Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/77/2024MA0067

Optimization of autonomous driving image detection based on RFAConv and triplet attention

Zhipeng Ling[1a,*], Qi Xin[1b], Yiyu Lin[2], Guangze Su[3], Zuwei Shui[4]

[1a]Computer Science, University of Sydney, Sydney, Australia
[1b]Management Information Systems, University of Pittsburgh, Pittsburgh, PA, USA
[2]Computer Science and Engineering, Santa Clara University, CA, USA
[3]Information Studies, Trine University, Phoenix, USA
[4]Information Studies, Trine University, Phoenix, USA

*Corresponding author:zlin5578@alumni.sydney.edu.au

Abstract. YOLOv8 plays a crucial role in the realm of autonomous driving, owing to its high-speed target detection, precise identification and positioning, and versatile compatibility across multiple platforms. By processing video streams or images in real-time, YOLOv8 rapidly and accurately identifies obstacles such as vehicles and pedestrians on roadways, offering essential visual data for autonomous driving systems. The study introduces two pivotal modifications: the C2f_RFAConv module and the Triplet Attention mechanism. Firstly, the proposed modifications are elaborated upon in the methodological section. The C2f_RFAConv module replaces the original module to enhance feature extraction efficiency, while the Triplet Attention mechanism enhances feature focus. Subsequently, the experimental procedure delineates the training and evaluation process, encompassing training the original YOLOv8, integrating modified modules, and assessing performance improvements using metrics and PR curves. The results demonstrate the efficacy of the modifications, with the improved YOLOv8 model exhibiting significant performance enhancements, including increased MAP values and improvements in PR curves. Lastly, the analysis section elucidates the results and attributes the performance improvements to the introduced modules. C2f_RFAConv enhances feature extraction efficiency, while Triplet Attention improves feature focus for enhanced target detection.

Keywords: YOLOv8, Autonomous driving, C2f_RFAConv, Triplet Attention mechanism, Performance improvement.

1. Introduction

With the rapid development of autonomous driving technology, the accurate perception and understanding of the surrounding environment of the vehicle has become crucial. Image detection, as the core component of the automatic driving system, directly affects the decision-making and safety performance of the vehicle. In recent years, deep learning, especially convolutional neural networks (CNNs), have made remarkable progress in image detection tasks, such as the [1]YOLO series models (You Only Look Once), which perform well in real-time and detection accuracy. However, faced with the complex and changeable traffic environment, existing methods still face challenges in feature extraction and multi-scale information processing. In practical applications of autonomous driving,

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/77/2024MA0067

image detection needs to deal with a large number of complex scenes, including different lighting conditions, weather changes, and dynamic targets. These factors present the following challenges to detection models:

1. Multi-scale target detection: Targets in the traffic environment vary in size, ranging from distant pedestrians to nearby vehicles, requiring detection models to have strong multi-scale perception abilities.

2. Real-time processing: The automatic driving system requires efficient computing performance to ensure real-time processing and response, thereby avoiding safety hazards caused by delays.

3. Robustness: [2]The model needs to maintain stable performance in various extreme conditions, including rain, night, and occlusion.

4. Accuracy of feature extraction: In complex traffic scenes, accurate extraction of useful features is crucial for target recognition and classification.

Therefore, to address the above challenges, this paper proposes an improved YOLOv8 model that combines the RFAConv and Triplet Attention mechanisms. RFAConv (Receptive Field Attention Convolution) and the Triplet Attention mechanism are introduced to optimize the performance of the YOLOv8 model and improve the image detection capability of the automatic driving system. This paper will detail the design and implementation of these two mechanisms and verify their effectiveness through experiments.

2. Related Work

2.1. Automatic driving image detection technology

Normally, autopilot uses eight cameras to identify objects in the real world. The images captured by the cameras include pedestrians, vehicles, animals, and obstacles, which are important not only for the safety of the drivers of unmanned autonomous vehicles, but also for others. It is important that the camera is able to identify these objects in a timely and accurate manner.

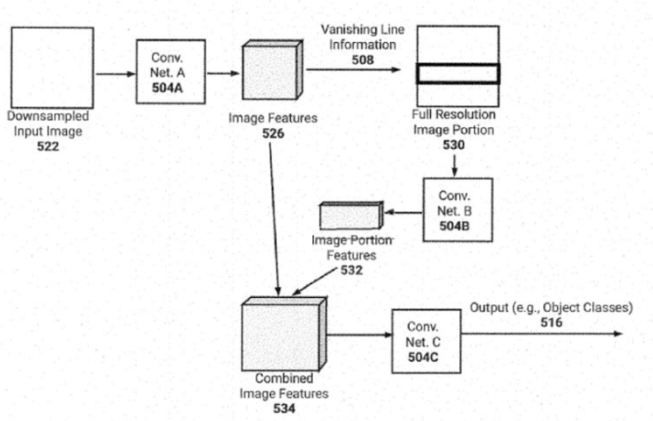


Figure 1. Automatic driving image detection

Breakthroughs in deep learning and computer vision have facilitated the rapid development of autonomous driving, which promises to free up drivers, reduce traffic congestion and improve road safety. However, the potential of autonomous driving has yet to be fully unlocked, largely due to unsatisfactory perceptual performance in real-world driving scenarios. As a result, even though autonomous vehicles (AVs)[3] are already being used in many restricted and controlled environments, deploying them in urban environments still faces serious technical challenges. For example, camera-based systems lack accurate depth information, while liDAR[4] only systems are hampered by low resolution input data, especially over long distances. Table 1 shows this

The YOLOv8 model excels in image detection for autonomous driving due to its real-time processing capabilities, high detection accuracy, and robustness in handling complex and dynamic traffic environments. By efficiently processing multi-scale information and accurately extracting features,

DOI: 10.54254/2755-2721/77/2024MA0067

YOLOv8 enables reliable target recognition and classification, contributing to the safety and efficiency of autonomous driving systems.

2.2. YOLOv8 model

The YOLO (You Only Look Once) series of models became very famous in the field of computer vision. The object detection and tracking model YOLOv8 can quickly and accurately identify and locate multiple objects in an image or video frame, as well as track their movement and classify them. In addition to detecting objects, YOLOv8[5] can also distinguish the exact contours of objects, perform instance segmentation, estimate human posture, and help identify and analyze specific patterns in medical images, among other computer vision tasks.

The main functions of YOLOv8 model include:

1. High-speed target detection: YOLOv8 continues to maintain the high-speed detection characteristics of YOLO series models, capable of real-time processing of video streams or high-speed analysis of targets in static images.

2. High-precision recognition: Through the improved algorithm and network structure, YOLOv8 improves the accuracy of target detection, including better boundary frame positioning and classification accuracy.

3, multi-platform compatibility: YOLOv8[6] supports deployment through a variety of formats such as ONNX, OpenVINO, CoreML and TFLite, enhancing the availability and compatibility of the model, enabling it to run on a variety of hardware and platforms.

4. Multi-task capability: In addition to target detection, YOLOv8 also supports tasks such as instance segmentation, image classification and pose estimation, providing a one-stop solution for a variety of visual recognition needs.

In the field of segmented tasks, a milestone in the field is the full convolutional network (FCN). In addition, U-Net and SegNet are common models in drivable area segmentation tasks. However, lane line segmentation is different from driveable area segmentation because lane lines have a distinct elongated feature in road images. The segmentation of lane lines requires effective analysis of Low-Level and multi-scale features.

Recently, models such as PointLaneNet and [7]MFIALane have gained a lot of attention in the field of lane segmentation. Although they achieve excellent results in their respective fields, integrating them together is challenging due to the different resolutions of the required features, with segmentation operations performed at the pixel level and detection tasks using grid cells in a one-stage approach and selective search in a two-stage approach. Despite their different focal points, both segmentation and detection tasks need to extract initial features from the input image. Therefore, a Backbone network can be shared. Integrating three different necks and heads into a model with a shared Backbone saves a lot of computational resources and inference time compared to using separate models for each task.

2.3. Triplet Attention mechanism principle

The recent proliferation of attention mechanisms across various computer vision tasks underscores their efficacy in leveraging interdependencies among channels or spatial positions. In this paper, we explore a lightweight yet effective attention mechanism termed Triplet Attention, which captures cross-dimensional interactions using a tripartite structure to compute attention weights. We validate the effectiveness of our approach on challenging tasks including image classification on ImageNet-1k and object detection on [8]MSCOCO and PASCAL VOC datasets. Furthermore, through intuitive examination of GradCAM and GradCAM++ results, we offer deeper insights into the performance of Triplet Attention.

DOI: 10.54254/2755-2721/77/2024MA0067

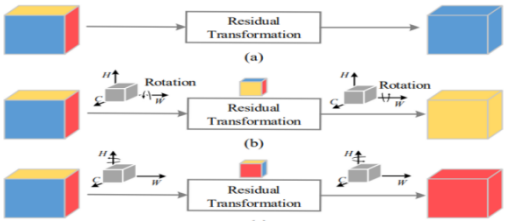


Figure 2. Schematic diagram of Triplet Attention

The basic principle of Triplet Attention lies in leveraging a tripartite structure to capture cross-dimensional interactions in input data, thereby computing attention weights. This method effectively establishes interdependencies between channels or spatial positions, with minimal computational cost.

This process effectively enhances the model's ability to focus on relevant features across different dimensions, facilitating more effective feature extraction and representation. For example, in image classification tasks, Triplet Attention can help the model to attend to important spatial and channel information simultaneously, improving classification accuracy across various classes and scenarios.

2.4. Triplet Attention and other simple attention mechanisms

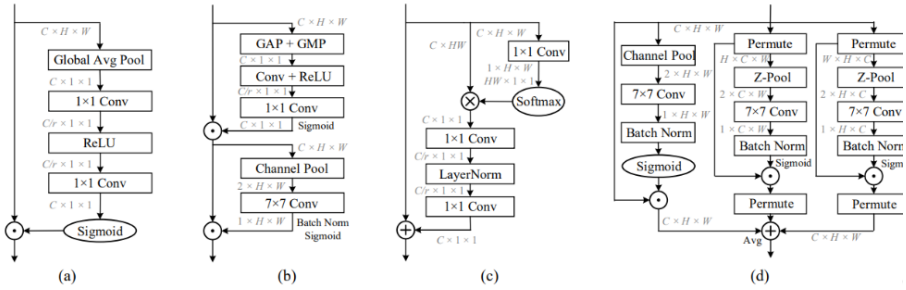


Figure 3. Comparison of Triplet Attention with other simple attention mechanisms

In the realm of attention mechanisms for enhancing neural network performance in computer vision tasks, various modules have been proposed. Convolutional Block Attention Module (CBAM), Additionally, its modular design facilitates seamless integration into existing network architectures, thereby augmenting their understanding and processing capabilities for complex data structures.

In summary, Triplet Attention stands out among other attention mechanisms in computer vision tasks due to its unique architecture and efficient computation. This comparative analysis sets the stage for further exploration of Triplet Attention's effectiveness in the experimental section.

3. Methodology

3.1. Focus on the spatial characteristics of receptive field

The receptive field spatial feature refers to the local region of input data that the convolutional layer can "see" in the convolutional neural network[10] (CNN). In CNN, the output of each convolution operation is a small window, or a local receptive field, based on the input data. Specifically, the RFAConv module utilizes an adaptive receptive field mechanism to weight the input data at each location to increase attention to important features, thereby enhancing the model's ability to understand complex scenes.With the RFAConv module, we are able to extract key information from input data more accurately, providing a more reliable feature representation for subsequent target detection and recognition tasks.

DOI: 10.54254/2755-2721/77/2024MA0067

3.2. Mechanism to solve parameter sharing problem

RFAConv convolution solves the problem of parameter sharing by introducing an attention mechanism that allows the network to assign a specific weight to each perception. Specifically, RFAConv uses spatial attention to determine the importance of each position in the receptive field and adjusts the weight of the convolution kernel accordingly. [11-12]This dynamically adjusted parameter sharing mechanism enables the network to better capture the differences between different regions, thus enhancing the model's understanding and representation of image content.

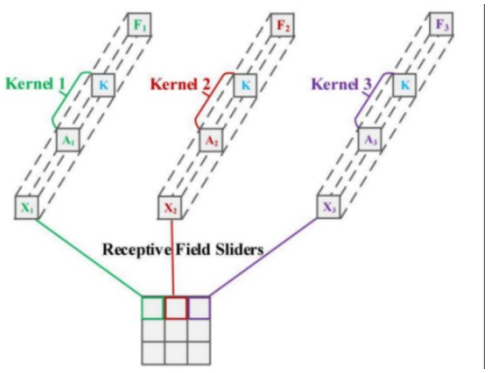


Figure 5. Schematic diagram of RFAConv convolution operation

During the RFAConv convolution operation, the convolution kernel parameter K is multiplied by the attention weight A to generate a customized convolution kernel for each receptor field position. Specifically, this process combines the attention mechanism with the convolutional kernel to produce a customized convolutional kernel for each receptive field location. For example, Kernel1, Kernel2, and Kernel3 in the figure are obtained by multiplying the common convolution kernel parameter K with the corresponding attention weights A1, A2, and A3, respectively.

RFAConv dynamically adjusts the weights of convolutional kernels by leveraging receptive field attention mechanism, providing customized attention for feature extraction in each region. This adjustment allows the network to handle each local region more finely, better capturing and responding to specific features of the input data rather than simply applying the same weights to the entire image. Such an approach enhances the network's understanding and representation of features, thereby improving learning and prediction outcomes.

In summary, through this approach, [13]RFAConv enhances the model's expressive power, allowing it to more accurately adapt to and represent the features of input data, especially when dealing with complex or variable image content. This flexible parameter adjustment mechanism provides a new pathway for improving the performance and generalization capability of convolutional neural networks.

4. Experimental Procedure

4.1. Experimental design

The main objective of this experiment is to improve the performance of YOLOv8 in target detection by improving its network structure. We use C2f_RFAConv to replace the C2f module in the original YOLOv8 network, and introduce RFAConv and Triplet Attention modules. We will verify the improvement effect by comparing the performance indicators before and after the improvement.

1. Experimental environment:

- Hardware: GPU (e.g. NVIDIA Tesla V100)
- Software: Python 3.x, PyTorch, YOLOv8 framework

2. Data set:

- Training and testing using the COCO dataset.

DOI: 10.54254/2755-2721/77/2024MA0067

4.2. Experimental Model

1.Original YOLOv8 network training:

Using the original YOLOv8 network structure, we trained on the COCO dataset, saving the model and log files during the training process. After the training, we recorded the final performance indicators with MAP(50) values of 0.326 and MAP(50-95) values of 0.187. In order to visually demonstrate the detection performance of the model, the precision rate-recall ratio (PR) curve of the original YOLOv8 was drawn and saved in this experiment.

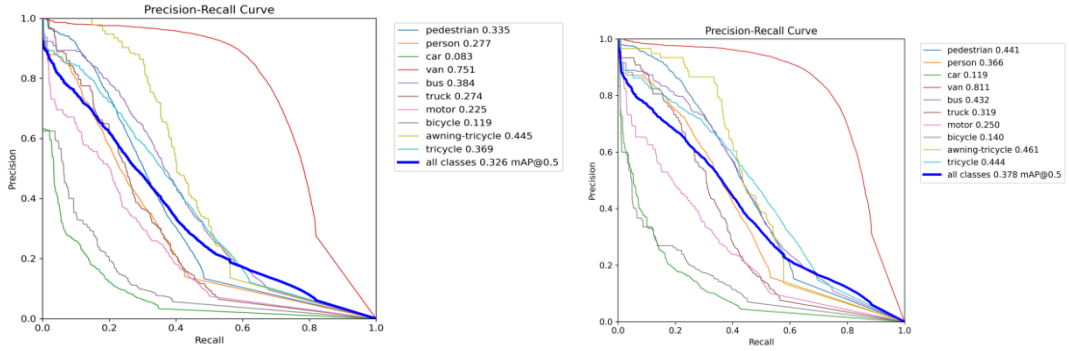


Figure 6. Accuracy - recall rate (PR) curve of the original YOLOv8 and YOLOv8-C2f_RFAConv

Figure PR curve shows the relationship between accuracy rate and recall rate of the model under different thresholds. The closer the curve is to the upper right corner, the better the performance of the model. In the PR curve, we can see the trend of the accuracy rate and recall rate of various targets with the detection threshold, which helps us to evaluate the performance of the model under different detection difficulties. Through these indicators and curves, we can fully understand the detection effect of the original YOLOv8 on the COCO data set, and provide a benchmark and reference for subsequent improvement.

The implementation code for the C2f_RFAConv module is provided, which contains detailed definitions of the C2f_RFAConv and RFAConv classes. RFAConv makes convolution operations more efficient by introducing attention mechanisms and multi-scale feature extraction. We integrated these new modules into the YOLOv8 network to make sure the code was up and running. Through training on the COCO dataset, we verify the correctness and validity of the model.

2.Performance comparison and analysis:

By comparing the performance indicators of the original YOLOv8 and the improved YOLOv8, we mainly focus on the changes of MAP(50) and MAP(50-95). The results show that the improved YOLOv8's MAP(50) has been improved from 0.326 to 0.385, and MAP(50-95) has been improved from 0.187 to 0.217. After calculating the performance improvement of each category, it is found that the detection accuracy of all categories has been improved in different degrees. The reasons for the performance improvement after the introduction of C2f_RFAConv and Triplet Attention modules are analyzed in detail.

4.3. Experimental Result

- Original YOLOv8: MAP(50): 0.326-MAP(50-95): 0.187

DOI: 10.54254/2755-2721/77/2024MA0067

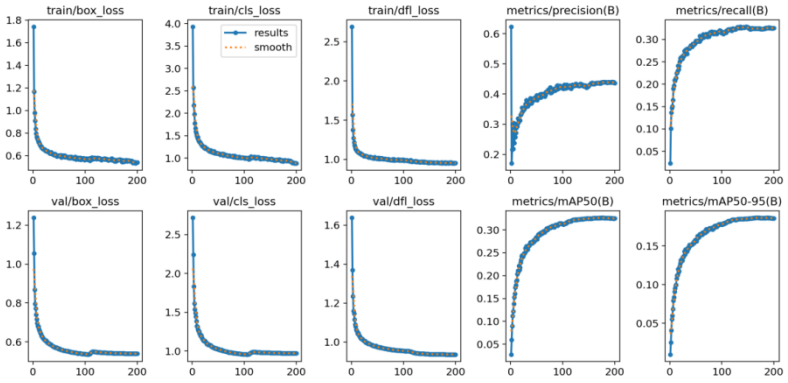


Figure 9. Training results of the original YOLOv8 model

- Improved YOLOv8-C2f_RFAConv-Triplet Attention-P2: MAP(50): 0.385-MAP(50-95): 0.217

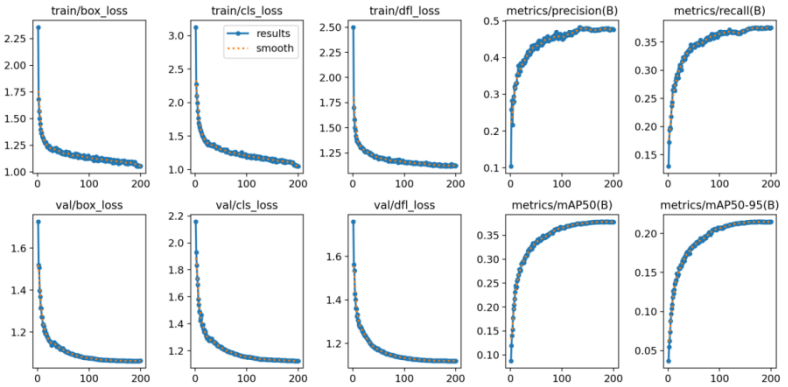


Figure 10. Result of improved YOLOv8-C2f_RFAConv-Triplet Attention-P2

Comparing the performance indicators of the original YOLOv8 and the improved YOLOV8-C2F_RfaconV-triplet Attention, the MAP(50) of the improved model has been improved from 0.326 to 0.385, with an increase of 0.059, and the MAP(50-95) has been improved from 0.187 to 0.217. This trend further validates the effectiveness of introducing C2f_RFAConv and Triplet Attention modules, which makes the model perform better in feature extraction and attention mechanism.

In this experiment, the addition of detection head P2 layer usually has higher resolution and is able to capture more details of small size targets, helping to improve small target detection by providing richer spatial information.

C2f_RFAConv uses weighted feature extraction in RFAConv to capture important feature information more accurately and improve the expression ability of feature maps. The multi-scale feature extraction mechanism makes the model more flexible when dealing with objects of different sizes, and enhances the overall detection ability of the model.The combination of C2f_RFAConv and Triplet Attention module greatly enhances the feature extraction capability and detection accuracy of YOLOv8, making the model perform better in various detection tasks.

5. Conclusion

In conclusion, YOLOv8 stands as a cornerstone technology in the realm of autonomous driving, offering rapid and precise target detection capabilities vital for ensuring road safety and efficiency. This study has presented a comprehensive enhancement approach for the YOLOv8 model, addressing its crucial role and the challenges faced in autonomous driving scenarios. By introducing the C2f_RFAConv module and the Triplet Attention mechanism, significant improvements have been achieved in feature

DOI: 10.54254/2755-2721/77/2024MA0067

extraction efficiency and feature focusing, respectively. Ultimately, the combination of these modifications results in a marked enhancement in the overall performance of the YOLOv8 model, promising improved safety and efficacy in autonomous driving systems.

In summary, the findings of this study underscore the pivotal role of advanced computer vision techniques, such as those integrated into the YOLOv8 model, in advancing the capabilities of autonomous driving systems. The successful implementation of the C2f_RFAConv module and the Triplet Attention mechanism has not only demonstrated the adaptability and versatility of the YOLOv8 framework but has also showcased its potential for further refinement and optimization. By leveraging these enhancements, the improved YOLOv8 model exhibits enhanced feature extraction efficiency and improved target detection accuracy, paving the way for safer and more efficient autonomous driving experiences in real-world scenarios.

References

[1] Laroca, R., Severo, E., Zanlorensi, L. A., Oliveira, L. S., Gonçalves, G. R., Schwartz, W. R., & Menotti, D. (2018, July). A robust real-time automatic license plate recognition based on the YOLO detector. In 2018 international joint conference on neural networks (ijcnn) (pp. 1-10). IEEE.

[2] Huang, R., Pedoeem, J., & Chen, C. (2018, December). YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers. In 2018 IEEE international conference on big data (big data) (pp. 2503-2510). IEEE.

[3] Du, J. (2018, April). Understanding of object detection based on CNN family and YOLO. In Journal of Physics: Conference Series (Vol. 1004, p. 012029). IOP Publishing.

[4] Raj, T., Hanim Hashim, F., Baseri Huddin, A., Ibrahim, M. F., & Hussain, A. (2020). A survey on LiDAR scanning mechanisms. Electronics, 9(5), 741.

[5] Ćorović, A., Ilić, V., Đurić, S., Marijan, M., & Pavković, B. (2018, November). The real-time detection of traffic participants using YOLO algorithm. In 2018 26th Telecommunications Forum (TELFOR) (pp. 1-4). IEEE.

[6] Shinde, S., Kothari, A., & Gupta, V. (2018). YOLO based human action recognition and localization. Procedia computer science, 133, 831-838.

[7] Hussain, Muhammad. "YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection." Machines 11.7 (2023): 677.

[8] Sarda, Abhishek, Shubhra Dixit, and Anupama Bhan. "Object detection for autonomous driving using yolo [you only look once] algorithm." 2021 Third international conference on intelligent communication technologies and virtual mobile networks (ICICV). IEEE, 2021.

[9] Benjumea, A., Teeti, I., Cuzzolin, F., & Bradley, A. (2021). YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles. arXiv preprint arXiv:2112.11798.

[10] Liang, S., Wu, H., Zhen, L., Hua, Q., Garg, S., Kaddoum, G., ... & Yu, K. (2022). Edge YOLO: Real-time intelligent object detection system based on edge-cloud cooperation in autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems, 23(12), 25345-25360.