Automation and Machine Learning (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2024.050111
ISSN 2516-5003 Vol. 5 Num. 1

Pedestrian and vehicle detection method in infrared scene based on improved YOLOv5s model

Jie Yang, Wenzhun Huang*

School of Electroinc Information, Xijing University, Xi'an, China
**Corresponding author: 1051917312@qq.com*

Keywords: Object detection, Infrared scene, YOLOv5s model, Small goals, Obscure the target, SIoU loss function, Decoupling head

Abstract: A infrared pedestrian and vehicle target detection method based on an improved YOLOv5s model is proposed to address the issues of false alarms and missed detections caused by small pedestrian and vehicle targets, occlusion, and low visibility in nighttime driving and complex environments. To address the issue of missed detection of small targets, a small target detection layer is introduced, which increases the three detection layers of the original model to four layers to better handle the detection problem of small-sized targets; The SIoU loss function has been introduced to improve the accuracy of multi-scale object detection, allowing the model to better process different types of targets separately, enhancing the flexibility and generalization ability of the model; At the same time, in response to the contradiction between different tasks of the model, which leads to low pedestrian and vehicle detection accuracy and slow convergence speed, a decoupling head is introduced in the YOLOv5s head to improve the model detection accuracy and positioning speed; Experiments were conducted on the FLIR dataset, and the results showed that the improved YOLOv5s model algorithm increased Precision by 1.6% compared to the original YOLOv5s model algorithm, with Recall and mAP @ 0 5 and mAP @ 0 5: 0 95% has increased by 3.4%, 2.3%, and 5.2%, reaching 90.0%, 88.2%, 93.9%, and 58.9%, respectively.

1. Introduction

With the development of artificial intelligence, the demand for intelligent vehicle functions is increasing, making accurate detection of pedestrians and vehicles one of the key challenges. Traditional target detection methods have some drawbacks due to their inability to adapt to complex environments. Although researchers have made certain achievements in the field of object detection, further research is needed to address the low accuracy of vehicle object detection in complex environments such as insufficient lighting and obstructed vision[1]. It still faces the following problems: when the size of the same target is detected in an infrared environment, the recognition scale of the same target will also change; When there is occlusion between objects, problems such as false detection and missed detection may occur. In addition, due to low image recognition, color loss, and blurred features in infrared scenes, the performance of object detection is not ideal.

To address the above issues, this article proposes an improved model based on YOLOv5s, with a focus on the characteristics of infrared road scenes and corresponding improvements made

accordingly. The loss function of YOLOv5 network has been modified by replacing the original CIoU with SIoU, which enables the network to better learn the location information of the target. Meanwhile, we introduce an understanding coupling head to alleviate the conflict between classification and regression tasks, enabling the model to quickly capture multi-scale features of the target. In addition, we have introduced a small object detection layer to enhance the network's feature extraction ability and detection accuracy for small targets.

2. YOLOv5s model and its improvement strategy

2.1. YOLOv5s Network Architecture

The YOLOv5s network structure is shown in Figure 1, mainly composed of three parts: Backbone, Neck, and Head. Backbone is composed of CBS module, C3 module, and SPPF module, responsible for continuously shrinking feature maps and extracting multi-scale feature information. CBL is a convolutional block composed of Conv, Batch Normalization (BN) layer, and SiLu activation function; The C3 module consists of three convolutional layers and a BottleNeck module[2]; The SPPF module is an improvement on the basis of Spatial Pyramid Pooling (SPP), which makes the network model computation faster. The function of Neck is to obtain relatively shallow features from Backbone, and then fuse them with deep semantic features to obtain rich feature information. Neck's structure includes FPN and PAN. FPN samples feature maps from top to bottom and fuses Backbone feature maps to capture strong semantic features; PAN samples feature maps of different sizes from bottom to top and down, including image semantics and feature information, ensuring accurate prediction of images of different sizes and transmitting low-level position information to higher layers. Head utilizes the feature information passed forward to generate bounding boxes with features on the image, displaying categories and probabilities.

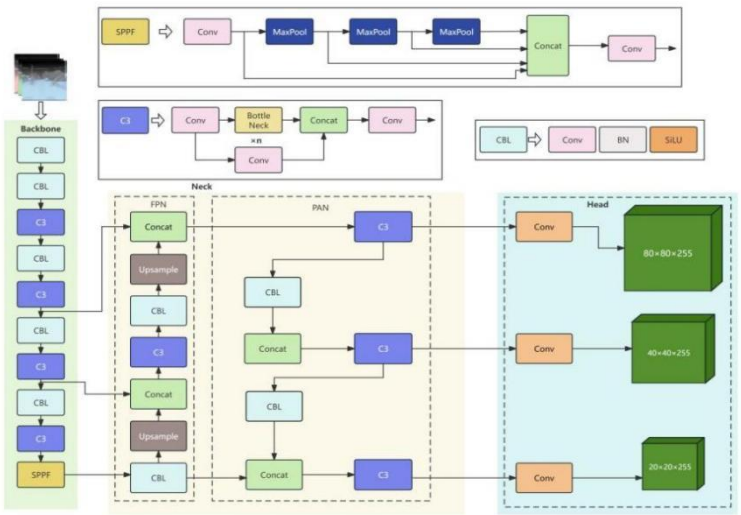


Figure 1: YOLOv5s network structure.

2.2. YOLOv5s Improvement Strategy

2.2.1. Decoupling head

The implementation of the standard detection head in the YOLOv5s model is achieved through the fusion and sharing of regression and classification branches. However, the focus and feature information of the regression and classification tasks are not exactly the same, resulting in spatial

misalignment. The classification task focuses more on using the extracted features to determine the category of the target in the image, thus focusing on the prominent areas in the feature map; The regression task focuses on coordinates and edge information of the target to correct bounding box parameters, which can lead to conflicts between the two tasks. To solve the above problem, this paper introduces a decoupling head in the YOLOv5s head to replace the original coupling head, weaken the contradiction between classification and regression tasks, and accelerate network convergence and improve detection accuracy.As shown in Figure 2. The improved decoupling head processes classification and regression tasks separately through two paths that are non-interference and independent of each other.

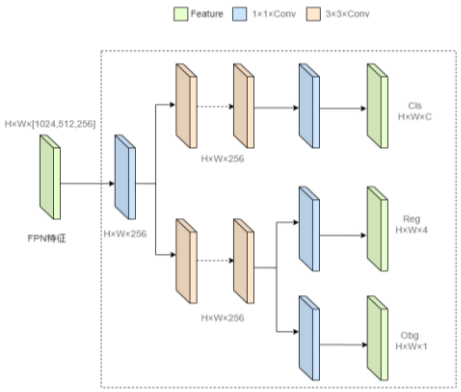


Figure 2: Decoupling head structure diagram.

2.2.2. Small target detection layer

Due to the small size variation of targets in infrared scenes, such as humans and distant cars, it is easy to cause missed detections and decreased detection accuracy. In contrast, the YOLOv5s network model uses multiple downsampling layers to enhance the receptive field, resulting in a large downsampling factor. Therefore, deep feature maps are difficult to learn the feature information of small targets, and there is a significant difference in scale between small targets and the background throughout the entire model learning process, Making large targets and areas with obvious features dominant during the training process may overlook the weak features of small targets in the network model, which affects the detection performance of small targets[3-4]. To solve this problem, we introduced a small object detection layer on the original model structure, increasing the original three-scale detection layer to four layers to further upsample and expand the feature map. This helps to combine shallower positional information with deeper semantic information more closely, improving the network's sensitivity to small targets.

2.2.3. Improvement of loss function

The existing YOLOv5 network model uses CIoU as the localization loss function, taking into account factors such as center point offset, height ratio, and the difference in overlapping area between the real box and the predicted box.However, when the width and height of the predicted box are linear with the width of the true box, it will result in the predicted box's width and height not increasing or decreasing simultaneously, without considering the matching of its predicted box. When the scale of the detection target changes significantly, the detection accuracy of the model will also be affected by the slower convergence speed. Replacing the CIOUs loss function with SIOU can effectively solve this problem.

The SIOU loss function consists of four loss functions, namely Angle Cost, Distance Cost, Shape Cost, and Intersection over Union (IoU) Cost, where Angle Cost considers the vector angle between

the target bounding boxes[5]. The introduction of this cost term aims to consider the rotation angle between bounding boxes, in order to capture more comprehensive information about the target shape and improve the model's robustness to bounding box rotation when identifying target positions. Specifically, the calculation of Angle Cost usually involves predicting the vector angle between the predicted bounding box and the actual bounding box[6], and the function expression is:

$$\Lambda = 1 - 2\sin^2(\arcsin(\frac{C_h}{\sigma}) - \frac{\pi}{4}) \quad (1)$$

Where Λ is the angle loss; C_h is the height difference between the center point of the real box and the predicted box; σ is the distance between the center point of the real box and the predicted box; $C_h = \max(b_{Cy}^{gt}, b_{Cy}) - \min(b_{Cy}^{gt}, b_{Cy})$, $\sigma = \sqrt{(b_{Cx}^{gt} - b_{Cx})^2 + (b_{Cy}^{gt} - b_{Cy})^2}$,The Angle cost function calculates the contribution of angle cost to the loss function, as shown in Figure 3,When $\alpha < \frac{\pi}{4}$, α will be minimized, otherwiseβ will be minimized, $\beta = \frac{\pi}{2} - \alpha$.

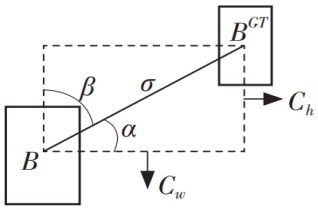


Figure 3: Angle cost function.

The calculation formula for Distance Cost is as follows, where Λ is the angle loss; ρ_t is a normalized indicator for measuring the coordinate deviation between the predicted box center point and the true box center point, where t=x represents the horizontal axis and t=y represents the vertical axis, $\rho_x = (\frac{b_{Cx}^{gt} - b_{Cx}}{C_\omega}), \rho_y = (\frac{b_{Cy}^{gt} - b_{Cy}}{C_h}), \gamma = 2 - \Lambda, C_\omega$ and C_h represent the width and height of the minimum bounding rectangle of the real box and the predicted box, respectively[7].

$$\Delta = \sum_{t=x,y}(1 - e^{-\gamma\rho_t}) \quad (2)$$

The calculation formula for Shape Cost is as follows:$\omega_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})}, \omega_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})}$,W and h are the width and height of the prediction box, respectively.

$$\Omega = \sum_{t=w,h}(1 - e^{-\omega t})^\theta \quad (3)$$

The expression for Intersection over Union (IoU) Cost is:

$$L_{IoUCost} = 1 - IoU \quad (4)$$

The calculation formula for IoU is as follows, as shown in Figure 4.

$$IoU = \left|\frac{B \cap B^{GT}}{B \cap B^{GT}}\right| \quad (5)$$

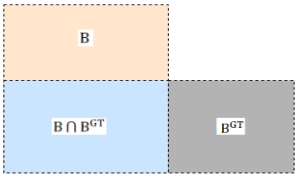


Figure 4: Schematic diagram of IoU calculation.

The final SIoU overall regression loss function is:

$$L_{box} = 1 - IoU + \frac{\Delta+\Omega}{2} \tag{6}$$

By improving the loss function of the YOLOV5s model, replacing the decoupling head, and adding a small object detection layer, the final improved model structure is shown in Figure 5:

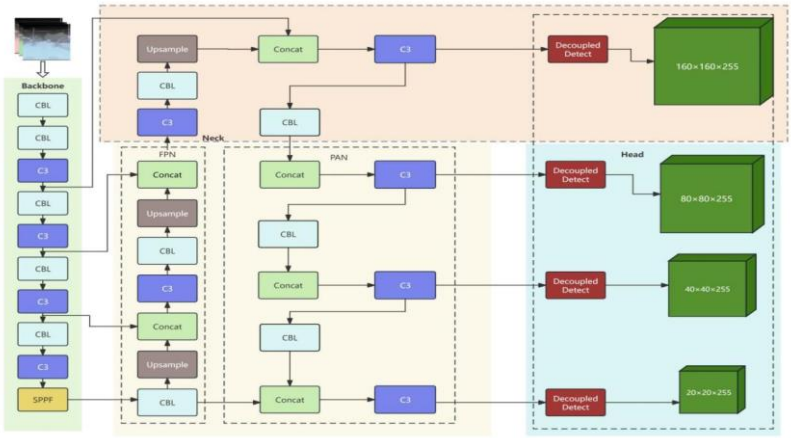


Figure 5: Improved YOLOV5s model

3. Experimental results and analysis

3.1. Experimental platform and parameter configuration

This experiment is a PyTorch 1.10.0 deep learning framework Python 3.8 and CUDA12.0 programming software built under the Win11 operating system. The specific configuration is shown in Table 1.

Table 1: Experimental platform and parameter configuration.

hardware and software platform	Type
operating system	Ubuntu20.04
CPU	AMD EPYC 9754
GPU	GeForce RTX 3090
Graphics memory	24G
Memory	128G
framework	Pytorch 1.10.0
programming environment	Python 3.8.5

3.2. Dataset preparation

The pedestrian and vehicle datasets in the infrared scene of this experiment are from the FLIR thermal infrared dataset launched by FLIR company. The images in this dataset are all from short video clips, and the annotation tool LabelImg is used to annotate the dataset. The annotation categories are "person" and "car", which represent the detected pedestrians and vehicles, respectively. The annotated information is saved in XML format in the same path, Finally, 80% of the annotated images were randomly selected as the training set, 10% as the testing set, and 10% as the validation set[7-10].

3.3. Parameter Setting and Evaluation

In order to optimize the performance of the model, improve generalization ability and training effectiveness, it is necessary to set the hyperparameters in the model configuration file before training the network model, as shown in Table 2. The image size (img) selected in the model is 640 * 640, with 100 iterations (epochs) and a batch size of 8.

Table 2: Parameter Setting and Evaluation.

Parameter	numerical value
batch-size	32
img	640*640
epochs	300

In terms of object detection, the evaluation indicators are usually accuracy P, recall R, and mean average precision (mAP) to evaluate the performance of the algorithm. The calculation formula for evaluation indicators is:

$$P = \frac{T_P}{T_P + F_P} \tag{7}$$

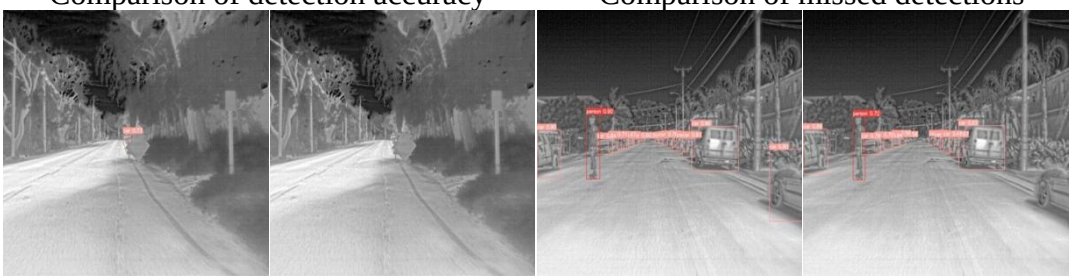
$$R = \frac{T_P}{T_P + F_N} \tag{8}$$

$$m_{AP} = \frac{\sum P}{N} \tag{9}$$

The improved YOLOv5s algorithm was compared with the original algorithm through a test set, as shown in Figure 6. The improved algorithm showed significant improvements in detection accuracy, missed detections, small object detection, and dense scenes compared to the original algorithm.

(a) (b) (a) (b)

Comparison of detection accuracy Comparison of missed detections



(a) (b) (a) (b)

Small target detection Dense scenes

(a) Improved YOLOv5s algorithm (b) Original YOLOv5s algorithm

Figure 6: Comparison before and after improvement.

3.4. Comparative experiment

In order to further verify the detection performance of the improved network model, it was compared with the mainstream network models YOLOv5n, YOLOv5s, YOLOv7, and YOLOv8 in the YOLO series. The experimental results are shown in Table 3. Compared to the YOLOv5s algorithm, the improved algorithm in this article has a 2.3% increase in mAP and a 3.4% increase in recall. For small object detection, such as detecting vehicles at long distances, the accuracy has increased from 87.2% to 89.7%. Compared with commonly used YOLOv7 and YOLOv8, there is a corresponding improvement in recall, accuracy, and mAP indicators.

Table 3: Comparative experiments of various models.

experiment	network model	Precision			Recall	mAP
		all	person	car		
1	YOLOv7	87.4	86.4	88.3	80.9	89.6
2	YOLOv8	88.2	88.4	88.1	81.4	89.9
3	YOLOv5n	87.0	86.5	87.4	79.9	88.3
4	YOLOv5s	88.4	89.6	87.2	84.8	91.6
5	Ours	90.0	90.2	89.7	88.2	93.9

4. Conclusion

This article proposes an infrared pedestrian and vehicle detection method based on an improved YOLOv5s model. By introducing a small object detection layer, the problem of low accuracy in small object detection in infrared scenes is better handled. At the same time, the SIoU loss function is introduced to improve the robustness and generalization ability of the model. A decoupling head is introduced at the head of the original YOLOv5s model to improve detection accuracy and localization speed. Through experimental comparison on the FLIR dataset. This method is superior to existing methods and suitable for the current demand for infrared scene detection. In future work, we will continue to improve the detection accuracy of our method and make lightweight modifications to the model to improve detection speed.

References

[1] Jiaxiang Q, Ziming W, Yimin H.An embedded device-oriented fatigue driving detection method based on a YOLOv5s [J].Neural Computing and Applications, 2023, 36(7):3711-3723.

[2] Yu C, Shin Y.SAR ship detection based on improved YOLOv5 and BiFPN [J].ICT Express, 2024, 10(1):28-33.

[3] Jun X, Renjie G, Yuanpei Z, et al.Carbonate Rock Fracture Identification Method Based on an Improved YOLOv5 Algorithm[J].Pure and Applied Geophysics, 2024, 181(1):189-201.

[4] Kyedong L, Sik K P.Deep Learning Model Analysis of Drone Images for Unauthorized Occupancy Detection of River Site [J].Journal of Coastal Research, 2024, 116(sp1):284-288.

[5] Xing B, Sun M, Ding M, et al.Fish sonar image recognition algorithm based on improved YOLOv5.[J].Mathematical biosciences and engineering : MBE, 2024, 21(1):1321-1341.

[6] L. L M, L. Z, X. C, et al.Research on surface defect detection method of metallurgical saw blade based on YOLOV5[J].Metalurgija, 2024, 63(1):121-124.

[7] Yanru F, Yuliang C, Huijun Y.A detection algorithm based on improved YOLOv5 for coarse-fine variety fruits [J].Journal of Food Measurement and Characterization, 2023, 18(2):1338-1354.

[8] Ziwei W, Yi H, Jianxiang D, et al. YOLOv5-Based Seabed Sediment Recognition Method for Side-Scan Sonar Imagery [J]. Journal of Ocean University of China, 2023, 22(6):1529-1540.

[9] Luxuan B, Bo L, Jue W, et al.Multi-branch stacking remote sensing image target detection based on YOLOv5[J].The Egyptian Journal of Remote Sensing and Space Sciences, 2023, 26(4):999-1008.

[10] Yunfeng J, Zhizhan L, Ruili W.Research on lightweight pedestrian detection based on improved YOLOv5[J]. Mathematical Models in Engineering, 2023, 9(4):178-187.