Research on pedestrian detection using optimized mask R-CNN algorithm in low-light road environment

K C Lai1, J Zhao1,2, D J Liu2, X N Huang1 and L LWang1

1Department of Mechanical Engineering, Guizhou university, Guiyang, 550025, China
2Department of Transportation, Southeast University, Nanjing, 210000, China

E-mail: zhaoj@gzu.edu.cn

Abstract. Aiming to the performance degradation of the object detection algorithms in low-light environment, the image fusion module (MSRCR-IF) proposed in this paper is introduced into the object detection network based on the object detection algorithm of Mask R-CNN. This proposed fusion algorithm adjusts Region Proposal Network (RPN) and delete instance mask branch to achieve better pedestrian detection performance of algorithm in low-light environment. The experimental results reveal that the algorithm proposed in this paper has better detection performance than other current mainstream algorithms in COCO2017 data set, and the average detection accuracy of 85.05% was achieved under the self-built low-light road environment data set, which was 4.66% higher than before improvement. In order to verify the effectiveness of the improved algorithm, a real car data test was conducted, and the test results showed that the improved method can effectively improve the detection effect of pedestrians in low light conditions.

1. Introduction

Pedestrian detection has always been the focus of computer vision, and the research of this technology can be applied to unmanned driving scenarios, as well as applications including intelligent video surveillance and intelligent robots. In recent years, with the rapid development of deep learning, its powerful feature expression capabilities have made deep learning-based detection algorithms occupy an important position in the field of target detection. At present, the target detection algorithm frameworks that are widely used include R-CNN, Fast R-CNN, Faster R-CNN, SSD, YOLOv3 and Mask R-CNN. Although the pedestrian detection model based on deep learning has been continuously innovated and improved in recent times, experiments have found that the performance of these pedestrian detection algorithms tends to decline in low-light environments [1].

In order to improve driving safety in low-light environments, more and more researchers and scholars have begun to focus on pedestrian detection technology at night and in low-light environments to improve the robustness of pedestrian detection algorithms. Jun Liu [2] et al. proposed a new method for real-time pedestrian detection and tracking in a dynamic environment by integrating RGB and depth data; Liu [3] et al. designed four ConvNet fusion structures based on the Faster RCNN network to fuse color image and thermal image features, and reduced the missed detection rate of KAIST data by 3.5% compared with other network results; Jorg Wagner [4] et al. studied two deep fusion structures are discussed and the potential of deep models for multispectral pedestrian detection is discussed. Studies have shown that pre-trained late fusion structures are significantly better than existing mainstream solutions. Rashed H [5] et al propose a robust and real-time CNN architecture for Moving Object Detection (MOD) under low-light conditions by capturing motion information from both camera and LiDAR sensors.
The above methods mainly have the following three problems: 1. The existing detection network has poor adaptability to different illuminations. 2. The processing process is complicated and time-consuming, and the performance improvement is not obvious when applied to the detection network. 3. These algorithms often rely on additional hardware devices (such as thermal cameras, depth cameras, etc.). Therefore, in response to the above problems, this paper uses Mask R-CNN as the basis without adding additional equipment, and adds an image fusion module (Multi-Scale Retinex with Color Restoration using Image Fusion, MSRCR-IF) to the detection model to highlight Image target information under different lighting, so as to ensure that the target detection network has better detection capabilities for images with better lighting and images with poor lighting. In addition, in order to improve the accuracy and speed of target recognition, the Region Proposal Network (RPN) is improved and the Mask Branch is removed.

2. Mask R-CNN target detection network
In 2017, He K [6] et al. proposed Mask R-CNN, an object instance segmentation network framework. The target detection network is improved on the basis of Faster R-CNN and is one of the best detection algorithms at present. The Mask R-CNN algorithm mainly includes five parts, which are The Feature Extraction Network (Resnet101), Feature Pyramid Network (FPN), Region Proposal Network (RPN), and Region of Interest Align, (RoI Align), and Functional Networks.

The backbone network of Mask R-CNN is formed by the combination of deep convolutional neural network Resnet101 and FPN (Feature Pyramid Network). The FPN structure contains three parts: top-down, bottom-up and horizontal link. (1) The bottom-up connection path can obtain the image feature map (Feature map), which can extract deeper feature semantic information of the image. (2) The top-down connection path is to upsample the more abstract and more semantic high-level feature map extracted by the convolutional neural network. (3) Connect horizontally to fuse the feature map obtained by upsample with the same size feature map obtained from bottom to top. The combination of these three network structures can fuse the features of each level, so that it has strong semantic information and spatial information, and avoids the loss of processing information, which plays a very important role in the feature learning process. Its structure is shown in Figure 1.

![Figure 1. Feature Pyramid Networks (FPN).](image)

The function of the regional submission network RPN is to generate a series of candidate frames with each pixel as the center according to the image features extracted by the convolutional neural network. The size of each frame is determined by the two parameters of Scales and Ratio. Among them, Scales represents the size of the box, and there are 5 sizes; Ratio represents the aspect ratio of the box, and there are 3 types of ratios. The RPN network will also modify the center, width, and height of each candidate frame to obtain a candidate frame with higher accuracy. For the precise positioning of the candidate frame, the Mask R-CNN network is different from the RoI Pooling operation used in the Faster R-CNN network but adopts RoI Align, which can retain the decimal position of the candidate frame and achieve more precise positioning in space. It solves the problem of regional mismatch caused by the two Quantify in RoI Pooling operation, and effectively improves the accuracy of the detection model. Finally, the accurate detection frame and classification information are obtained, and the mask area for target recognition is generated to realize target detection.
3. Mask R-CNN target detection framework improvement

The low-light road pedestrian detection network based on Mask R-CNN proposed in this paper adds an image fusion module to Mask R-CNN, and improves the mid-region submission network and instance segmentation branch module of the detection network. The network structure is represented in Figure 2.

Figure 2. Improved Mask R-CNN detection network framework.

3.1. Image fusion module

The image fusion module is inspired by the articles of Jiang B [7], Ying Z [8] and others. Its purpose is mainly to determine the degree of exposure of the input image, and to help solve the imbalance of exposure and image features in the image. Outstanding issues. The calculation process is shown in Equation 1:

\[ W_1 \times P + W_2 \times P^{M} = R \]

Where \( P \) is the original image, \( P^{M} \) is the image after the MSRCR (Multi-Scale Retinex with Color Restoration) processing, \( R \) is the result image after fusion, \( W_1 \) is the weight of the original image, \( W_2 \) is the weight of the image after MSRCR processing, \( * \) is represent the product, and \( W_1 + W_2 = 1 \). The entire algorithm flow framework is shown in Figure 3.

Figure 3. Image fusion module.

The image fusion module will assign a larger weight value to well-exposed areas and a smaller weight value to under-exposed areas by adjusting the and weight values. In this way, the information retention of the image under different lighting is ensured. The formula for calculating the weight matrix is shown in Equation 2:

\[ W_1 = T^\mu \]

Where \( T \) is scene lighting estimation, \( \mu \) is a parameter that controls the degree of enhancement. And we can get \( T \) through the optimal equation:
\[
\min \|L - T\|_2^2 + \alpha \|M \cdot \nabla T\|_1
\]  
(3)

Where \(\|\|_1\) and \(\|\|_2\) represent the norm of the corresponding calculated value, which \(\alpha\) is a constant; \(\nabla T\) and \(T\) represent the gradient in the horizontal and vertical directions. \(L(x)\) represents the initial value of light estimation, the expression is shown in Equation 4. \(M(x)\) is a weight matrix. Its expression is shown in Equation 5, \(\omega(x)\) is the central area of the pixel \(x\), \(\varepsilon\) is a small constant to avoid the denominator being zero.

\[
L(x) = \max_{c \in \{R,G,B\}} P(x)
\]  
(4)

\[
M(x) = \frac{1}{\sum_{y \in \omega(x)} \nabla L(y) + \varepsilon}
\]  
(5)

By finding the optimal solution, the image fusion module can not only ensure that well-exposed areas are not over-exposed, but also increase the exposure of under-exposed areas of the picture, achieving overall balance of image exposure and retention of important information. Therefore, adding this module before the feature pyramid network FPN can effectively help the feature pyramid network FPN to obtain stable image features and information, so as to improve the robustness of the network and the effect of detecting people in low-light environments.

### 3.2. Improvement of RPN network

At the beginning of the Mask R-CNN target detection algorithm proposed network RPN in the design area, in order to be able to detect more shaped target items, the scale of the candidate frame is set to the following five types: [32×32, 64×64, 128×128, 256×256, 512×512], the candidate box has three aspect ratios: [1:2, 1:1, 2:1]. The algorithm in this paper is mainly used for pedestrian detection in road scenes, and most of the candidate boxes for pedestrians are elongated (including: pedestrians walking on the road, pedestrians riding bicycles, motorcycles, etc.), so this important setting is for pedestrian detection tasks it has obvious shortcomings. Therefore, this paper improves the RPN of the regional recommendation network. By adding a set of vertical candidate boxes to replace horizontal candidate boxes, the candidate regions provided by the final RPN network are more accurate. Figure 4 (a) and (b) show the schematic diagrams of the candidate frame ratio before and after improvement, respectively.

![Figure 4. Schematic diagram before and after the improvement of the size of the candidate box.](image)

### 3.3. Remove mask branch

For the task of target detection, we pay more attention to identifying the category and position information of the target object, and in the Mask R-CNN target detection framework, the mask branch and the classification prediction branch are parallel to each other, and removing the mask branch will not affect the training and prediction results of other layers of the network. Therefore, this article chooses to remove the mask branch in the Mask R-CNN network, so that the convolutional neural network focuses on the extracted feature maps for classification and positioning tasks. In this way, we
can save a lot of training and prediction time. In summary, the improved Mask R-CNN network structure is shown in Figure 2.

4. Data set and network training

4.1. Data set preparation
Because this article is aimed at pedestrian detection in road scenes, it is necessary to collect pedestrian image information in different road scenes. In this training model, a total of 5000 pictures were selected as the data set, of which the training set contained 4000 images and the verification set 1000. Part of the data set selects pedestrian images under different lighting environments from the public COCO2017 data set, and the other part selects from the self-built low-light road environment data set. And by using the labelme image annotation tool to manually label the pictures and generate the training files needed to train the network model.

4.2. Network training
The experimental algorithm in this paper is based on the deep learning network framework TensorFlow, and the code running environment is Python3.5.2. Completed under the experimental configuration of ubuntu16.04 operating system equipped with NVIDIA RTX-2080TI GPU. To speed up the training and prevent overfitting, the specific parameters during training are set as follows: Momentum is 0.9, Weight Decay Coefficient is 0.0001, Batch Size is 32, and Learning Rate is 0.001 and the number of iterations 50,000. At the same time, the Mask R-CNN pre-training weight file h5 is used during network training, and the weight values are allocated to the corresponding category layer. Using a pre-trained weight can avoid re-training a lot of time, and can also obtain a more stable detection effect.

5. Experimental results and analysis
In order to verify the effect of the improved network on pedestrian detection in low-light environments, 500 pieces of data were selected for testing on the COCO2017 data set and the self-built low-light road environment data set. This article uses the following two indicators to evaluate the performance of the target detection network, including average detection accuracy (Average Precision, AP) and detection rate (Detection Rate, DR). The detection accuracy (Precision, P) is defined as shown in equation (6).

\[
P = \frac{TP}{TP + FP} \times 100\%
\]

Where: TP represents the number of targets that the model correctly detects; FP represents the number of targets that the model has falsely detected.

5.1. Experimental verification before and after improvement
In order to verify the effectiveness of the improved regional recommendation network RPN and the deletion of the instance split branch, this section will compare the COCO2017 data set by adjusting the parameters of the regional recommendation network under the trained weight model. The results are shown in Table 1:

| Method   | Number of candidate regions | AP(%) | DR(ms) |
|----------|----------------------------|-------|--------|
| RPN      | 2000                       | 90.47 | 218    |
| Improve RPN | 2000                      | 90.97 | 204    |
| RPN      | 500                        | 87.74 | 195    |
| Improve RPN | 500                       | 88.42 | 188    |
It can be seen from Table 1 above that we adjusted the number of candidate regions. After the improvement, the RPN network and the RPN network before the improvement have improved the accuracy by 0.5% and 0.68% under the same number of candidate regions. At the same time, the improved network detection time is about 10ms higher than before.

5.2. Comparison experiment verification

In order to verify the effectiveness of the image fusion module proposed in this paper, the current mainstream target detection algorithms SSD, YOLOv3 and the improved algorithm before and after this paper are compared for simulation experiments. The test results are shown in Tables 2 and 3:

| Algorithm   | AP (%) | DR (ms) |
|-------------|--------|---------|
| SSD         | 72.53  | 78      |
| YOLOv3      | 86.72  | 53      |
| Mask R-CNN  | 90.47  | 218     |
| Ours        | 91.54  | 204     |

Table 2. Comparison of COCO test set results.

It can be seen from Table 2 that the algorithms before and after improvement have better detection accuracy than the mainstream target detection algorithms such as SSD and YOLOv3 under the COCO data set. At the same time, the average detection accuracy of the improved algorithm is 1.07% higher than that before the improvement, indicating that that the improvement of the algorithm can still guarantee a better detection effect under normal lighting.

| Algorithm   | PA (%) | DR (ms) |
|-------------|--------|---------|
| SSD         | 33.53  | 78      |
| YOLOv3      | 58.25  | 53      |
| Mask R-CNN  | 80.39  | 218     |
| Ours        | 85.05  | 204     |

Table 3. Comparison of low-light road environment test set results.

Comparing the experimental results of several target detection algorithms from Table 3, it can be found that the detection performance of SSD and YOLOv3 algorithms under low-light environment conditions declines rapidly. The average detection accuracy is 33.53% and 58.25%, respectively. The detection effect is average, and leakage is easy to occur. Detection or misdetection. However, Mask R-CNN still performs well in this data set due to its superior algorithm structure, reaching an average detection accuracy of 80.39%. The performance of the methods proposed in this paper are all improved compared with other algorithms, achieving an average detection accuracy of 85.05%, which is 4.66% higher than the algorithm before the improvement. This fully demonstrates that the improved algorithm proposed in this paper has better detection performance and robustness for target detection in low-light environments.

5.3. Visualization of experimental results

In order to show the visual results of the comparative experiment, this paper selected several pictures from the test data set for comparative analysis, and uniformly adjusted the confidence threshold of target detection to 0.5. The comparison of the results of each algorithm is shown in Figure 5 below. From the experimental comparison in Figure 5, we can clearly see that the SSD algorithm and the YOLOv3 algorithm are prone to missed detection in a low-light road environment, and the detection accuracy is relatively low. Comparing Figure 5(d) and (e), it can be found that the improved algorithm can not only detect some smaller pedestrian targets, but also has higher and more accurate detection accuracy than before.
6. Real vehicle experimental test
In order to test the accuracy and effectiveness of the algorithm in this paper, a certain brand of electric control car is used as an experimental platform to conduct data collection and experiment on the campus of Guizhou University, as shown in Figure 6. The experimental platform is mainly equipped with sensors such as lidar Velodyne, monocular camera, and GPS positioning module. The speed of the test vehicle is controlled at about 25km/h. The test environment is cloudy, dark, and the driving environment is simple.

![Figure 5](image)

**Figure 5.** Comparison results of pedestrian detection experiments of various algorithms in low light environment.

![Figure 6](image)

(a) Electric control platform vehicle (b) Electric control equipment

**Figure 6.** Experimental platform car.
Figure 7 shows the test results of the real scene. The left side is the detection result of the original Mask R-CNN algorithm, and the right side is the improved detection result. The detection speed of each frame is about 0.2s. From Fig. 7(a)(b), it can be clearly seen that the improved algorithm has better detection effect for far away pedestrian targets, and the detection accuracy is also improved, while in Fig. 7(c) It can be seen that the algorithms before and after the improvement both detect the rubber road cone on the left side of the road as pedestrians, but the improved false detection is significantly lower than the original algorithm. Therefore, it is verified that the improved target detection algorithm has better detection effect and robust performance for pedestrian detection on low-light roads.

![Figure 7. Comparison results of real car experiment data.](image)

7. Conclusion
Aiming at the problem of pedestrian detection algorithm performance degradation in low-light environments, based on Mask R-CNN, an image fusion module is added to the network, while the RPN network is improved for the characteristics of pedestrians and the instance segmentation branch is deleted to finally achieve the overall improvement of the pedestrian detection capability of the target network. The experimental results show that the improved method proposed in this paper is better than before the improvement under the COCO2017 data set and the low-light road environment data set, and the effectiveness of the improved algorithm is verified in the actual vehicle experimental data test.

Acknowledgments
The authors acknowledge the financial support received from National Nature Science Foundation of China (Grant No.51965008), Major Science and Technology projects of China of Guizhou ZNWLQC [2019]3012.

References
[1] Kruthiventi S S S, Sahay P, and Biswal R 2017 Low-light pedestrian detection from RGB images using multi-modal knowledge distillation. *Ieee Image Proc.* 978-1-5090-2175-8
[2] Liu J, Liu Y, Cui Y, and Chen Y Q 2014 Real-time human detection and tracking in complex environments using single RGBD camera *Ieee Image Proc.* 978-1-4799-2341-0
[3] Liu J, Zhang S, Wang S, and Metaxas D 2016 Multispectral Deep Neural Networks for Pedestrian Detection *British Machine Vision Conf.* 46 206-17
[4] Wagner J, Fischer V, Herman M, and Behnke S 2016 Multispectral pedestrian detection using deep fusion convolutional neural networks *24th Esann Comm. Com. Inf. Sc.* 94 69-77
[5] Rashed H, Ramzy M, Vaquero V, Sallab A E and Yogamani S 2019 FuseMODNet: real-time camera and LiDAR based moving object detection for robust low-light autonomous driving IEEE/CVF ICCVW 978-1-7281-5023-9

[6] Kaiming H, Georgia G, Piotr D, and Ross G 2017 Mask r-cnn IEEE PP 1-1

[7] Jiang B, Woodell G A, and Jobson D J 2015 Novel multi-scale retinex with color restoration on graphics processing unit J. Real-Time Image Pr.10(2) 239-53

[8] Ying Z, Li G, Ren Y, Wang R, and Wang W 2017 A new low-light image enhancement algorithm using camera response model IEEE ICCVW 978-1-5386-1034-3