Analysis the improvements of YOLOv5 algorithms: NRT-YOLO, MR-YOLO and YPH-YOLOv5

Wenlei Tao

Sussex Artificial Intelligence, Zhejiang Gongshang University, Hangzhou 310000, China

wt218@sussex.ac.uk

Abstract. Computer Vision (CV) is a fundamental aspect of artificial intelligence, with applications spanning multiple domains. The YOLO (You Only Look Once) algorithm has significantly contributed to real-time object recognition in CV. This paper explores the evolution of the YOLO algorithm, focusing on the improvements brought by three specialized variants: NRT-YOLO, MR-YOLO, and TPH-YOLOv5. NRT-YOLO addresses the challenge by introducing the C3NRT module, enhancing precision while maintaining low complexity. MR-YOLO optimizes YOLOv5 for industrial quality control, improving speed and accuracy. TPH-YOLOv5 enhances object detection in drone-captured images by introducing additional prediction heads and transformer modules. Despite these advancements, YOLO algorithms have limitations, including difficulty in detecting small objects and issues in complex scenes. Nevertheless, the research sheds light on the continuous evolution of YOLO algorithms, offering insights into real-time object detection, with applications in manufacturing, healthcare, transportation, and more. The significance of this research lies in showcasing the adaptability and potential of the YOLO framework. As YOLO continues to evolve, it promises to revolutionize various industries and applications, providing robust, real-time object detection solutions.

Keywords: CNN, YOLOv5, C3NRT, MobileNetV3, transformer prediction head.

1. Introduction

Computer Vision is one of the core technologies of AI, involving the acquisition, processing and analysis of images and videos in the real world, so that machines can extract meaningful contextual information from the physical world, and realize the vision of machines like humans. It covers a wide range of crucial CV technology fields, such as edge detection, motion detection, face identification, image recognition, pattern recognition, and optical character recognition. The 1950s saw the start of research on computer vision technology, which then had a period of fast advancement in the 1970s. The publication of David C. Murray (Vision) in 1982 marked the emergence of computer vision as an independent discipline. The 1990s saw the successful use of computer vision to industries that had very low standards for accuracy and robustness, such as video conferencing, archaeology, surveillance, and security. Through the cross-integration of AI, the field of computer vision has realized more sophisticated applications of scenarios since the turn of the twenty-first century. Today, computer vision technology is widely employed in a variety of industries, including manufacturing, healthcare, and transportation.

The science of computer vision has greatly advanced since convolutional neural networks (CNN) were first used for object detection [1]. There have been many great CNN-based algorithms put out, including Faster R-CNN, YOLO [2], and SSD. These techniques have produced outstanding results for natural object detection in labelled datasets like MS COCO and PASCAL VOC [3]. Additionally, numerous detectors utilising transformers are presented, including deformable DETR and YOLOS, because of the inclusion of transformer components into the field of view. These techniques can offer greater accuracy than CNN-based networks in specific detecting applications. The YOLO series algorithm, a traditional one-stage detection technique, stands out among these as having clear advantages in processing real-time data and trials with high engineering numbers, which has led to its increased use in computer vision research in recent years.

This paper first discusses the basic concept for YOLO [4] and the improvement features of YOLOv5. Next, this study will take three improved models of YOLOv5, RNT-YOLO [5], MR-YOLO [6], and YPH-YOLOv5 [7], as research objects to explore the rich possibilities existing in this algorithm. Finally, this study will summarize some limitations of YOLO model in use, and look forward to the future development direction of YOLO framework, which will affect the continuous development of real-time object detection system.

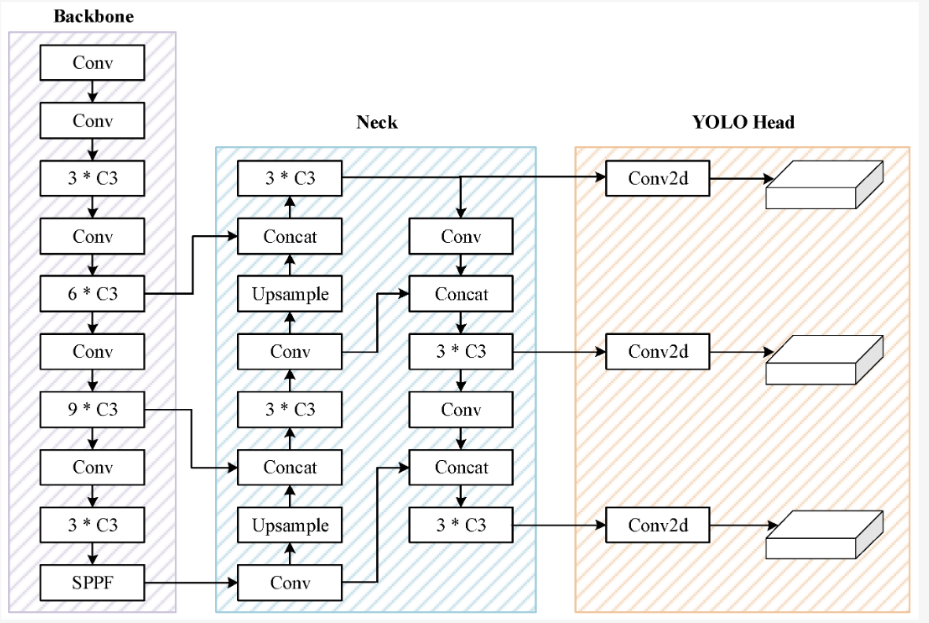

Figure 1. The architecture of YOLOv5 [9].

2. Basic descriptions

Joseph, for the first time, real-time end-to-end object detection is utilized [8]. The expression "You only see once" (YOLO) refers to the fact that the network can finish the detection task list by employing a synchronous method that involves a sliding window followed by a classifier. Running hundreds or thousands of photos is necessary for this, or more sophisticated techniques. A classifier is run in the second phase after the first step has detected regions of potential items or regions. The YOLO predictive detection output also employs two distinct outputs as opposed to the Fast R-CNN, which only employs a straightforward regression-based output. Regression frame coordinates and classification probability. YOLOv5 is the YOLO series released by Glenn Jocher in 2020, a few months after YOLOv4 [9]. It was developed with Pytorch instead of Darknet, based on YOLOv4. The backbone is an improved CSP Darknet53, consisting of improved CSP-PAN and SPPF. The neck uses improved CSP-PAN and SPPF, while the head is similar to YOLOv3. YOLOv5 is open source and actively maintained by Ultralytics, with over 250 contributors. Ultralytics offers many tagging, training and deployment integrations, as well as iOS and Android mobile versions. It also uses a number of enhancements including Mosaic, copy-paste, random affine, blending, HSV enhancement, and random horizontal flipping. The sensitivity of the grid is improved, making it more stable against runaway gradients. The architecture of YOLOv5 is illustrated in Figure 1.

3. NRT-YOLO

With regard to object detection technology, faced with the characteristics of clusters of tiny objects, unsatisfactory results often appear. Because the location of the imaging sensor is generally far away, the object of focus observation usually looks small and densely distributed. In response to this demand, researchers began to try to use single-stage detector YOLO. At this time, YOLOv5 showed great potential in detecting small objects [10]. Inspired by this, some researchers used YOLOv5l as a benchmark network, replacing C3 blocks in the original backbone network. The architecture of NRT-YOLO is illustrated in Figure 2.

The C3NRT module is one of the key innovations of RNT-YOLO, which employs an ingenious structure consisting of multiple residual structures nested. This includes a combination of transformer encoders and bottleneck modules, which work together to effectively improve performance. This module is designed to overcome the challenges, including the need for higher detection accuracy. The goal of the entire RNT-YOLO approach is to significantly improve the performance of YOLOv5 in the detection of small objects, making it more reliable and powerful in practical applications. The design of this algorithm not only fully considers the special properties of small objects, but also aims to improve the acquisition of global information to detect and locate small objects more accurately. NRT-YOLO is suitable for small remote sensing objects because of its high precision and low complexity. In future studies, more and different data sets will be used to demonstrate the generalization ability of the NRT-YOLO and C3NRT modules.

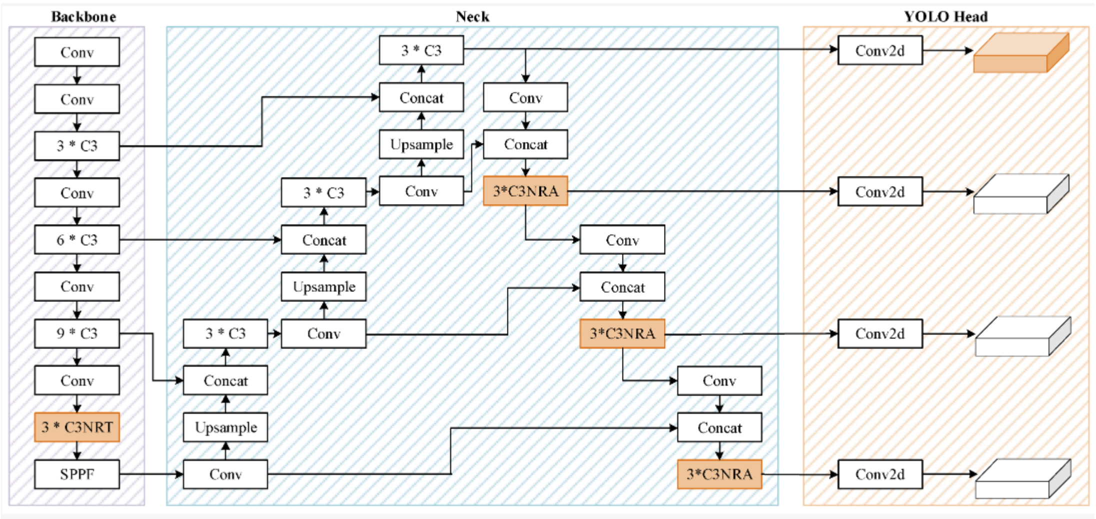

Figure 2. The architecture of NRT-YOLO [10].

4. MR-YOLO

The quality detection of equipment is crucial in the manufacturing of various high-precision industrial gadgets. In response to the problems, research has switched to a new technology to detect product quality. More and more industrial researchers are using YOLOv5 because to its high detection accuracy and lightning-fast speed. Even the highly complicated and minimally computed YOLOv5s model struggles with insufficient real-time performance. The quick and accurate magnetic ring detection network model MR-YOLO (YOLOv5 for Magnetic Ring) based on YOLOv5s was developed. To perform object detection, the YOLOv5 algorithm specifically makes use of a different CNN model. The network outputs three separate scales of prediction results, with each scale corresponding to N channels and includes prediction information; third, the network processes the network prediction results in order to process them and obtain the detected target. The network management system then processes the outcomes of the network prediction and obtains the found target. MR-YOLO introduces the MobileNetV3 [11] module to replace the YOLOv5 backbone, reducing the number of parameters and calculations, and speeding up the detection speed. At the same time, one further optimized the volume of the model, replacing the complex SPPF module with a lightweight SE-focused module. The backbone output uses PanNet as the core to enter the neck for bi-directional feature fusion, and then enters the head to detect targets of different sizes (seen from the architecture in Figure 3).

The YOLOv5 network now uses the portable MobileNetV3 module to significantly simplify the model while preserving detection accuracy and speed. In order to help the network choose better features, SE attention modules were also added to the backbone. The training data are supplemented with random noise using the Mosaic data enhancement approach. To increase the localization accuracy and regression speed of the algorithm, the SIoU loss function is utilized in place of the CIoU loss function. Experiments show that compared with YOLOv5 algorithm, MR-YOLO algorithm has better model expression and detection effect.

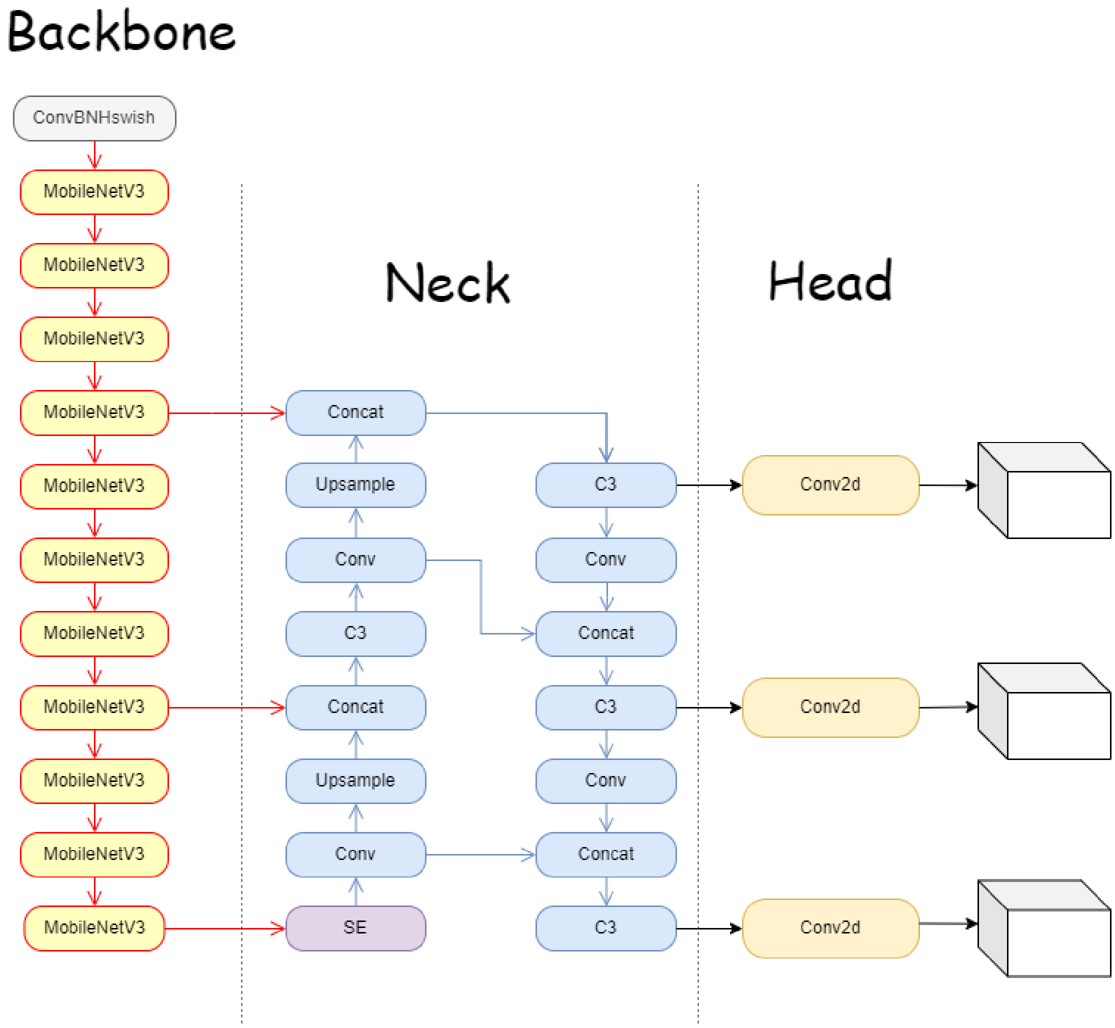

Figure 3. The architecture of MR-YOLO [11].

5. TPH-YOLOv5

TPH-YOLOv5 is an improved object detection model, built on YOLOv5, designed to address the three main challenges faced in object detection tasks in images captured by drones. Key improvements to the model include: First, to enhance sensitivity to tiny objects, TPH-YOLOv5 introduces an additional prediction head specifically designed to detect small, small, medium, and large objects. This additional head makes use of high-resolution features and multiple layers of advanced features to improve the detection performance of tiny objects. Second, in order to make more efficient use of context information, TPH-YOLOV5 replaces the original CNN-based prediction head with a transformer-based prediction head, called the transformer prediction Head (TPH). The transformer module helps capture the long-term relationship between the object and its surroundings, thereby improving detection accuracy in chaotic scenarios. In addition, to further improve performance, TPH-YOLOv5 employs several techniques, including convolutional Block Attention modules (CBAM), data enhancement, multi-scale testing (ms testing), multi-model integrated inference strategies, and two auxiliary classifiers for confounding classes, optimized specifically for confounding classes [11]. Overall, TPH-YOLOv5 significantly improves the performance of target detection in drone captured images by adding prediction heads, introducing transformer modules, optimizing attention mechanisms, and introducing various tricks. This allows it to better adapt to different scales and high-density scenarios, while maintaining flexibility and effectiveness. The architecture of TPH-YOLOv5 is illustrated in Figure 4.

TPH-YOLOv5 came in fourth place. Extensive testing on two benchmark datasets demonstrates that both of our models produce new SOTA outcomes and that tph-yolov5 ++, an upgraded variant, can perform as well as or better than TPH-YOLOv5 while dramatically reducing computational and memory costs.

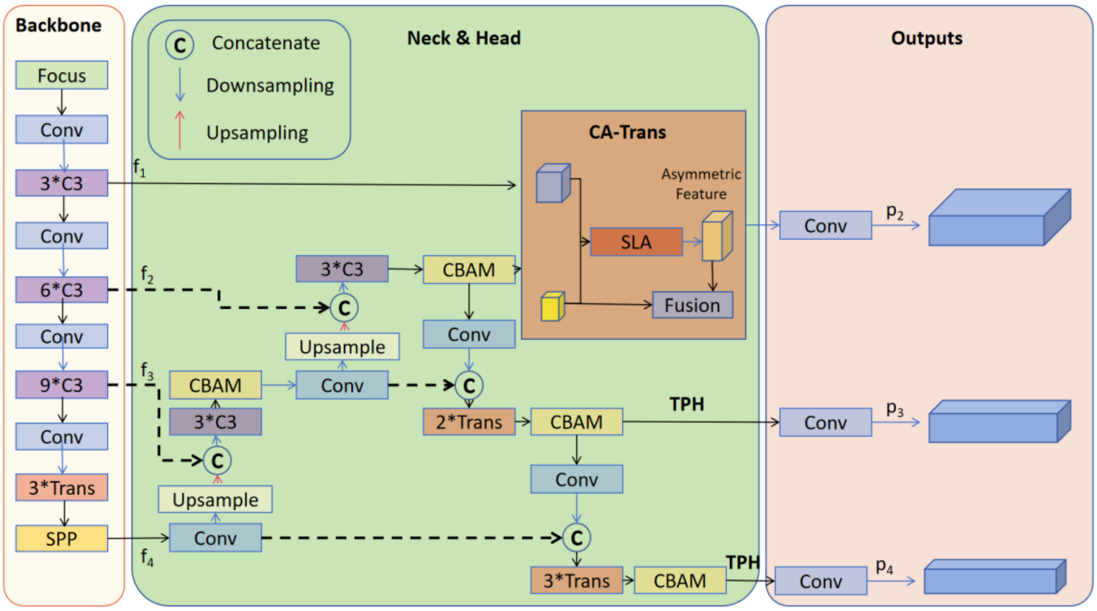

Figure 4. The architecture of TPH-YOLOv5 [11].

6. Limitations and prospects

The YOLO (You Only Look Once) algorithm has made remarkable progress in computer vision, but it also comes with some limitations. First, YOLO may not perform well when dealing with small objects because it uses grid cells for object detection, which tends to ignore small objects or misclassify them. Second, when there is overlap between targets, YOLO may generate multiple overlapping bounding boxes, resulting in inaccurate detection. In addition, YOLO has a limited ability to generalize to objects with different shapes, aspect ratios, and viewing angles, which can lead to problems with false detection or missed detection in complex scenes.

However, YOLO still has a wide range of applications and exciting prospects. Researchers are constantly improving and expanding the algorithm to overcome its limitations. Future YOLO variants may employ a more complex convolutional neural network architecture, introducing multi-scale feature fusion and attention mechanisms to improve detection performance and accuracy. In addition, techniques such as Generative Adversarial Networks (GANs) are available to implement to generate more realistic training data, helping to improve YOLO's generalization ability. With advances in hardware technology, YOLO can also be more widely used in embedded devices and edge computing, providing more powerful object detection capabilities for areas such as autonomous driving, intelligent surveillance, drones, and robotics. In addition, YOLO can be combined with technologies such as semantic segmentation and three-dimensional object detection to achieve a more comprehensive understanding of the scene. In summary, although the YOLO algorithm has some limitations, its future prospects in CV are still very bright. Through continuous research and improvement, YOLO is expected to be useful in more areas and provide more accurate, robust and efficient object detection solutions.

7. Conclusion

This paper delves into the evolution of the YOLO algorithm, from its inception to the YOLOv5 model, and explores the enhancements brought by three specialized variants: NRT-YOLO, MR-YOLO, and TPH-YOLOv5. These variants address specific challenges in object detection across diverse domains, ranging from remote sensing to industrial quality control and drone-captured scenarios. Throughout the paper, we've witnessed how YOLO has evolved as a pivotal technology in computer vision, with continuous improvements aimed at enhancing detection accuracy, speed, and adaptability. While each variant caters to distinct application areas, they collectively exemplify the versatility and potential of the YOLO framework.

However, it's essential to acknowledge that YOLO and its variants still face limitations, such as their performance on small objects and complex scenes. Looking forward, future research should aim to overcome these limitations and enable YOLO to excel in even more challenging scenarios. These results pave a path for the advancements and innovations within the YOLO family of algorithms, offering insights into the current state of object detection in computer vision. As YOLO continues to evolve and adapt, it holds the promise of revolutionizing various industries and applications by providing robust, real-time object detection solutions.

References

[1] Krizhevsky A, Sutskever I and Hinton G E 2017 Commun. ACM vol 60 pp 84–90.
[2] Redmon J, Divvala S, Girshick R and Farhadi A 2016 Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp 779–788.
[3] Everingham M, van Gool L, Williams C K I and Winn J Z 2010 Int J Comput Vis vol 88 pp 303–338.
[4] Terven J and Cordova-Esparza D 2023 arXiv preprint arXiv:230400501.
[5] Liu Y, He G, Wang Z, Li W and Huang H 2022 Sensors vol 22 p 4953.
[6] Lang X, Ren Z, Wan D, Zhang Y and Shu S 2022 Sensors vol 22 p 9897.
[7] Zhao Q, Liu B, Lyu S, Wang C and Zhang H T2023 Sensors vol 15 p 1687.
[8] Redmon J, Divvala S, Girshick R and Farhadi A 2016 Proceedings of the IEEE conference on computer vision and pattern recognition pp 779–788.
[9] Jocher G. YOLOv5 by Ultralytics, Retrieved from: https://githubcom/ultralytics/yolov5.
[10] Guo Z, Wang C, Yang G, Huang Z and Li G 2022 Sensors vol 22 p 3467.
[11] Sandler M, Howard A, Zhu M, Zhmoginov A and Chen L C 2018 Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp 4510–4520.