scene complexity and where the target is easily occluded, the model introduces the C2PSA (Convolutional Block with Parallel Spatial Attention) module with a parallel spatial attention mechanism. This module can adaptively enhance the feature representation of key regions of the target. In addition, the SPPF (Spatial Pyramid Pooling-Fast) module has also been optimized in terms of computation, which reduces the computational cost of pooling operations while ensuring the multi-scale feature extraction effect. When dealing with small target detection tasks or target detection tasks with occlusion, YOLOv11 can output stable and excellent detection results.

3.12.YOLOv12

YOLOv12 [12] is the first YOLO model with attention mechanism as its core. The model introduces an innovative region attention module and residual efficient layer aggregation network, which improves detection accuracy. However, the attention computation is heavy, which also leads to a decrease in inference speed, poor CPU performance, and stability issues during training. The official positioning is as a "community model" for research and exploration, and it is not included in the production version that can be stably deployed.

4. Applications

YOLO's series of object detection algorithms, with their superior speed and accuracy, are profoundly transforming

several key fields. In agriculture, it acts as an "electronic eye," accurately identifying crop diseases, pests, weeds, and fruit maturity, supporting smart agriculture and precision pesticide application. In industry, it enables real-time automated detection of product surface defects, significantly improving production efficiency and yield. In security, it can identify abnormal behavior and intrusions in real time, building intelligent protection networks. YOLO is becoming a core engine driving the intelligent upgrading of industries.

4.1. Agriculture

Alhwaiti et al. [13]used the YOLO deep learning model to achieve efficient identification of crop diseases, and accurately located and classified leaf lesions through target detection technology. Compared with traditional methods, YOLO has the advantages of strong real-time performance and high detection accuracy, and can run stably in complex agricultural environments. Studies have shown that this method significantly improves the efficiency of disease identification, helps to realize intelligent and precise management of agricultural production, and is of great significance for increasing crop yield and reducing economic losses. Yu et al. [14] introduced the feature extraction capabilities of structure-optimized networks such as GSC2f on the basis of YOLOv8. This method can effectively deal with problems such as complex backgrounds, large changes in light, and significant differences in lesion scale in agricultural scenarios. Experimental results show that BGM-YOLO achieves a good balance between accuracy and detection speed, and has higher recognition accuracy and stronger robustness than the original model, which can provide reliable technical support for intelligent monitoring of agricultural diseases and precision agriculture.

4.2. Industry

Wang J et al. [15] proposed a lightweight YOLO target detection model for surface defect detection in electronic manufacturing. By optimizing the network structure, the computational complexity is reduced while ensuring detection accuracy, enabling real-time detection. Experimental results show that this method has high recognition rate and stability for various defect types, effectively improving the efficiency of industrial quality inspection and reducing labor costs. Lu M et al [16] proposed an improved model for steel surface defect detection, introducing dynamic convolution and optimized loss function on the basis of YOLOv8 to enhance feature extraction capabilities. This method can effectively cope with multi-scale and weak defect problems in complex industrial environments. Experimental results show that WSS-YOLO outperforms the original model in both detection accuracy and robustness, and is suitable for real-time industrial quality inspection scenarios.

4.3. Security

Gawande U et al [17] proposed an improved YOLOv5 model combined with motion feature maps for personnel detection and suspicious behavior recognition. By fusing spatiotemporal information, the detection accuracy and stability in complex monitoring scenarios were improved. Experimental results show that this method achieves a good balance between real-time performance and recognition performance, and can be effectively applied to intelligent security monitoring systems. Wang et al [18] proposed an improved YOLOv4 model for real-time weapon detection in videos. By optimizing the network structure and feature extraction methods, the ability to recognize weapons in small targets and complex backgrounds was improved. Experimental results show that this method performs well in terms of accuracy and real-time performance, and can be effectively applied to counter-terrorism and public safety monitoring scenarios.

5. Conclusion

The YOLO object detection algorithm has evolved through multiple generations, achieving a good balance between detection accuracy and real-time performance, and has become an important research direction in the field of computer vision. This paper systematically reviews the development history and key technological improvements of YOLOv1 to YOLOv12, and analyzes its practical application value in typical application scenarios such as agriculture, industry, and security. Although YOLO still faces certain challenges in small object detection, adaptability to complex scenes, and computational resource consumption, its performance is expected to improve further with the continuous development of attention mechanisms, anchor-free design, and end-to-end optimization. In the future, the YOLO algorithm will continue to evolve in terms of intelligence, lightweight design, and multi-task integration, providing stronger technical support for the intelligent upgrading of various industries.

References

[1] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[2] Redmon J, Farhadi A. YOLO9000: better, faster, stronger [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[3] Farhadi A, Redmon J. Yolov3: An incremental improvement [C]//Computer vision and pattern recognition. Berlin/Heidelberg, Germany: Springer, 2018, 1804: 1-6.

[4] Wen C, Wen J, Li J, et al. Lightweight silkworm recognition based on Multi-scale feature fusion [J]. Computers and electronics in agriculture, 2022, 200: 107234.

[5] Wang H, Jin Y, Ke H, et al. DDH-YOLOv5: improved YOLOv5 based on Double IoU-aware Decoupled Head for object detection [J]. Journal of Real-Time Image Processing, 2022, 19(6): 1023-1033.

[6] Norkobil Saydirasulovich S, Abdusalomov A, Jamil M K, et al. A YOLOv6-based improved fire detection approach for smart city environments [J]. Sensors, 2023, 23(6): 3161.

[7] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7464-7475.

[8] Hussain M. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection [J]. Machines, 2023, 11(7): 677.

[9] Wang C Y, Yeh I H, Mark Liao H Y. Yolov9: Learning what you want to learn using programmable gradient information [C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2024: 1-21.

[10] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection [J]. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011.

[11] Cheng S, Han Y, Wang Z, et al. An underwater object recognition system based on improved yolov11 [J]. Electronics, 2025, 14(1): 201.

[12] Tian Y, Ye Q, Doermann D. Yolov12: Attention-centric real-time object detectors [J]. arXiv preprint arXiv: 2502.12524, 2025.

[13] Alhwaiti Y, Khan M, Asim M, et al. Leveraging YOLO deep learning models to enhance plant disease identification [J]. Scientific Reports, 2025, 15(1): 7969.

[14] Yu C, Xie J, Tony F J A. BGM-YOLO: An accurate and efficient detector for detecting plant disease [J]. Plos one, 2025, 20(5): e0322750.

[15] Wang J, Dai H, Chen T, et al. Toward surface defect detection in electronics manufacturing by an accurate and lightweight YOLO-style object detector [J]. Scientific Reports, 2023, 13(1): 7062.

[16] Lu M, Sheng W, Zou Y, et al. WSS-YOLO: An improved industrial defect detection network for steel surface defects [J]. Measurement, 2024, 236: 115060.

[17] Gawande U, Hajari K, Golhar Y. Novel person detection and suspicious activity recognition using enhanced YOLOv5 and motion feature map [J]. Artificial Intelligence Review, 2024, 57(2): 16.

[18] Wang G, Ding H, Duan M, et al. Fighting against terrorism: A real-time CCTV autonomous weapons detection based on improved YOLO v4 [J]. Digital Signal Processing, 2023, 132: 103790.