Vehicle Detection in Low Illumination Based on Attention Mechanism and RetinexNet

Rongdan Qu (18553567366qrd@gmail.com)
Shandong University of Science and Technology
Xingke Li (18763905001a@gmail.com)
Shandong University of Science and Technology

Case Report

Keywords: YOLOv5, Attention mechanism, RetinexNet, Target detection, DIoU

DOI: https://doi.org/10.21203/rs.3.rs-1727102/v3

Additional Declarations: No competing interests reported.

MDPI

Article

Vehicle Detection in Low Illumination Based on Attention Mechanism and RetinexNet

Rongdan Qu *,Xingke Li

Intelligent Equipment College, Shandong University of Science and Technology, Taian 271019, China
* Correspondence: 1855356366qrd@gmail.com(R.Q.);18763905001a@gmail.com(X.L.)

Abstract: This study proposes a method to detect vehicles by enhancing the YOLOv5 network structure and incorporating RetinexNet to address the problem of limited detection capabilities of target identification algorithms in low illumination conditions, such as at night time. CBAM attention module is implemented in the network's Neck detection layer to extract the vehicle's primary features, reduce the extraction of unused features, and improve the vehicle's detection performance. The DIoU is introduced as the loss function of the model to solve the imprecise location of the prediction box and speed up the convergence of the model. RetinexNet is applied to enhance the detection capabilities of low-illumination images by enhancing and denoising them. The experimental results indicate that the enhanced model detects vehicles with 90% accuracy in low-light conditions and has a strong detection performance overall.

Keywords: YOLOv5; Attention mechanism; RetinexNet; Target detection; DIoU)

1. Introduction

Vehicle detection technology is an important research topic in modern intelligent transportation systems for computer vision technology. With the development of computer vision, digital image processing technology, and intelligent transportation technology, automatic vehicle recognition technology is increasingly being valued. However, even though existing systems have high requirements for images, problems such as complex backgrounds, uneven lighting, low resolution, and dirty and old vehicles are always encountered, especially in low-light or bad weather conditions, which leads to difficult vehicle model recognition and low image quality and license plate recognition rate. Therefore, how to improve the recognition rate of vehicles under low-light conditions to cope with complex and changing application environments is an important issue in the research of vehicle detection and recognition systems.

In some areas, infrared cameras are used to detect vehicles. Although this method can effectively obtain night-time environments, it is prone to interference from vehicle lights, and infrared cameras are usually expensive. Therefore, for the purpose of successful vehicle detection, some scholars have tried to use certain algorithms to analyze the images obtained by ordinary CCD and CMOS cameras. RITA C [1] et al. used the position of the headlights to determine the vehicle, but this method is easily affected by external light sources such as ground reflections, with low accuracy. O'Malley R [2,3] et al. tried to obtain vehicle position information by adjusting the high red threshold for the car lamp adaptively based on the changes in its light. The above methods have, to a certain extent and from different prospectives, improved the detection accuracy by pairing the headlights and taillights [4,5], but the detection errors are still large, and the accuracy is still difficult to be guaranteed.

To solve the shortcomings of traditional nighttime vehicle detection, the low-light images collected can be enhanced first to highlight the feature information in the image and improve the accuracy of subsequent target detection. At present, image enhancement

Citation: Rongdan, Q. ;Xingke, L. Vehicle Detection in Low Illumination Based on Attention Mechanism and RetinexNet. *Journal Not Specified* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

algorithms are mainly divided into traditional methods and deep learning methods. Currently, there are many traditional enhancement methods. Liu [6] et al. used LogAbout for lighting compensation, leading to higher correct detection rate. This idea can quickly and effectively solve the problem of lighting compensation. Pang-Ting Huang [7] et al. proposed a dynamic illumination object (DIO) detector to overcome the uncertainty in lighting conditions. B. Froba [8] et al. introduced light-invariant local structural features for target detection. R. E. Sequeira [9] greatly improved the accuracy through the scattering noise process. Nevertheless, the needs of complex scenes or subsequent target detection cannot be met in most cases. For example, the histogram equalization method (Figure 1(a)) can be used to change the gray-level histogram of the original image from a relatively concentrated gray-level range to a uniform distribution across the entire gray-level range, resulting in a decrease in the gray level of the transformed image and the loss of some details, which is not conducive to subsequent target detection. The gray world algorithm (Figure 1(b)) is prone to failure when the scene color is not rich, especially when large single-colored objects appear. Finally, the automatic white balance method (Figure 1(c)) is not suitable for shooting single-colored objects and is prone to color deviation.

(a) (b) (c)

With the emergence of deep learning, low-light image enhancement algorithms based on convolutional neural networks have shown significant improvements over traditional methods. I. S. Kim [10] proposed a method that combines an image quality enhancement network with an object detection network to obtain high-quality images from ordinary surveillance cameras used in different locations, thus improving safety in low-light areas at night. Y. Qu [11] et al. proposed an image transformation optimization network based on a cyclic generative adversarial network, which significantly improves detection accuracy and increases the number of detected targets in low-light environments. W. Wang [12] et al. used block matching and 3D filtering methods for image denoising and sharpening. H. Kuang [13] et al. proposed a biologically-inspired image enhancement method and weighted feature fusion technology, in which the classifier is trained with a support vector machine and the image enhancement method has a good effect on vehicle detection. Although these detection methods have improved detection accuracy in low-light environments from different perspectives, it is still difficult to maintain good performance in complex low-light environments.

RetinexNet [14] deep network provides a good method for low-light enhancement, but there are still some deficiencies. The enhanced image is prone to color distortion and image blur, which is not conducive to subsequent detection. Therefore, by converting the RGB color space to the HSV color space before image enhancement, the color components are separated from the brightness and saturation components to ensure that the color components remain unchanged while adjusting the brightness and saturation components. This avoids color distortion caused by adjusting the three color channels separately in the RGB color space. In response to the blurring of the enhanced image caused by the denoising operation of the Enhance-Net in RetinexNet, Laplace is introduced for sharpening to enhance the edge information of the image and improve the recognition rate of vehicle detection. After image enhancement, an improved YOLOv5 network is used for vehicle detection. Specifically, inspired by the way the human brain processes visual information, an attention mechanism is added to the original network model to improve the accuracy of object detection. Hu J [15] et al. proposed introducing SENet to model the correlation

between feature channels, so as to strengthen important features to improve accuracy. However, this method only focuses on channel factors and does not fully utilize global contextual information. On the other hand, CBAM [16] uses attention mechanisms in both channel and spatial dimensions, achieving better results than SENet that only focuses on channel attention mechanisms. For the loss function used in model training, directly regressing the Euclidean distance between the center points of the two boxes accelerates convergence and obtains relatively fast speeds.

In summary, this work has made the following significant contributions:

1) YOLOv5, a highly popular detection algorithm, is used for optimization through unique feature techniques.

2) Furthermore, the detection of vehicles under low-light conditions is achieved, and experimental studies on several solutions have been done to determine their effectiveness, which is highly relevant to real-world scenarios.

2. YOLOv5 Model

YOLOv5-v5.0 is the YOLO family of algorithms, proposed by Redmon et al. Previous versions include YOLOv1 [17], YOLOv2 [18], YOLOv3[19], and YOLOv4 [20]. It has four models, YOLOv5s, YOLOv5m, YOLOv5l ,and YOLOv5x. In this article, the YOLOv5s model will be used for experimental research, and the network structure is shown in Figure 1, which is relatively easy to train due to its minimum network depth and minimum feature map width. YOLOv5s consists of the Input, Backbone, Neck and Prediction. The input side of the network is enhanced with Mosaic to process the data, and YOLOv5s is able to recalculate and adaptively scale the anchor box compared to the previous version of YOLO where the anchor box value could only be manually modified. Backbone mainly consists of the focus structure, the CSP1_x structure and the spatial pyramid pooling (SPP)[21].Focus performs a replication operation followed by slicing to obtain a binary downsampling feature map without losing feature information. cSP1_x has a residual structure[22] to optimize the gradient information of the network and avoid the disappearance of gradients due to network deepening. SPP converts input images of different sizes into fixed-size images. Neck uses the FPN3[23] and PANet[24] architecture for feature fusion, which enables

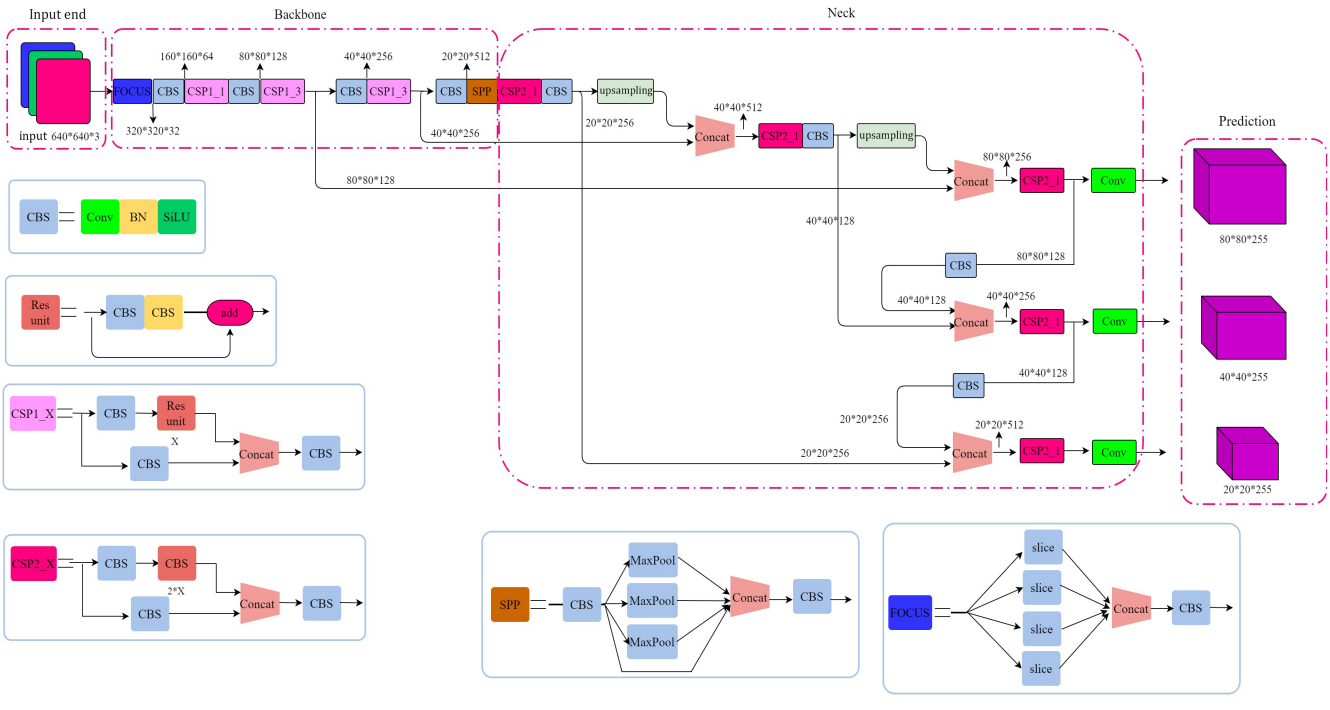

Figure 1. YOLOv5s network architecture diagram.

effective multi-scale feature fusion. The CSP2_x structure is used, replacing the residual structure with a CBL structure based on CSP1_x. The CBL consists of a convolutional layer, BN normalization and a ReLU activation layer, thus enhancing the fusion of the extracted features. Non-maximum suppression processing in Prediction uses CIoU_loss[25] as a loss function for the B-box, eliminating redundant bounding boxes.

3. Attention Module

Adding the attention mechanism of CBAM to YOLOv5s' network model to improve the accuracy of target detection is mainly inspired by the way the human brain processes visual information.

The Convolutional Attention Module (CBAM) is a simple, lightweight and effective attention module for feed-forward convolutional neural networks that improve on the problem that the attention generated by SENet[26] on the feature map channel can only focus on the feedback capability of some layers on the channel layer by applying attention in both the channel and spatial dimensions inferentially and multiplying the generated attention map with the input feature image to be used for adaptive feature refinement, significantly improving the network model's ability to extract image features while only increasing the computational effort by a negligible amount. The CBAM module can be integrated into most of the mainstream networks at this stage and can be trained end-to-end with the underlying convolutional neural network, therefore, this article chooses to integrate this module into the YOLOv5 network to highlight the main features and reduce the extraction of unnecessary features, so as to effectively improve the detection accuracy of the network. The structure of the CBAM module is shown in Figure 2.

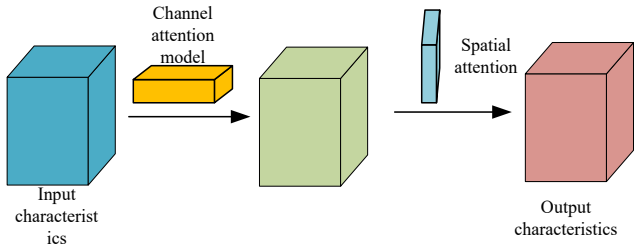

Figure 2. CBAM structure chart.

The specific process is to take a given feature map F $\in \mathbb{R}^{C\times H\times W}$and first compress the feature map under maximum pooling and average pooling through a 1-dimensional channel attention mechanism $M_c \in \mathbb{R}^{C\times 1\times 1}$, thus generating two different spatial information describing the feature map average pooling feature $F_{avg} \in \mathbb{R}^{C\times 1\times 1}$and maximum pooling feature $F_{\max} \in \mathbb{R}^{C\times 1\times 1}$ .It is then passed into a shared network consisting of a hidden layer and a multilayer perceptron (MLP), with the hidden layer set to an activation size of $\mathbb{R}^{C/r\times 1\times 1}$, where r is the reduction ratio to reduce the parameter overhead, to obtain two feature vectors. Finally, they are accumulated and operated by the sigmoid activation function to obtain the channel weights $M_c$which are then multiplied with each pixel of the given feature map F to complete the adaptive feature refinement and obtain the new feature $mapF_1$. Figure 3 represents the structure of the channel attention module CAM, and the expression of the channel attention mechanism is :

$$\begin{aligned} M_c(F) &= \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{\mathrm{avg}}\right)\right) + W_1(W_0(F^c_{\max}))\right) \end{aligned} \tag{1}$$

where σ the S-shaped function,$W_0 \in \mathbb{R}^{C/r\times C}MLP$ is the multilayer perceptron, AvgPool is the average pooling, and MaxPool is the maximum pooling.

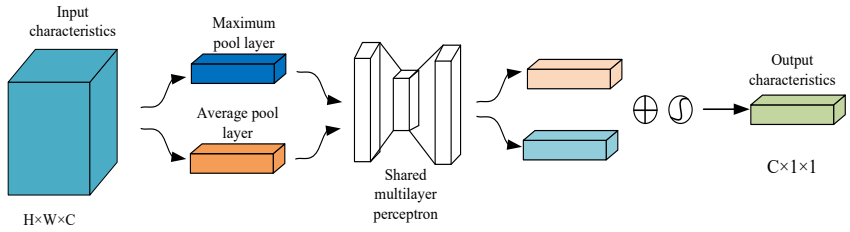

Figure 3. Channel Attention Module CAM structure.

The new feature map $F_1 \in \mathbb{R}^{C\times H\times W}$is then passed through the spatial attention mechanism, which first performs global maximum pooling and global average pooling operations along the channel axis, splices the two obtained feature maps into a complete feature and then convolves it through a 7×7 convolution kernel to reduce its dimension to 1 channel, and finally normalizes it by the sigmoid activation function and multiplies it with the input in the channel to generate The final spatial attention feature $M_s \in \mathbb{R}^{1\times H\times W}$. Figure 4 represents the structure of the spatial attention module SAM, with the expression for the spatial attention mechanism as :

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7\times7}([\text{AvgPool}(F)\,\text{MaxPool}(F)])\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F^s_{\text{avg}}; F^s_{\text{max}}\right]\right)\right) \end{aligned} \tag{2}$$

where σ is the sigmoid activation function and $f^{7\times7}$ is the convolution operation through a convolution kernel of size 7×7.

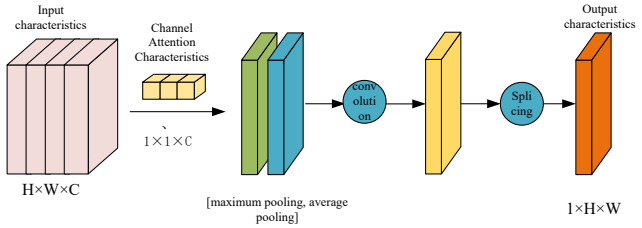

Figure 4. Structure of the spatial attention module SAM.

4. RetinexNet Model

Image processing algorithms for detection in low-illumination environments are mainly divided into traditional methods and deep learning methods. Traditional methods mainly make use of the more significant lighting features in the current environment, such as image enhancement using infrared light and visible light image fusion[27],but their details such as the color of the image cannot be well described; Retinex is an effective and feasible illumination image enhancement, but the majority of Retinex-based methods suffer from model capacity limitations due to artificial constraints and parameters when applied to various environments. Since convolutional neural networks have been widely used in image processing, noise and color distortion in images have been effectively improved, and the RetinexNet deep network provides a good method for low-light enhancement.

RetinexNet is a data-driven Retinex decomposition method that consists of two network structures Decom-Net and Enhance-Net. decomposition net Decom-Net decomposes the acquired image into reflection-independent and structure-aware balanced illumination, consisting of a low-light/normal-light image for the same reflectance and a smooth The two constraints of the illumination map are learned.

The network structure is shown in Figure 5, which takes a low light image S_{low} and a normal light image S_{normal} as input, and predicts the reflectance R_{low} and illuminance I_{low} for S_{low} and S_{normal} respectively.

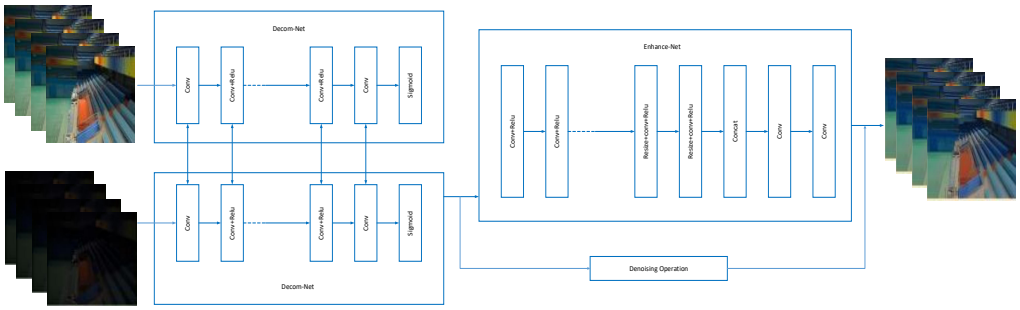

Figure 5. RetinexNet Network Framework.

During training, a pair of low-light normal-light images is acquired, and the decomposition of the low-light and its corresponding normal-light images is learned under the condition that the low-light and normal-light images share the same weights. Both the image components of the low-light image and the normal-illumination image can be used to reconstruct the image S-component, and the decomposition network structure is shown in Table 1.A 3×3 convolutional layer is first used to extract the features, then several 3×3 convolutional layers with ReLu as the activation function are used to map the RGB image The reflectance and illumination are then mapped from the feature space by a 3×3 convolutional layer ,and constrained to [0,1] by a Sigmoid layer.

Table 1. S-component Decomposition Network Structure

Input	Operate	Convolution kernel	Output channel	Step	Output
RGB	rgb to hsv	–	–	–	H,S,V
S	conv	3	64	1	feats0
feats0	conv & ReLU	3	64	1	feats1
feats1	conv & ReLU	3	64	1	feats2
feats2	conv & ReLU	3	64	1	feats3
feats3	conv & ReLU	3	64	1	feats4
feats4	conv & ReLU	3	64	1	feats5
feats5	conv & sigmoid	3	64	1	R,I

In turn,the loss $\mathcal{L}$ is given by :

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{ir}\mathcal{L}_{ir} + \lambda_{is}\mathcal{L}_{is} \tag{3}$$

where $\mathcal{L}_{\text{recon}}$ denotes reconstruction loss, $\mathcal{L}_{\text{ir}}$ denotes constant loss, $\mathcal{L}_{\text{is}}$ denotes illumination smoothness loss, and $\mathcal{L}_{\text{ir}}$ and $\mathcal{L}_{\text{is}}$ denote the coefficients that balance reconstruction smoothness and illumination smoothness consistency.Reconstruction of loss $\mathcal{L}_{\text{recon}}$ as :

$$\mathcal{L}_{\text{recon}} = \sum_{i=\text{low,normal}} \sum_{j=\text{low,normal}} \lambda_{ij} \left\| R_i \circ I_j - S_j \right\|_1 \tag{4}$$

The reflectivity consistency loss $\mathcal{L}_{ir}$ is :

$$\mathcal{L}_{ir} = \| R_{\text{low}} - R_{\text{normal}} \|_1 \tag{5}$$

In order to improve the perception of image structure, Total Variation minimization is often used as a gradient minimization method for the whole image to be used as a smoothing prior to image recovery, but for regions with strong structure or large differences in image brightness, the illumination map gradient has been uniformly decreasing, so the direct use of this method is not reliable. So the original total variance minimization method is weighted to give a final illumination smoothness loss$\mathcal{L}_{is}$ is :

$$\mathcal{L}_{is} = \sum_{i=l} \left\| \nabla I_i \circ \exp\left(-\lambda_g \nabla R_i\right) \right\| \tag{6}$$

where ∇ denotes the horizontal and vertical gradients and λ_g denotes the equilibrium structural strength factor.

Enhance-Net is used to adjust the illumination map to maintain regional consistency and to crop the layout distribution through multi-scale cascades while introducing reflectance denoising to eliminate noise in dark areas or noise that may be amplified during enhancement. The encoder-decoder structure introduces a multi-scale cascade to adjust the illumination from a hierarchical perspective. The structure enables successive downsampling of the input image to small scales, with the downsampling block consisting of a convolutional layer with a step size of 2 and a ReLU activation layer, in order to obtain a large-scale perspective view of the illumination distribution of the photograph and to improve the adaptive nature of the network. Up-sampling is performed to reconstruct the local light distribution using large-scale information.

By introducing multiscale connectivity and thus global illumination consistency, when there are M upsampled blocks with resizable convolutional layers (consisting of a nearest neighbor difference operation, a convolutional layer with a step size of 1 and a Relu activation layer), each block will extract a C-channel feature map and then connect these different scale features to the $C \times M$ channel features by resizing them to the final scale using the nearest neighbor difference method The $C \times M$ channel feature map is then connected to the $C \times M$ channel feature map by the nearest neighbor difference method. This is then reduced to C channels by 1×1 a convolutional layer, and finally a 3×3 convolutional layer to complete the reconstruction of the light map. The loss function $\mathcal{L}$ of this network is :

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{is} \tag{7}$$

where $\mathcal{L}_{\text{recon}}$ denotes reconstruction loss, i.e.

$$\mathcal{L}_{\text{recon}} = \left\| R_{\text{low}} \circ \hat{I} - S_{\text{nprmal}} \right\|_1 \tag{8}$$

The gradient map of R_{low} is weighted as $\hat{I}$.

5. Loss function optimisation

The loss function is able to calculate the difference between the predicted and real data of the model. The GIoU[28] function was used in the original YOLOv5 network to calculate the objectness score, which solves the problem that when the real box in IoU does not completely overlap with the position of the predicted box, it will lead to a loss value of 1 for both, thus The GIoU and its loss calculation formula is :

$$\text{GIoU}(B, B_{gt}) = \text{IoU}(B, B_{gt}) - \frac{|C - (B \cup B_{gt})|}{|C|} \tag{9}$$

$$\begin{aligned} L_{GIoU}(B, B_{gt}) &= 1 - \text{GIoU}(B, B_{gt}) \\ &= 1 - \text{IoU}(B, B_{gt}) + \frac{|C - (B \cup B_{gt})|}{|C|} \end{aligned} \tag{10}$$

where B is the prediction frame, B_{gt} denotes the ground truth and IoU is calculated as :

$$IoU = \frac{|B \cap B_{gt}|}{|B \cup B_{gt}|} \tag{11}$$

C in Eqs. (9) and (10) is the minimum external rectangular box, which is calculated as :

$$C = \left(x_2^C - x_1^C\right) \times \left(y_2^C - y_1^C\right) \tag{12}$$

Where x_1^p, x_2^p, y_1^p, y_2^p are :

$$\begin{aligned} x_1^C &= \min\left(x_1^B, x_1^{B_{gt}}\right), x_2^C = \max\left(x_2^B, x_2^{B_{gt}}\right) \\ y_1^C &= \min\left(y_1^B, y_1^{B_{gt}}\right), y_2^C = \max\left(y_2^B, y_2^{B_{gt}}\right) \end{aligned} \tag{13}$$

When the real frame contains the prediction frame, while the size of the prediction frame and the real frame is fixed, IoU cannot calculate the area ratio of the prediction frame to the real frame well, the reason for this is that the prediction frame does not matter in which position of the real frame, as shown in Figure 6. The values calculated by Eqs. (9) and (11) are unchanged, and the relative positions of the two cannot be determined, resulting in inaccurate target positioning and affecting the convergence effect and accuracy of the model.

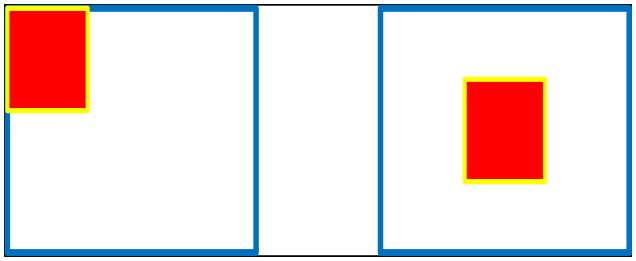

Figure 6. Schematic diagram of the location of the true and predicted boxes.

Based on the above problem, the DIoU[29] loss function algorithm is introduced, which differs from GIoU in that it uses the magnitude of the ratio of the square of the distance between the predicted frame and the centroid of the real frame to the square of the diagonal length of the smallest external rectangular frame C of both as part of the loss value calculation. DIoU with its loss calculation formula is :

$$DIoU\left(B, B_{gt}\right) = \mathrm{IoU}\left(B, B_{gt}\right) - \frac{\rho^2\left(B, B_{gt}\right)}{C_1^2} \tag{14}$$

$$\begin{aligned} L_{DIoU}\left(B, B_{gt}\right) &= 1 - D\,\mathrm{IoU}\left(B, B_{gt}\right) \\ &= 1 - \mathrm{IoU}\left(B, B_{gt}\right) + \frac{\rho^2\left(B, B_{gt}\right)}{C_1^2} \end{aligned} \tag{15}$$

where ρ is the Euclidean distance between the centre point of the prediction frame and the centre point of the detection frame, and C_1 is the slope distance of the smallest external rectangular frame of the two frames. The calculation equations are respectively as follows :

$$\rho^2\left(B, B_{gt}\right) = \left(x_1^p - x_2^p\right)^2 + \left(y_1^p - y_2^p\right)^2 \tag{16}$$

$$C_1^2 = \left(x_2^c - x_1^c\right)^2 + \left(y_2^c - y_1^c\right)^2 \tag{17}$$

Where x_1^p, x_2^p, y_1^p and y_2^p are :

$$\begin{aligned} x_1^p = x_2^B - x_1^B, y_1^p = y_2^B - y_1^B \\ x_2^p = x_2^{B_{gt}} - x_1^{B_{gt}}, y_2^p = y_2^{B_{gt}} - y_1^{B_{gt}} \end{aligned} \tag{18}$$

DIoU enables the centroid of the prediction frame to move closer and closer to the centroid of the real frame, which in turn allows the model to converge faster.

6. Experimental results and analysis

6.1. Experimental environment and experimental data

All experiments in this paper were run under Windows 10 with Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz, NVIDIA GeForce GTX 1650 graphics card, 16GB RAM hardware environment, using Pytorch1.10 deep learning framework, CUDA10.2 parallel computing

architecture and cuDNN deep neural network acceleration library. The experimental dataset contains 2052 images, and the samples are shown in Figure 7. The vehicles in the images were labeled with labelimg, mainly cars, vans, and buses.

Figure 7. Schematic diagram of the sample data set.

The XML format file generated by labelimg was converted into a txt file for training, and the ratio of training set to test set was divided into 8:2. Batch_size was set to 2, the initial learning rate was 0.003, all models were trained with 100 epochs, and some hyperparameters were set as shown in Table 2.

Table 2. Training hyperparameter settings

Parameter	Numerical Value
Initial learning rate	0.0032
Termination rate	0.12
Warm-up learning rounds	2.0
Preheat learning initial bias learning rate	0.05
Epoch	100

6.2. Evaluation indicators

In order to test the performance of the improved model, Precision (P), Recall (R) and mean average precision (mAP) are used as indicators for model performance evaluation in this paper. The specific formulae are.

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \tag{19}$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \tag{20}$$

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{21}$$

In Eq. (19) (20) TP indicates true positive, i.e. correct identification of positive targets. FP indicates false positive, i.e. identification of negative targets as positive. FN indicates false negative, i.e. identification of positive targets as negative. In equation (21) C indicates the number of categories and AP is the area enclosed by the $P - R$ curve, i.e.

$$AP = \int_0^1 P(R)dR \tag{22}$$

6.3. Results of training

Both the improved model and the original YOLOv5 model were trained under the same conditions, and the resulting mAP curves are shown in Figure 8, with the red curve representing the improved model and the blue curve representing the original YOLOv5 model. As can be seen from the curves, both models converge rapidly in the first 20 rounds, and then level off after 50 rounds until the training is completed, with both models well trained and no obvious overfitting or underfitting. By testing the models in the improvement process, the results obtained are shown in Table 3.

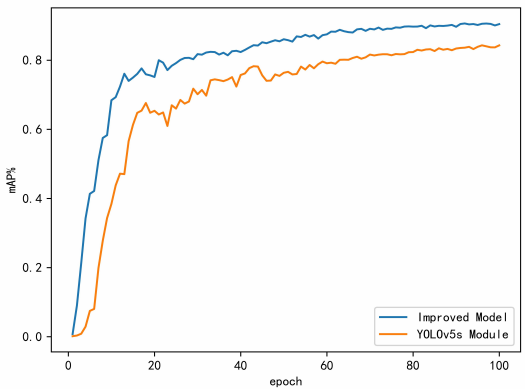

Figure 8. mAP curves for the original YOLOv5 model and the improved model.

The data in the table shows that the improved network has a significant improvement in target recognition accuracy relative to the original network, with an increase of about 6%, and although the improved network has a slightly reduced *FPS*, it basically meets the demand for real-time performance.

Table 3. Training hyperparameter settings.

Model	*mAP*	*FPS*
YOLOv5s	0.8429	24
YOLOv5s+ RetinexNet	0.8948	19
YOLOv5s +RetinexNet+CBAM	0.9004	18
YOLOv5s +RetinexNet+CBAM+DIoU	0.9052	19

Figure 9 shows the comparison curves of the CBAM and SENet modules in the network. The training results show that the *mAP* of the CBAM module is slightly better than that of the SENet module, and the YOLOv5 network incorporating the CBAM module can significantly improve the ability of the network model to extract image features.

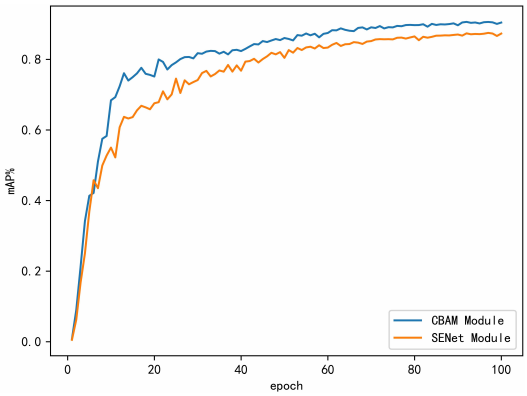

Figure 9. CBAM vs SENet mode comparison graph.

Figure 10 shows the effect of partial vehicle detection, with the original YOLOv5 detection results on the left and the improved model on the right. The original network before improvement cannot effectively detect vehicles in dim photos, while the improved model enhances and denoises the image by RetinexNet to improve the overall brightness of the image, and then enhances the recognition accuracy of vehicle targets by adding the YOLOv5 detection network with the CBAM .attention module, so that it can accurately detect vehicles in the images.

Figure 10. Vehicle inspection result.

7. Conclusion

The detection model based on attention mechanism and RetinexNet is proposed to address the problem of low accuracy of traditional detection algorithms in low illumination environments.

1)The introduction of the CBAM attention module in the YOLOv5 detection network can effectively solve the detection of vehicles in complex environments and optimize the loss function algorithm to improve the detection accuracy of targets.

2)The introduction of Decom-Net and Enhance-Net in RetinexNet to decompose and enhance low-light images improves the ability of the model to detect vehicles at night.

3)The improved algorithm mAP has high accuracy which reaches to 0.9052. The proposed method improves the detection of vehicles in low illumination environments and can meet the needs of a wide range of applications at this stage.

Author Contributions: Conceptualization,R.Q.;Methodology,R.Q.,L.X.; Software, L.X.; Formal Analysis,R.Q.,L.X.; Data Curation,R.Q; Visualization,L.X.; Investigation,,R.Q.; Writing - Original Draft,R.Q.;Writing - Review and Editing,R.Q.,L.X.All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cucchiara, R.; Piccardi, M.; Mello, P. Image analysis and rule-based reasoning for a traffic monitoring system. *IEEE transactions on intelligent transportation systems* **2000**, *1*, 119–130.
2. O'Malley, R.; Jones, E.; Glavin, M. Rear-lamp vehicle detection and tracking in low-exposure color video for night conditions. *IEEE Transactions on Intelligent Transportation Systems* **2010**, *11*, 453–462.
3. O'malley, R.; Glavin, M.; Jones, E. Vision-based detection and tracking of vehicles to the rear with perspective correction in low-light conditions. *IET Intelligent Transport Systems* **2011**, *5*, 1–10.

4. Pham, T.A.; Yoo, M. Nighttime Vehicle Detection and Tracking with Occlusion Handling by Pairing Headlights and Taillights. *Applied Sciences* **2020**, *10*, 3986.
5. Parvin, S.; Rozario, L.J.; Islam, M.E.; et al. Vision-based On-Road Nighttime Vehicle Detection and Tracking Using Taillight and Headlight Features. *Journal of Computer and Communications* **2021**, *9*, 29.
6. Liu, H.; Gao, W.; Miao, J.; Li, J. A Novel Method to Compensate Variety of Illumination in Face Detection. In Proceedings of the Joint Conference on Information Sciences, 2002.
7. Huang, P.T.; Chan, Y.M.; Fu, L.C.; Huang, S.S.; Hsiao, P.Y.; Wu, W.Y.; Lin, C.; Chang, K.C.; Hsu, P.M. Pedestrian detection system in low illumination conditions through Fusion of image and range data. *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)* **2014**, pp. 2253–2254.
8. Fröba, B.; Ernst, A. Face detection with the modified census transform. *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.* **2004**, pp. 91–96.
9. Sequeira, R.E.; Gubner, J.A.; Saleh, B.E.A. Image detection under low-level illumination. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* **1993**, *2 1*, 18–26.
10. Kim, I.S.; Jeong, Y.; Kim, S.H.; Jang, J.S.; Jung, S.K. Deep Learning based Effective Surveillance System for Low-Illumination Environments. *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)* **2019**, pp. 141–143.
11. Qu, Y.; Ou, Y.; Xiong, R. Low Illumination Enhancement For Object Detection In Self-Driving *. *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)* **2019**, pp. 1738–1743.
12. Wang, W.; Peng, Y.; Cao, G.; Guo, X.; Kwok, N. Low-Illumination Image Enhancement for Night-Time UAV Pedestrian Detection. *IEEE Transactions on Industrial Informatics* **2020**, *17*, 5208–5217.
13. Kuang, H.; Zhang, X.S.; Li, Y.; Chan, L.L.H.; Yan, H. Nighttime Vehicle Detection Based on Bio-Inspired Image Enhancement and Weighted Score-Level Feature Fusion. *IEEE Transactions on Intelligent Transportation Systems* **2017**, *18*, 927–936.
14. Wei, C.; Wang, W.; Yang, W.; Liu, J. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560* **2018**.
15. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
16. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
17. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
18. Redmon, J.; Farhadi, A. YOLO9000: better, faster, stronger **2017**. pp. 7263–7271.
19. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* **2018**.
20. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* **2020**.
21. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* **2015**, *37*, 1904–1916.
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
23. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
24. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8759–8768.
25. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2020, Vol. 34, pp. 12993–13000.
26. Yongsheng, Q.; sun guobing.; Yuwu, W.; Shengchun, S. night environment image enhancement based on infrared and visible image fusion. *Journal of Harbin University of Commerce (NATURAL SCIENCE EDITION)* **2021**.
27. Land, E.H.; McCann, J.J. Lightness and retinex theory. *Josa* **1971**, *61*, 1–11.
28. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression **2019**. pp. 658–666.
29. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression **2020**. *34*, 12993–13000.