ISSN 2616-5775 Vol. 6, Issue 10: 31-38, DOI: 10.25236/AJCIS.2023.061005

Research on Vehicle Detection and Recognition Algorithm Based on Improved YOLOv5

Yujiao Liu[a], Xuan Guo[b]

College of Electronic Information, Dongguan Polytechnic, Dongguan, 523808, Guangdong, China
[a]lyj_198898@126.com, [b]guoguo_x@hotmail.com

Abstract: *This paper aims to study and improve the pedestrian and vehicle detection and recognition algorithm based on YOLOv5. Firstly, the network structure of YOLOv5 is introduced, including the backbone network, neck network, and post-processing algorithm. In order to address the challenges of pedestrian and vehicle detection, this paper carefully improves the backbone network, neck network, and post-processing algorithm. Experimental results show that the improved algorithm achieves higher accuracy and better performance in pedestrian and vehicle detection tasks. By comparing the performance of different modules before and after improvement, as well as comparing with other algorithms, the superiority of the algorithm is validated. This research is of great significance for improving the application of pedestrian and vehicle detection and recognition algorithms in areas such as traffic management, intelligent monitoring, and autonomous driving, and provides useful references for related research in these fields.*

Keywords: *Improved YOLOv5; Pedestrian Vehicle Detection; Recognition Algorithm*

1. Introduction

With the rapid development of computer vision and deep learning technologies, pedestrian and vehicle detection and recognition play an important role in areas such as traffic management, intelligent monitoring, and autonomous driving. Currently, deep learning-based object detection algorithms have made significant progress, and the YOLO series algorithms have gained widespread attention for their speed and accuracy. However, YOLOv5, as the latest version in the YOLO series, still faces challenges in the accuracy and efficiency of pedestrian and vehicle detection. By verifying through experiments and comparing the performance of the algorithm before and after improvement, we can conclude that the improved algorithm achieves higher accuracy and efficiency in pedestrian and vehicle detection and recognition tasks. The results of this research are of great significance for the further development and practical application of pedestrian and vehicle detection technology.

2. YOLOv5 Network Architecture

YOLOv5 (You Only Look Once) is a grid-based object detection algorithm that divides the input image into an S × S grid. If the center point of an object falls within a grid cell, that grid cell is responsible for predicting the bounding box and class information for that object. Each grid cell predicts B bounding boxes and one class, where the bounding boxes contain the position and confidence information of the objects. To enhance detection accuracy and adapt to different scenarios, the YOLOv5 network predicts 3 bounding boxes per grid cell, each containing coordinate information (x, y, w, h), a confidence score, and C conditional class probability information [1].

Currently, YOLOv5 provides four models: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, which gradually increase the width and depth of the network.

At the input stage, the input image undergoes preprocessing, which includes adaptive scaling to a unified standard size and the addition of adaptive anchor box calculations. During the training process, the algorithm dynamically calculates the optimal anchor box values for different training sets. In order to enrich the dataset, improve the robustness of the network, and enhance the performance of small object detection, the Mosaic data augmentation method is employed. This method randomly selects four images for operations such as rotation, scaling, and cropping, and then randomly stitches them together to generate new training data. This approach not only enriches the dataset but also improves

ISSN 2616-5775 Vol. 6, Issue 10: 31-38, DOI: 10.25236/AJCIS.2023.061005

the learning and detection capabilities of the network. The effect of Mosaic data augmentation is shown in Figure 1.

Figure 1. Mo-saic data enhancement effects

The backbone network of YOLOv5 consists of four modules: Focus, CBL, CSP, and SPP, which are used to extract features of the objects in the image. The CSP module adopts the idea of Cross-Stage Partial network, which solves the problem of redundant gradient information in the network design during the inference process. This enhances the learning capability of the network and optimizes the computational cost. Additionally, the Focus structure is introduced to perform slicing operations, which downsample the image by a factor of 2 before entering the backbone network, resulting in four feature maps. This process consolidates the width and height information of the image into the spatial channel without losing feature information, allowing the network to extract richer features. For example, in the YOLOv5s model, the original 608×608×3 image is input into the Focus structure, and after the slicing operation, a 304×304×12 feature map is generated [2]. Then, it undergoes convolution with 32 convolutional kernels to create the final 320×320×32 feature map. The schematic diagram of the slicing operation in the Focus module can be referred to in Figure 2.

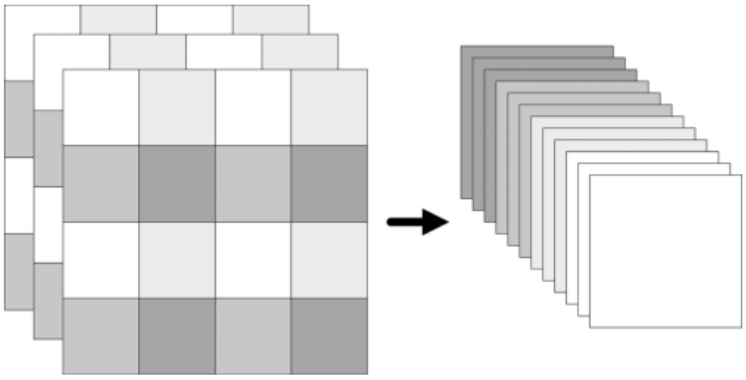

Figure 2. Schematic diagram of the slicing operation of the Focus module

In conclusion, YOLOv5 is an efficient and accurate object detection algorithm that achieves fast and precise object recognition through grid division and multi-bounding box prediction. The design and optimization of its backbone network further enhance the capability of feature extraction. With data augmentation and adaptive anchor box calculations, YOLOv5 demonstrates increased robustness and adaptability. In the future, the development of YOLOv5 will continue to drive the application and advancement of object detection technology in various fields [3].

The YOLOv5 network architecture is a deep learning model used for object detection, and it has strong overall detection performance. However, it is not specifically designed for vehicle detection, and there is still room for improvement in detecting dense vehicle targets and low-light environments. In order to achieve real-time and accurate vehicle detection in applications such as autonomous driving, while meeting the requirements of network lightweight, this study selects the YOLOv5s network as the foundation and makes improvements to enhance the accuracy of vehicle detection.

The neck network in the YOLOv5 network architecture plays an important role in generating the feature pyramid, enhancing the model's detection capability for objects of different scales, and recognizing the same object at different sizes and scales. This network adopts the FPN-PAN network structure. FPN combines upsampling to fuse deep semantic information with bottom-level target position information, enhancing the network's feature learning ability. Based on this, the PAN structure

ISSN 2616-5775 Vol. 6, Issue 10: 31-38, DOI: 10.25236/AJCIS.2023.061005

is integrated to propagate strong localization features from bottom to top, comprehensively improving the network's learning performance for semantic information and localization information. In this way, the neck network can generate multi-scale feature vectors to predict image features, generate bounding boxes, and predict the category and confidence of objects in the image.

The output layer (Head) in the YOLOv5 network consists of three detection heads, which downsample the original image by 8, 16, and 32 times, respectively, generating three different-sized feature vectors. These feature vectors are used to predict image features, generate bounding boxes, and predict the category and confidence of objects in the image. By introducing multiple detection heads, the YOLOv5 network can capture objects at different scales and improve the accuracy of object detection.

To address the shortcomings of YOLOv5 in vehicle detection, this study selected the YOLOv5s network as the base and made improvements to enhance the accuracy of vehicle detection. The focus of the improvements mainly lies in adjusting the network's architecture and parameters to meet the requirements of vehicle detection. Through training and optimization on the dataset, the improved network can better handle complex detection tasks, such as dense vehicle targets and low-light environments.

3. Improvement of Pedestrian Vehicle Detection Algorithm Based on YOLOv5s

3.1. Backbone Network (Backbone) Improvement

The backbone network plays a crucial role in deep learning-based object detection models as it is responsible for extracting the features of the objects in the image. In YOLOv5, the idea of the VOVNet network structure is incorporated to improve the backbone network. The CSP-OSA module is introduced to replace the original CSP1 structure, further enhancing the network's performance and detection accuracy. Traditional backbone networks mainly utilize Convolutional Neural Networks (CNNs) to extract features. However, in this paper, inspired by the design principles of VOVNet, the CSP-OSA module is proposed to construct a new backbone network structure. The CSP-OSA module effectively inherits some advantages of VOVNet, making the network more suitable for object detection tasks. The CSP-OSA module combines the concepts of Cross Stage Partial (CSP) and Object Spatial Attention (OSA). The CSP idea solves the problem of redundant gradient information during the inference process in network design, thereby enhancing the learning capability and runtime efficiency of the network. The OSA module introduces target spatial attention, which enables adaptive learning and focus on important target regions in the feature maps, improving the network's perception and representation of the objects.

The alternative improvement of the CSP-OSA module allows the backbone network to extract target features more effectively. By introducing the CSP-OSA module, the network can accurately capture feature information of different scales and dimensions, resulting in more rich and accurate representations of the objects [4]. Additionally, the integration of VOVNet's design principles in the CSP-OSA module further enhances the learning capability and computational efficiency of the network. The CSP-OSA module is an improved module designed in this paper. Based on the OSA module, it enhances the network's feature extraction capabilities and the richness of gradient combinations by adding the CSPNet structure. In the CSP-OSA module, the feature maps of the base layers are divided into two parts. One part is directly connected to the end of the stage through convolution, while the other part is processed by the OSA module. The fusion of these two feature maps enables more rich gradient combinations, thereby further improving the network's feature extraction capabilities.

The CSP-OSA module is based on the OSA module and introduces the CSPNet structure, aiming to further enhance the network's capability in feature extraction through feature segmentation and fusion. Specifically, the CSP-OSA module divides the feature maps of the base layers into two parts, where one part is directly connected to the end of the stage, and the other part is processed by the OSA module. This design allows the network to combine the information from both parts of the feature maps and achieve more comprehensive feature extraction.

In summary, the CSP-OSA module adds the CSPNet structure on top of the OSA module. By performing feature segmentation and fusion, the network can better utilize features of different scales and levels, thus improving the feature extraction capability and detection performance. The introduction of the CSP-OSA module is of great significance for the development of deep learning-based object detection models, providing strong support for more accurate and efficient object

ISSN 2616-5775 Vol. 6, Issue 10: 31-38, DOI: 10.25236/AJCIS.2023.061005

detection in practical applications. The module structure diagram, as shown in Figure 3, illustrates the composition of the CSP-OSA module and the process of feature fusion [5].

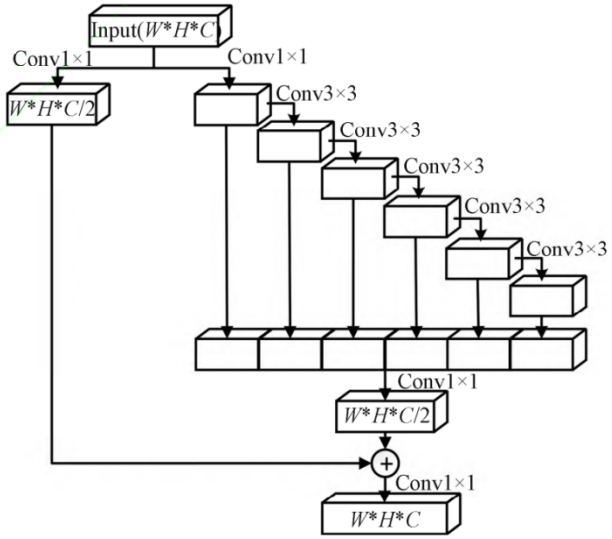

Figure 3. CSP-OSA module structure

3.2. Improvements to the Neck network

The neck network is an important component that connects the backbone network and the output layer in object detection models. It enhances the features extracted by the backbone network and outputs them to the final output layer, thus improving the detection accuracy of the network. In the field of deep learning, a commonly used feature enhancement module is the Feature Pyramid Network (FPN), which is widely used in tasks such as object detection and instance segmentation. Before the introduction of the FPN module, most object detection algorithms performed forward propagation in a bottom-up manner, only predicting on the top-level features. This led to insufficient utilization of the target position information in the bottom-level features, affecting the localization accuracy of the objects [6].

The FPN module addresses this issue by introducing a top-down structure and propagating deep semantic information to the shallow layers through lateral connections. This structure effectively utilizes the rich semantic information in the high-level features and combines it with the low-level features, enabling the network to achieve multi-scale object detection and effectively improve the detection capability of small objects. By individually predicting on each fused feature layer, the FPN module further enhances the expressive power of the network.

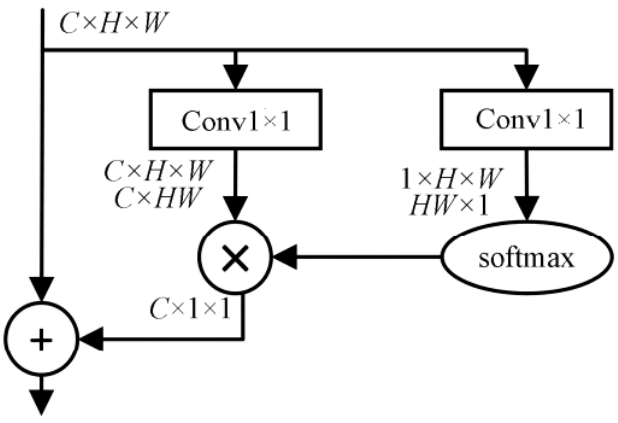

Figure 4. Simplified non-localized attention mechanism

The further improved FPN-PAN structure no longer uses the bottom-up structure of the PANet model. Instead, it aggregates the deep semantic information and shallow position information contained in all feature layers of FPN and enhances the features using an attention mechanism. Compared to the

ISSN 2616-5775 Vol. 6, Issue 10: 31-38, DOI: 10.25236/AJCIS.2023.061005

original Non-Local attention mechanism, the attention mechanism used in this paper has been simplified by removing one branch, making the model more concise and efficient. The improved structure is illustrated in Figure 4.

The introduction of the FPN-PAN structure further enhances the capability of the neck network in feature enhancement. By applying the attention mechanism, the model can more accurately capture important features in the image and achieve more precise localization and recognition of objects. This improvement allows the network to achieve better performance and accuracy in object detection tasks [7].

The neck network plays an important role in object detection tasks as it is responsible for generating the feature pyramid and enhancing the network's detection capability for objects of different scales. In YOLOv5, the neck network has been improved to further enhance the overall detection performance of the network. In YOLOv3, the feature pyramid network (FPN) structure was already used, which fuses deep semantic information with bottom-level target position information through upsampling, enhancing the network's feature learning ability. In YOLOv5, the FPN structure has been improved by incorporating the PA-Net network structure, further enhancing the effectiveness of the feature pyramid. The PA-Net structure introduces bottom-up feature propagation, transmitting rich positional information from shallow layers to deep layers, fully integrating high-level semantic information with shallow-level positional information, thereby further improving the overall detection performance of the network. However, although this improvement method has achieved significant effects in strengthening feature fusion and semantic and positional information integration, it still mainly focuses on information fusion between adjacent layers and lacks sufficient information exchange between non-adjacent layers. This limits the network's ability to detect small objects to some extent. To address this issue, this study proposes a feature enhancement model that enables information fusion between multiple feature layers.

This feature enhancement model introduces attention mechanisms to achieve information communication between multiple feature layers. Attention mechanisms can adaptively adjust the weights of features based on their importance, allowing the network to pay more attention to important features. In the neck network, attention mechanisms can be used to perform weighted fusion of features from non-adjacent layers, enhancing the detection capability of small objects. By enhancing the expression of features related to small objects, the network can more accurately detect and recognize small-sized targets. Additionally, further improving the performance of the neck network can be achieved by introducing more nonlinear transformations and multi-scale information fusion. For example, multi-scale convolutions can be used or more complex feature fusion mechanisms can be introduced to better capture target information at different scales.

3.3. Improvement of post-processing algorithms

The Non-Maximum Suppression (NMS) algorithm is an important post-processing step in object detection. Its purpose is to filter out redundant bounding boxes and find the best detection locations for objects. The traditional NMS algorithm sorts all the detection boxes based on their scores, selects the box with the highest score, and calculates the Intersection Over Union (IOU) between this box and the remaining detection boxes to filter out overlapping candidate boxes. When the IOU value is higher than a certain threshold, a group of bounding boxes is identified as the same object. This can be represented by the formula (1) as shown:

$$box = B_{\arg\max C_i} \tag{1}$$

In the task of vehicle detection in the field of autonomous driving, traditional NMS algorithms may not accurately localize the target vehicles due to complex situations such as dense and occluded target vehicles. This severely affects the determination of the relative positioning relationship between the target vehicles and the ego vehicle. To address this issue, this paper proposes the adoption of Weighted NMS algorithm for bounding box filtering, aiming to improve the overall performance of the algorithm and the localization accuracy of the bounding boxes.

The Weighted NMS algorithm improves upon the traditional NMS algorithm by introducing a weighting factor to adjust the scores of each detection box. This ensures that the detection box with the highest score has a higher priority in the subsequent filtering process. As a result, the suppressed non-maximum detection results may still contain the maximum value of the target features, which will not be directly ignored, thus avoiding inaccurate target localization. The formulas for the Weighted

ISSN 2616-5775 Vol. 6, Issue 10: 31-38, DOI: 10.25236/AJCIS.2023.061005

NMS algorithm are shown in formula (2), where the introduction of the weighting factor adjusts the scores of each detection box, enabling effective filtering of the detection results [8].

$$box = \frac{\sum_i \omega_i B_i}{\sum_i \omega_i}, B_i \in \{B \mid IOU(M,B) \geq thresh\} U \{box\}, \omega_i = C_i \times IOU(box, B_i) \tag{2}$$

The adoption of the Weighted NMS algorithm can effectively improve the performance of the object detection algorithm. By considering the weight of each detection box, the algorithm can more accurately localize the targets and eliminate the interference of overlapping boxes. In the field of autonomous driving, this improvement is crucial for accurately determining the relative positioning relationship between target vehicles and the ego vehicle. The application of the Weighted NMS algorithm enables more accurate localization of the output bounding boxes, thereby enhancing the overall performance of the algorithm and the detection capability of the targets.

4. Experimental Procedure and Analysis

4.1. Experimental data set

The experimental dataset used in this paper is a self-made dataset selected and modified based on the BDD100K autonomous driving dataset released by the University of California, Berkeley. This dataset contains diverse information, including different geographical, environmental, and weather factors. The dataset covers three vehicle object categories: car, bus, and truck. To ensure the rigor of the experiments, we divided the dataset into training set, test set, and validation set in a ratio of 7:2:1.

To improve the detection accuracy, generalization capability, and robustness of the model, we used the Z-Score normalization method to normalize the data. This ensures that the image data follows a standard normal distribution and reduces the impact of convergence speed and detection accuracy caused by large differences in the data. Additionally, we used the K-means clustering method to update the anchors of the network. Through this step, we can avoid the situation where the original anchors do not match the anchors of the current dataset due to changes in the dataset, thereby improving the accuracy and stability of the model. Furthermore, to increase the diversity and robustness of the training samples, we performed data augmentation by adjusting the saturation, exposure, and hue of the image data. This makes the training samples more representative and generalizable, thereby improving the model's performance in real-world scenarios [9].

4.2. Comparative analysis of improvements by module

In order to validate the impact of the improvement modules on the overall performance of the network, this paper adopts an incremental approach to gradually modify each improvement module and observe if it can improve the network performance. The experimental results are shown in Table 1.

Table 1. Impact of improved modules on network performance

network	mAP.5	mAP.5:.95	Inference time /s
YOLOv5s	0.6702	0.4541	0.008
I-YOLOv5s	0.6951	0.482	0.010
II-YOLOv5s	0.699	0.4802	0.011
III-YOLOv5s	0.7005	0.4811	0.011

The data in Table 1 indicates that with the improvement of each module in the network, both mAP.5 and mAP.5:0.95 values gradually increase. In particular, modifying the backbone network leads to the most significant improvement in network performance. It is noteworthy that despite the improvements in each module, the inference time did not significantly increase. These results suggest that by incrementally improving each module, we can significantly enhance the object detection performance of the network. The improvement in the backbone network is particularly crucial and has the greatest impact on enhancing network performance. Furthermore, the fact that the improvement modules do not significantly increase the inference time demonstrates that the improvement methods not only enhance

ISSN 2616-5775 Vol. 6, Issue 10: 31-38, DOI: 10.25236/AJCIS.2023.061005

performance but also have good efficiency.

4.3. Comparative analysis of YOLOv5 model before and after improvement

In order to understand the difference between the performance of the improved network III-YOLOv5s and the original network, the training results of the two models are compared, as shown in Table 2:

Table 2. Comparison of network performance before and after improvement

Network model	precision	recall	mAP	Inference time /s
YOLOv5s	0.4317	0.7281	0.6709	0.009
III-YOLOv5s	0.456	0.7543	0.7007	0.013

In the experiment, we tested the improved network III-YOLOv5s and plotted Fig. 5 to show its detection effect. From the figure, we can see that the improved network III-YOLOv5s is able to recognize the vehicle targets better, both at different vehicle densities and at different light intensities.

Figure 5. III-YOLOv5s actual detection effect

The improved network demonstrates good robustness when facing complex environments with different sizes of vehicle targets and varying lighting conditions. This means that the improved network possesses the capability to successfully perform vehicle detection in practical applications, and it exhibits relatively strong detection capabilities. These results indicate that the improved network can maintain high detection accuracy and reliability in different environmental conditions. Whether facing high-density vehicle crowds or poor lighting conditions, the improved network can effectively accomplish the object detection task. These experimental results validate the robustness and stability of the improved network, providing a reliable solution for vehicle detection in real-world scenarios [10].

5. Conclusion

This paper investigates the pedestrian and vehicle detection and recognition algorithm based on the improved YOLOv5. The backbone network, neck network, and post-processing algorithm are improved to enhance the accuracy and efficiency of detection. The experimental results demonstrate significant achievements in pedestrian and vehicle detection tasks with the improved algorithm, exhibiting higher accuracy and faster processing speed. Compared to the original YOLOv5 algorithm and other object detection algorithms, the improved algorithm possesses notable advantages in performance. The

ISSN 2616-5775 Vol. 6, Issue 10: 31-38, DOI: 10.25236/AJCIS.2023.061005

research findings have practical implications in areas such as traffic management, intelligent surveillance, and autonomous driving, providing effective methods and insights for advancing pedestrian and vehicle detection technology. Future research can further optimize the network architecture and parameter settings of the algorithm to address the challenges of pedestrian and vehicle detection and recognition in more complex scenarios, achieving higher levels of accuracy and efficiency.

Acknowledgements

Key project of Dongguan Polytechnic University-level foundation in 2022; Number: 2022a09.

References

[1] Li M L, Sun G B, Yu J X. A pedestrian detection network model based on improved YOLOv5[J]. Entropy, 2023, 25(2): 381.
[2] Xu H, Zheng W, Liu F, et al. Unmanned Aerial Vehicle Perspective Small Target Recognition Algorithm Based on Improved YOLOv5[J]. Remote Sensing, 2023, 15(14): 3583.
[3] Liu H, Sun F, Gu J, et al. Sf-yolov5: A lightweight small object detection algorithm based on improved feature fusion mode[J]. Sensors, 2022, 22(15): 5817.
[4] Sun P, Ding X. UAV image detection algorithm based on improved YOLOv5[C]//2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE). IEEE, 2022: 757-760.
[5] Li A, Sun S, Zhang Z, et al. A Multi-Scale Traffic Object Detection Algorithm for Road Scenes Based on Improved YOLOv5[J]. Electronics, 2023, 12(4): 878.
[6] Jia X, Tong Y, Qiao H, et al. Fast and accurate object detector for autonomous driving based on improved YOLOv5[J]. Scientific reports, 2023, 13(1): 1-13.
[7] Li Z, Namiki A, Suzuki S, et al. Application of low-altitude UAV remote sensing image object detection based on improved YOLOv5[J]. Applied Sciences, 2022, 12(16): 8314.
[8] Deng L, Bi L, Li H, et al. Lightweight aerial image object detection algorithm based on improved YOLOv5s[J]. Scientific Reports, 2023, 13(1): 7817.
[9] Wang H, Xu Y, He Y, et al. YOLOv5-Fog: A multiobjective visual detection algorithm for fog driving scenes based on improved YOLOv5[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-12.
[10] Lin Q, Zhang S, Xu S. Construction of Traffic Moving Object Detection System Based on Improved YOLOv5 Algorithm[C]//2023 2nd International Conference on 3D Immersion, Interaction and Multi-sensory Experiences (ICDIIME). IEEE, 2023: 268-272.