# Real time object recognition based on YOLO model

**Zeyu Guan**

School of Automation, Nanjing University of Aeronautics and Astronaut, Nanjing, China

zyguan@nuaa.edu.cn

**Abstract.** With the rapid development of computer technology, the concept of computer vision has been proposed. Since then, many object recognition methods have been developed to lay the foundation for computer vision. Object recognition is vital in various computer vision applications, such as autonomous driving, surveillance systems, robotics, and other areas. The You Only Look Once (YOLO) model has gained significant attention due to its ability to achieve real-time object detection and localization in images and videos. This paper comprehensively reviews real-time object recognition based on the YOLO model. We discuss the YOLO architecture's underlying principles and advantages over traditional object detection methods. Then, according to the article by Joseph Redmon, the inventor of YOLO, the benefits of each version of the YOLO model and the performance optimization compared to the previous work are briefly introduced in the order of release. Furthermore, this paper explores its applications in different domains.

**Keywords:** Component, Object Recognition, You Only Look Once Model, Computer Vision.

## 1. Introduction

1943, American psychologist W. McCulloch and mathematician W. Pitts first proposed the concept of artificial neural networks. Using mathematical models, they completed the theoretical modeling of neurons in artificial neural networks and then opened people's research on artificial neural networks.

Deep learning-based object detection algorithms are mainly divided into two categories. One is based on Region-Convolution Neural Network (RCNN) [1], Fast RCNN [2], two-stage detection algorithms represented by Faster RCNN [3], Mask RCNN [4], and so on. Based on feature extraction, this algorithm is formed into many candidate regions by independent network branches and then classified and regressed, which has excellent detection accuracy and recall. Still, due to its complex network structure, the detection speed is slightly lower. The other is based on YOLO [5], and the Single Shot MultiBox Detector (SSD) [6] represents the first-stage algorithm, different from the two-stage algorithm first positioning. The classification detection method, one core of the staging algorithm, is to directly extract features in the network to predict object classification and location, and its algorithm model is simple, which can significantly improve the detection speed when the detection accuracy decreases slightly.

Among them, the YOLO series one-stage detection algorithm realizes end-to-end detection, ensuring high image feature extraction accuracy while considering real-time performance. YOLOv1 creatively treats the object detection task as a regression problem with multiple bounding boxes in space and the probability of classes corresponding to the bounding boxes, which is very fast [7]. Still, each unit can only detect one type of objects. Subsequently, YOLOv2 absorbs the advantages of the SSD algorithm [5]. It introduces various techniques such as batch normalization, cluster generation prior boxes, and constrained prior box scale so that predictions are more accurate, detection speed is faster, and more objects are recognized. YOLOv3 uses the Darknet53 network as the backbone to complete the feature extraction of images and use multi-scale features to object detection balances speed and accuracy, accelerating the landing of target detection in the industry [8].

The first part of this article introduces the basic principles of the YOLO model and how it compares to the original object recognition technology. The second part introduces the advantages of each version of YOLO and ideas for improvement. The third part presents the practical application of the YOLO model in engineering.

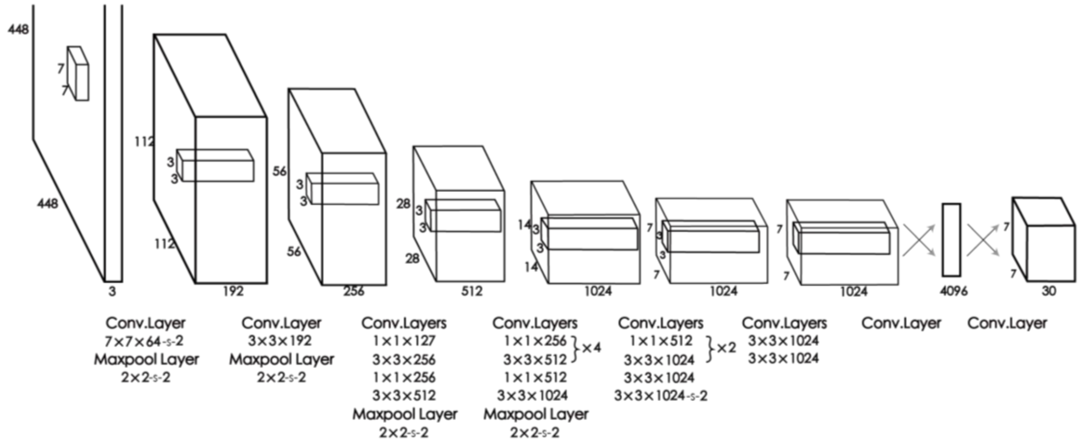

**Figure 1.** YOLO neural network structure [7].

## 2. The Principle of the YOLO Model and its Advantages

### *2.1. The principle of YOLO:*

The principle of YOLO, which stands for "You Only Look Once," is a real-time object detection algorithm commonly used in computer vision tasks. YOLO revolutionized object detection by introducing a single-stage, end-to-end deep learning approach that achieves high accuracy and fast processing speeds.

Here are the fundamental principles of YOLO:

(1) Single-stage detection: You can see the main structure of the YOLO neural network in Figure 1 [7]. YOLO takes an input image and divides it into a grid. Each grid cell predicts a fixed number of bounding boxes along with their corresponding class probabilities and confidence scores. Unlike traditional two-stage detectors, the entire detection process is performed in a single pass through the neural network.

(2) Bounding box prediction: YOLO predicts bounding boxes using regression. YOLO indicates the bounding box coordinates for each grid cell relative to the cell location and size. These coordinates are usually defined as offsets from the top-left corner of the grid cell.

(3) Class prediction: In addition to bounding boxes, YOLO also predicts the probability of each class for each bounding box. It uses softmax activation to calculate the class probabilities.

(4) Confidence score: YOLO introduces a confidence score to estimate the accuracy of the predicted bounding boxes. The confidence score represents the likelihood of an object being present within a bounding box. It is calculated as the product of the class probability and the intersection over union (IoU) between the predicted and ground truth boxes.

(5) Non-maximum suppression (NMS): YOLO applies non-maximum suppression to eliminate duplicate detections and select the most accurate bounding box. It removes bounding boxes with low

confidence scores and high overlap with other packages, keeping only the most confident and non-overlapping boxes.

(6) Training process: YOLO is trained on a large labeled dataset where each object is labeled with a bounding box and class label. The training process involves optimizing the neural network parameters using a loss function that combines localization loss (related to the accuracy of the predicted bounding boxes) and classification loss (related to the accuracy of the predicted class probabilities).

(7) Anchor boxes: YOLO utilizes anchor boxes to improve the detection of objects with different aspect ratios and scales. These anchor boxes are pre-defined bounding boxes of various shapes and sizes placed at each grid cell. YOLO predicts offsets to these anchor boxes to better align the predicted bounding boxes with the ground truth.

(8) Feature extraction: YOLO employs a deep convolutional neural network (CNN) as its backbone for feature extraction. The network processes the input image through multiple convolutional and pooling layers to capture rich hierarchical features at different scales. These features are then used to predict object classes and bounding boxes.

### *2.2. Advantages of YOLO Model*

(1) Real-Time Object Detection: One of the primary advantages of the YOLO model is its ability to achieve real-time object detection. Unlike traditional object detection methods that require multiple stages and complex pipelines, YOLO performs object recognition in a single pass through the network. This single-shot approach eliminates the need for time-consuming region proposal methods, resulting in significantly faster processing times. The YOLO model can process images or video frames in near real-time, making it highly suitable for immediate object detection and tracking applications.

(2) High Accuracy and Precision: Despite its real-time capabilities, the YOLO model maintains high accuracy and precision in object detection. By leveraging a unified architecture that simultaneously predicts bounding boxes and class probabilities, YOLO achieves accurate object localization and classification. The model benefits from the holistic understanding of the image context its grid-based design provides. This enables YOLO to capture fine-grained details and make precise predictions, leading to accurate and reliable object recognition results.

(3) Efficient Utilization of Hardware Resources: The YOLO model is designed to efficiently use hardware resources efficiently, enabling real-time object recognition on various devices. YOLO models are typically lightweight, with a small number of parameters compared to other complex architectures. This efficient design allows the YOLO model to run smoothly on resource-constrained devices such as embedded systems, drones, or smartphones. The ability to perform real-time object recognition with minimal computational requirements makes YOLO a practical and versatile solution for deployment in various environments.

(4) Simplicity and Ease of Implementation: The YOLO model offers a straightforward and intuitive architecture, making it relatively easy to understand and implement. With its single-shot approach, YOLO simplifies object detection by combining localization and classification into a single step. This simplicity facilitates rapid development and prototyping and allows researchers and developers to focus on enhancing specific aspects of the YOLO model or tailoring it for domain-specific applications. The simplicity of the YOLO model has contributed to its popularity and widespread adoption in the computer vision community.

Overall, the YOLO model offers several advantages, including real-time object detection, high accuracy and precision, efficient utilization of hardware resources, and simplicity in implementation. These advantages have positioned YOLO as a go-to choice for real-time object recognition tasks in various applications, enabling rapid and accurate detection of objects in images and videos.

## 3. Specific Research Work of Theyolo Model

### *3.1. The origin of the YOLO model——YOLOv1*

The YOLO model's ability to deliver accurate object detection results in real time has made it highly influential in the computer vision community. Its simplicity, speed, and accuracy have led to widespread adoption in various domains, driving advancements in real-time object recognition research.

YOLOv1 is a typical object detection one-stage method, which Joseph Redmon proposed in 2016 [7]. It is described in detail in the article: You Only Look Once: Unified, Real-Time Object Detection. The core idea of the YOLO algorithm is to treat the object detection problem as a regression problem with a convolutional neural network structure that can directly predict the bounding box and class probability from the input image. Using the regression method to do object detection, the execution speed is fast to achieve very efficient detection, the principle and idea behind it are also straightforward.

This paper proposes the concept of the Yolo model for the first time. The YOLO detection network includes 24 convolutional layers 2 fully connected layers to extract image features, and fully connected layers to predict image location and class probability values. The YOLO network borrows from the GoogLeNet classified network structure. The difference is that YOLO does not use the inception module but uses a 1×1 convolutional layer (here, the 1×1 convolutional layer exists for cross-channel information integration) + a 3×3 convolutional layer for a simple alternative. In addition, the authors also present a lighter detection network, fast YOLO, which has only 9 convolutional layers and 2 fully connected layers. With the titan x GPU, fast YOLO can achieve detection speeds of 155 fps, but the mAP value has also dropped from 63.4% to 52.7% of YOLO, but it is still much higher than the mAP value of previous real-time object detection methods (DPM).

### *3.2. Better, faster, stronger——YOLOv2*

Joseph Redmon updated Yolov1 in 2017, and Yolov2 came out [5]. The article introduces the most advanced real-time object detection system YOLO9000, which can detect more than 9000 object categories, using a novel, multi-scale training method. The same YOLOv2 model can run at varying sizes, offering an easy tradeoff between speed and accuracy.

YOLO suffers from a variety of shortcomings relative to state-of-the-art detection systems. Error analysis of YOLO compared to Fast R-CNN shows that YOLO makes many localization errors. Furthermore, YOLO has relatively low recall compared to region proposal-based methods. YOLO9000 is focused on improving memory and localization while maintaining classification accuracy.

Finally, the article proposes a joint training object detection and classification method. Using this approach, they train YOLO9000 simultaneously on the COCO detection and ImageNet classification datasets. Their joint training enables the YOLO9000 to predict the detection of object classes without labeled detection data.

### *3.3. An incremental improvement——YOLOv3*

In 2018, Joseph Redmom again made some optimizations on the YOLO model. Compared with YOLO v2, YOLO v3 has few innovative points, but it borrows some reasonable solutions to integrate into YOLO v2 [8]. Under the premise of maintaining the speed advantage, the prediction accuracy is improved, especially the recognition ability of small objects. This article has made a series of updates to YOLO, so that the system's performance has been improved simultaneously. The authors also trained a more accurate and better neural network. At 320×320, YOLOv3 runs in 22 ms at 28.2 mAP, as precise as SSD but three times faster. This article has made a series of updates to YOLO, so that the system's performance has been improved simultaneously. The authors also trained a more accurate and better neural network. The working principle of the four parts of Bounding Box Prediction, Class Prediction, Predictions Across Scales, and Feature Exquisite was introduced in detail. Then, the recognition simulation of small things was carried out, and the results were that YOLOv3 can perform well in the face of small objects that cannot be identified well in the previous YOLO version. However, when faced with medium-to-large body size targets, the performance is more general, and follow-up research is needed. Finally, the author analyzes the problems found in the research process and his solutions, providing error cases for later researchers to prevent later researchers from repeating the same mistakes.

## 4. Applications of YOLO Model

### *4.1. Autonomous Driving*

With the development of computer technology, the degree of equipment automation has increased. The concept of artificial intelligence is increasingly entering people's lives, and the Yolo model based on machine deep learning has also been more widely used.

One of the most prominent applications of the YOLO model is in autonomous driving systems [9]. Real-time object recognition is crucial for the perception module of self-driving vehicles to detect and track various objects, such as pedestrians, cars, traffic signs, and road markings. By leveraging the speed and accuracy of the YOLO model, autonomous vehicles can make informed decisions in real time, enhancing their ability to navigate complex road environments safely.

### *4.2. Surveillance Systems*

YOLO has also found extensive use in surveillance systems, where real-time object recognition plays a vital role in detecting and monitoring objects of interest [10]. The YOLO model enables efficient and accurate object detection in live video feeds, whether identifying intruders, tracking suspicious activities, or monitoring crowd movement. Its ability to process frames in real-time makes it ideal for applications requiring immediate responses, such as security and surveillance in public spaces, airports, or critical infrastructure.

### *4.3. Robotics and Industrial Automation*

The YOLO model has proven valuable for object recognition tasks in robotics and industrial automation. Robots equipped with YOLO-based object recognition systems can identify and localize objects in real time, enabling them to interact with their environment effectively. This capability is handy in tasks like pick-and-place operations, sorting objects on conveyor belts, or assisting in warehouse operations. YOLO-based object recognition empowers robots to adapt to changing environments and perform tasks precisely and efficiently.

### *4.4. Abbreviations and Acronyms Gaming*

The YOLO model also has augmented reality (AR) and gaming applications. By leveraging real-time object recognition, AR applications can seamlessly integrate virtual objects into the real world and provide interactive experiences. YOLO enables accurate detection and tracking of real-world objects, allowing virtual objects to interact convincingly with the physical environment. In gaming, YOLO-based object recognition can be used for gesture recognition, player tracking, and real-time interaction with virtual characters or objects.

### *4.5. Healthcare and Biomedical Imaging*

The YOLO model has demonstrated potential in various applications in the healthcare domain, such as medical imaging analysis and assisted diagnosis. Real-time object recognition can assist in detecting abnormalities, tumors, or specific anatomical structures in medical images or video streams. YOLO-based systems can aid radiologists and medical professionals by providing rapid and accurate insights facilitating early detection and intervention. Furthermore, the YOLO model can track surgical instruments during minimally invasive procedures, enhancing surgical precision and safety.

These applications represent just a few examples of how the YOLO model has been successfully applied in different domains. Its real-time capabilities, combined with high accuracy, have made it a go-to choice for object recognition tasks, enabling a wide range of applications that benefit from fast and reliable detection and classification of objects in images and videos.

## 5. Conclusion

The YOLO model has emerged as a powerful solution for real-time object recognition. YOLO introduced a groundbreaking approach by formulating object detection as a single-shot regression problem, enabling simultaneous object localization and classification. This architecture eliminates the need for computationally expensive region proposal methods in traditional object detection algorithms, making YOLO significantly faster.

The YOLO model divides the input image into a grid, directly predicting bounding boxes and class probabilities from the grid cells. This unique design allows YOLO to achieve real-time performance by performing object detection in a single pass through the network. Over the years, YOLO has undergone several iterations and improvements, leading to versions such as YOLOv1, YOLOv2, and YOLOv3. Since then, Yolo technology has continued to evolve rapidly. In the following years, new versions were continuously released, and the idea of object recognition was constantly optimized, which has developed to yolov8.

While highly influential in real-time object detection, the YOLO model faces limitations in fine-grained object detection, occlusion handling, and aspect ratio variations. However, future trends will likely address these shortcomings through multi-scale architectures, attention mechanisms, and contextual understanding. Advancements in one-stage detectors, few-shot learning, and cross-modal extensions are expected to make YOLO more versatile and adaptable. Additionally, a focus on explainability and fairness will ensure transparency and mitigate biases, expanding its applications across diverse industries in the future.

In summary, the YOLO model's ability to deliver accurate object detection results in real time has made it highly influential in the computer vision community. Its simplicity, speed, and accuracy have led to widespread adoption in various domains, driving advancements in real-time object recognition research.

## References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”.

[2] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” arXiv, Jan. 06, 2016. Accessed: Jul. 03, 2023. [Online]. Available: http://arxiv.org/abs/1506.01497

[3] L. Du, R. Zhang, and X. Wang, “Overview of two-stage object detection algorithms,” J. Phys.: Conf. Ser., vol. 1544, no. 1, p. 012033, May 2020, doi: 10.1088/1742-6596/1544/1/012033.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN.” arXiv, Jan. 24, 2018. Accessed: Jul. 03, 2023. [Online]. Available: http://arxiv.org/abs/1703.06870

[5] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger.” arXiv, Dec. 25, 2016. Accessed: Jul. 03, 2023. [Online]. Available: http://arxiv.org/abs/1612.08242

[6] W. Liu et al., “SSD: Single Shot MultiBox Detector,” in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., in Lecture Notes in Computer Science, vol. 9905. Cham: Springer International Publishing, 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection.” arXiv, May 09, 2016. Accessed: Jul. 03, 2023. [Online]. Available: http://arxiv.org/abs/1506.02640

[8] K. J. Oguine, O. C. Oguine, and H. I. Bisallah, “YOLO v3: Visual and Real-Time Object Detection Model for Smart Surveillance Systems(3s).” arXiv, Sep. 26, 2022. Accessed: Jul. 03, 2023. [Online]. Available: http://arxiv.org/abs/2209.12447

[9] C.-J. Lin, S.-Y. Jeng, and H.-W. Lioa, “A Real-Time Vehicle Counting, Speed Estimation, and Classification System Based on Virtual Detection Zone and YOLO,” Mathematical Problems in Engineering, vol. 2021, pp. 1–10, Nov. 2021, doi: 10.1155/2021/1577614.

[10] K. J. Oguine, O. C. Oguine, and H. I. Bisallah, “YOLO v3: Visual and Real-Time Object Detection Model for Smart Surveillance Systems(3s).” arXiv, Sep. 26, 2022. Accessed: Jul. 03, 2023. [Online]. Available: http://arxiv.org/abs/2209.12447