ISSN: 2664-9640

Detection of pedestrian targets in dark night based on YOLOv5

Zhongyi Liu, Meng Li, Dengfeng Wei *, Shaofa Zhou, Li Lu

School of Yangtze University, Hubei 434000, China.

* Corresponding Author

Abstract

In low-light conditions, night-time vehicle driving is more dangerous than daytime vehicle driving, and night-time monitoring is often not effective, and loopholes often occur. In order to improve the safety of night driving and improve the accuracy of monitoring equipment, more and more researchers have begun to study pedestrian detection at night. The issue faced by pedestrian detection at night is that it is difficult to effectively detect the target under low light, and the pedestrian target is small and the detection rate is low. For the above problems, this paper uses the target detection algorithm of YOLOv5 in pedestrian detection at night, performs Mosaic data enhancement on the input image data set, and uses the Focus structure to slice the image to increase its dimension, and then uses FPN+PAN for feature fusion. , to enhance the detection ability of small targets. Finally, the GLOU loss function is used to accelerate the convergence of the model.

Keywords

YOLOv5; pedestrian detection; dark ngiht.

1. Introduction

Object detection has always been one of the main tasks in the field of computer vision. In recent years, with the rapid development of convolutional neural network research, the performance and accuracy of object detection algorithms have been greatly improved. Target detection algorithms based on convolutional neural networks are generally divided into two categories, one is a two-stage target detection algorithm such as R-CNN[1], Fast R-CNN[2], Faster R-CNN[3], Mask R-CNN[4], R-FCN [5]etc. The steps of the two-stage target algorithm are that the first stage distinguishes foreground and background, generates candidate regions, and then performs feature extraction on the candidate stage in the latter stage for classification and regression Prediction. The second category is the single-stage target detection algorithm, which directly extracts features, and then classifies and regresses the target. Common algorithms such as SSD[6], YOLOv1[7], YOLOv2[8], YOLOv3[9], YOLOv4[10], RetinaNet[11]. Compared with the two-stage network, the single-stage network can directly generate the class probability and position coordinates of the target, that is, the target and the result can be obtained in a single time, and the detection speed is relatively fast. However, the accuracy of the single-stage network for small targets is relatively low. In recent years, researchers have been conducting research on single-stage networks, and proposed the FPN [12] feature pyramid structure to enhance the detection ability of small targets and enhance the contextual semantic information, and then the positioning information is not effectively transmitted. And YOLOv5 adds a self-oriented pyramid structure PAN to complement the FPN and fuse the underlying positioning features and imporve the precision of detection

ISSN: 2664-9640

2. Related work

The work of pedestrian detection has always been a very hot issue. Most models improve the accuracy of pedestrian detection mainly by using some excellent network models to extract features or using multi-scale fusion methods to improve the detection accuracy of the network.

Better featrue extraction network: Excellent feature extraction network, such as Fast-RCNN uses VGG[13] structure to extract features. The SSD network also uses the VGG network to extract the features of the target. Retinanet uses the Resnet with residual structure for network feature extraction. These excellent feature extraction networks can effectively extract features. YOLOv5 uses Darknet-53 as the feature extraction network, which is faster and more effective than the network described above.

Multi-scale feature fushion: The FPN structure can construct multi-scale semantic features by combining low-resolution and high-resolution features through Top-down and Horizontal connection strucute. Through this multi-scale strong semantic feature, the accuracy of target detection can be improved, especially for small targets. YOLOv5 adds a PAN network structure to the basic structure of FPN. The PAN network structure does not complement the previous FPN structure, and can further integrate low-resolution features with high-resolution features to improve the ability of target detection in target positioning. Thereby improving the detection efficiency.

3. YOLOv5 Algorithm

3.1. The source of the YOLOv5 algorithm

The YOLOv5 algorithm is developed on the basis of YOLOv3 and YOLOVv4. YOLOv5 is superior to the existing YOLOv3 and YOLOv4 algorithms in terms of accuracy and detection speed.

3.2. The network strucure of YOLOv5

The network structure of YOLOv5 mainly includes three modules, namely Backbone, Neck and Prediction. Among them, BockBone includes Focus structure, SPP structure, and Neck includes FPN and PAN structure. The structure is shown in Figure 1.

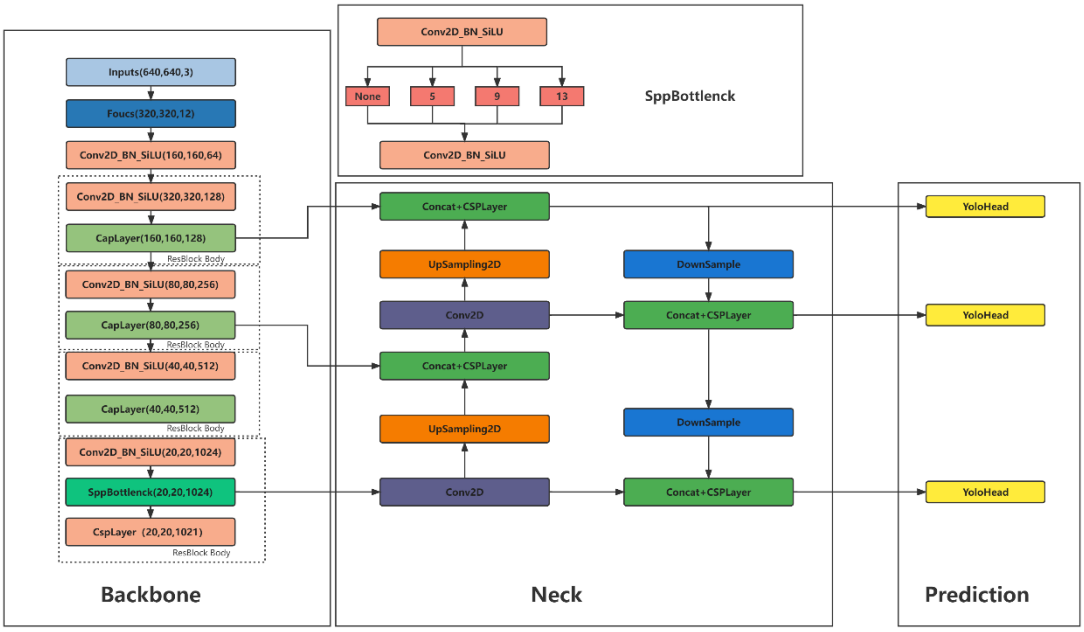


Figure1: YOLOv5 net structure

ISSN: 2664-9640

3.3. Focus

In YOLOv5, the image will be sliced before entering the Backbone. As shown in Figure 2a, the image will be divided into 4 blocks, and then spliced on the channel, each pixel in an image gets a value, Similar to the MAXPOOL operation, so that we can get four pictures, the four pictures are complementary, and no information is lost. The width and height information of the image is concentrated to the channel dimension, and the input channel is expanded by 4 times. That is, the stitched picture is equivalent to the original RGB three-channel turned into 12 channels. The specific slicing operation process is shown in Figure 2b.

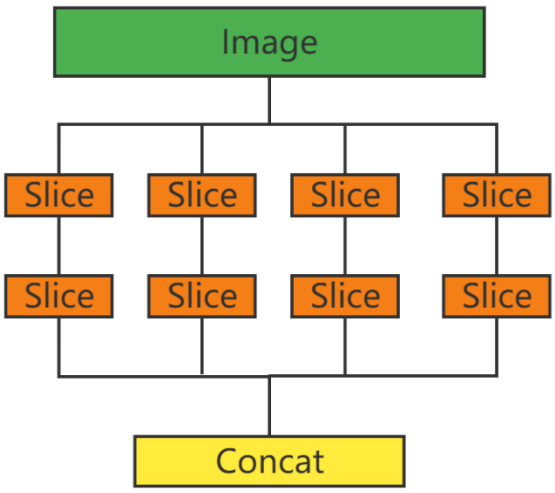


Figure2a: Focus structure

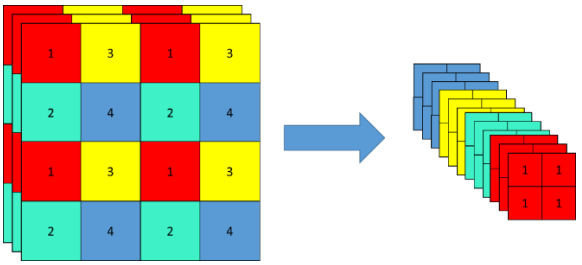


Figure2b: Focus structure

3.4. Spp

The main function of SPP is to solve the problem of the size of the input image. Previously, the target detection network such as R-CNN requires the input of pictures of fixed size. The trimming of these pictures will be clipped or deformed or scaled. Entering the network, this leads to the loss and deformation of the information of the picture to a certain extent, which limits the accuracy of the recognition. In the actual network implementation process, our convolutional layer does not need to input a fixed-size image, and can generate a feature map of any size, but we need a fixed-size input for the fully connected layer. Therefore, the SPP (Spatial Pyramid Pooling layer) structure can solve this problem. SPP inputs different receptive field features through three different kernel sizes, and then concatenates them on the channel to generate a fixed-size output. The specific SPP operation is shown in Figure 3

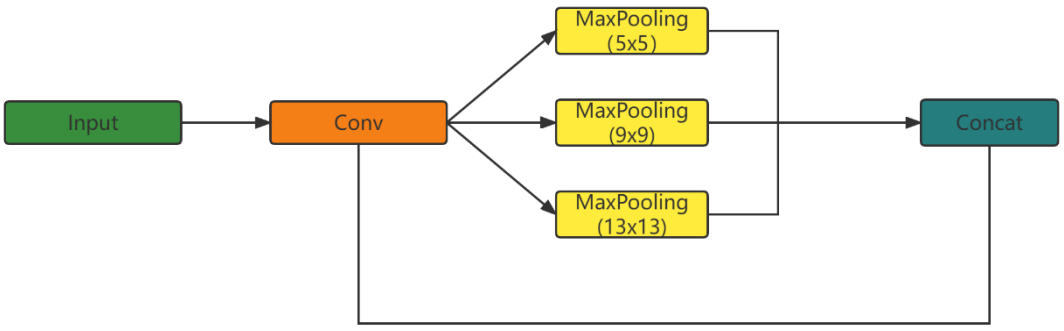


Figure3: SPP structure

ISSN: 2664-9640

3.5. Neck

The structure of the Neck part is composed of two parts: FPN (feature pyramid netword) and PAN (path augmentation network). The FPN part is to fuse the multiple feature layers output by Backbone from top to bottom through horizontal connections. This way of fusing low-resolution features with high-resolution features strengthens the semantic information of low-resolution features. PAN is complementary to the structure of FPN. The original FPN just fuses the low-resolution features output by Backbone with the previous layer, and then makes predictions without considering the semantic information of high-resolution feature maps. The PAN network structure does not complement the previous FPN structure, and the low-resolution features of the features can be combined with the features. The high-resolution features are further fused to improve the ability of target detection in target positioning, thereby improving the detection accuracy. Neck structure is shown in Figure 4

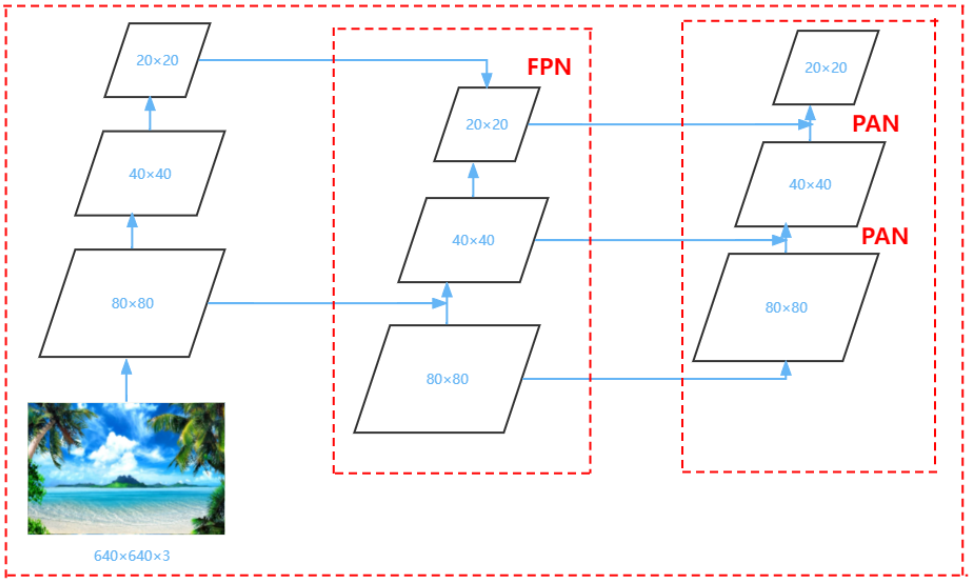


Figure4: Neck structure

3.6. Prediction

3.6.1Matching feature points

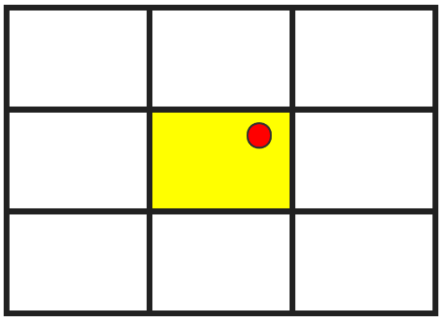


Figure5a: Matching feature points in YOLOv3

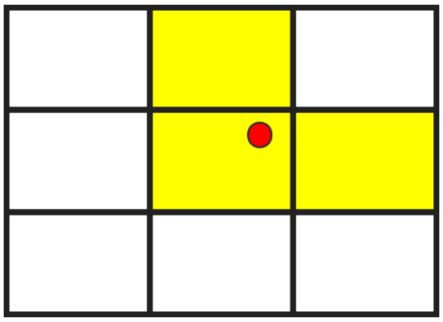


Figure5b: Matching feature points inYOLOv5

In YOLO3, each ground-truth box has a feature point in the upper left corner of the grid that is responsible for prediction. For the selected feature layer, first calculate that the ground truth box falls within the grid, and the feature point in the upper left corner of the grid is a feature

ISSN: 2664-9640

point responsible for prediction. In YOLOv5, the grid of the real frame responsible for prediction is not the grid of its center point, and two grids closest to the center point grid are added, and these three grids are considered to be responsible for predicting and changing the real frame The details are shown in Figure 5.

3.6.2Loss Function

The total loss function of YOLOV5 is shown in formula (1), which consists of three parts: localization loss, confidence loss and category loss. Among them, confidence loss and category loss are calculated by binary cross entropy loss function, localization loss, confidence loss The formulas of degree loss and type loss are shown in formula (2) (3) (4) respectively.

$$\text{Loss}_{\text{object}} = \text{Loss}_{\text{loc}} + \text{Loss}_{\text{conf}} + \text{Loss}_{\text{class}} \tag{1}$$

$$\text{Loss}_{\text{loc}} = 1 - \text{GIoU} \tag{2}$$

$$\begin{aligned}\text{Loss}_{\text{conf}} = &-\sum_{i=0}^{K\times K} I_{ij}^{\text{obj}} \left[\hat{C}_i^j \log C_i^j \right. \\ &\left. + \left(1-\hat{C}_i^j\right)\log\left(1-C_i^j\right)\right] - \\ \lambda_{\text{noobj}} &\sum_{i=0}^{K\times K}\sum_{j=0}^{M} I_{ij}^{\text{noobj}} \left[\hat{C}_i^j \log C_i^j \right. \\ &\left. + \left(1-\hat{C}_i^j\right)\log\left(1-C_i^j\right)\right]\end{aligned} \tag{3}$$

$$\begin{aligned}\text{Loss}_{\text{class}} = &-\sum_{i=0}^{K\times K} I_{ij}^{\text{obj}} \sum_{c\in \text{ classes}} \left[\hat{P}_i^j \log P_i^j \right. \\ &\left. + \left(1-\hat{P}_i^j\right)\log\left(1-P_i^j\right)\right]\end{aligned} \tag{4}$$

Where K refers to the feature map output by the network divided into KxK grids; M represents the number of anchors corresponding to each grid; I_{ij}^{obj} represents anchors with targets; I_{ij}^{noobj} indicates no target anchors; λ_{noobj} Indicates the weight of the corresponding confidence that there is no target anchor.

The traditional target detection algorithm uses the IOU loss function, but when the IOU does not intersect the real frame and the predicted frame, the value of the IOU is always 0. It cannot accurately reflect the relative position of the real frame and the predicted frame, which is not conducive to the back propagation of the model. . The loss part of YOLOv5 adopts the GIOU (generalized intersection over unnion loss) loss function. Compared with the IOU loss function, GIOU considers the overlapping area, center point distance and aspect ratio factors between the real frame and the predicted frame, which is more conducive to the Model training. The formulas and example diagrams of IOU and GIOU are equations (5) (6) and (6), respectively.

$$\text{IOU} = \frac{(\text{A}\cap\text{B})}{(\text{A}\cup\text{B})} \tag{5}$$

$$\text{GIOU} = \text{IOU} - \frac{\text{C} - (\text{A}\cup\text{B})}{\text{C}} \tag{6}$$

ISSN: 2664-9640

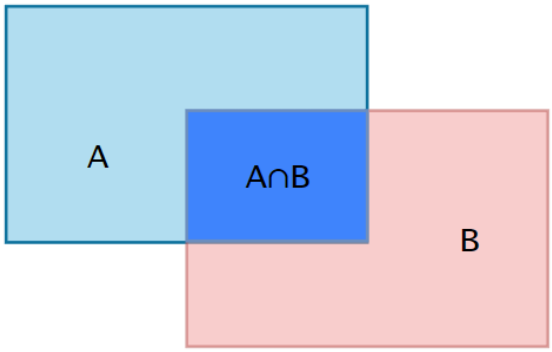


Figure6a: IOU

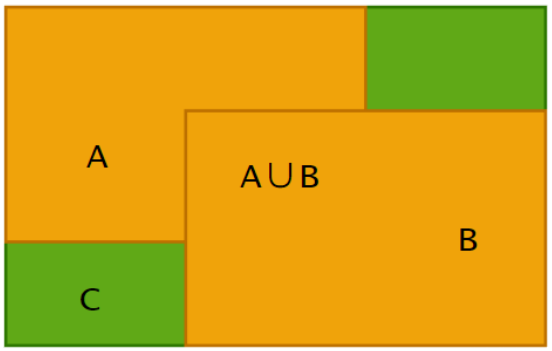


Figure6b: GIOU

4. Experiment and result

The YOLOv5 algorithm is applied to the night crowd detection, and compared with the YOLOv3 algorithm. Figure 1 shows the hardware configuration and experimental environment.

Tabel 1. Experimental environment

Item	Configuration
OS	Linux
GPU	3080
CPU	Intel(R) Xeon(R) Platinum 8255C
Framework	PyTorch 1.7.
Data annotation	LabelImag

4.1. Experimental Hyperparameter settings

We used the hyperparameters of Figure 2 in the YOLOv5 experiment, the image size of the network input was 1280x1021, the batch size was set to 16 according to the performance of the GPU and initial learning rate was set to 0.01. we user the SGD optimization algorithm to optimize the network parameters.

Tabel 2. Hyperparameter settings

Hyperparamter	Image Size	Batch Size	Epoch	Optimizer	Learing rate
Value/Type	1280x1280	16	120	SGD	0.01

4.2. DataSet

The data set comes from the video shot by the camera monitoring the intersection, and then extracts each frame of the video to form picture data, generating about 15,030 pictures. Then the data is divided into 9424 pieces of training set data, 2357 pieces of validation set data, and 3249 pieces of test set data. Then use the labeling software LabelImg to label the data set in YOLO format, and label the Person. After the labeling is completed, each picture corresponds to a txt file with the same name as the picture, and each line in the txt file represents a label instance. A total of 5 columns, from left to right, are: category, the ratio of the abscissa of the center of the marker box to the width of the picture, the ratio of the ordinate of the center of

ISSN: 2664-9640

the marker box to the height of the picture, the ratio of the width of the marker box to the width of the picture, the height of the marker box and the picture ratio of heights.

4.3. Evaluation Metrics and Implementation Results Analysis

This paper uses precision and recall and map as evaluation indicators. Precision and recall are contradictory. When the precision is very high, the recall is usually very low, and when the recall is very high, the precision is usually very low. Therefore, we add F1-sorce as the standard for integrating the two, and the specific formula is as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$F1 - Score = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{9}$$

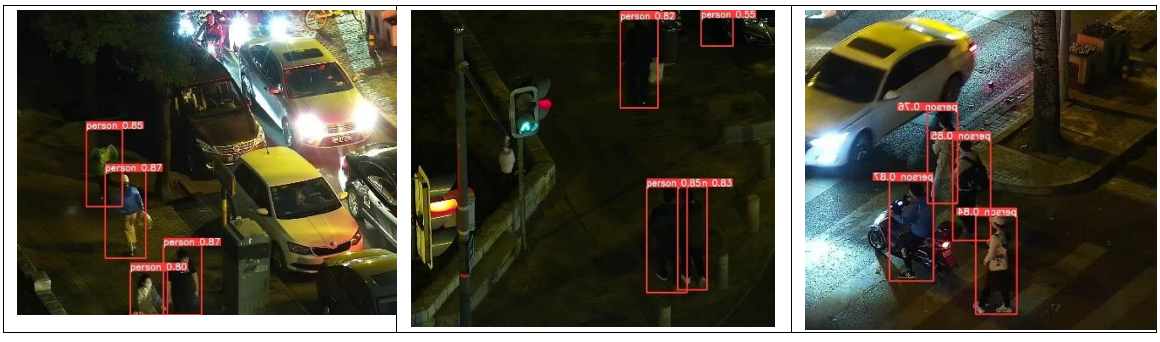
where true positive (TP) means that the aircraft is correctly predicted, false positive (FP) means that the actual category is a false alarm but the predicted class is the aircraft, and false negative (FN) means that the actual category is the aircraft, but the predicted category is a false alarm. Since there are many methods to eliminate false alarm targets.

4.4. Results

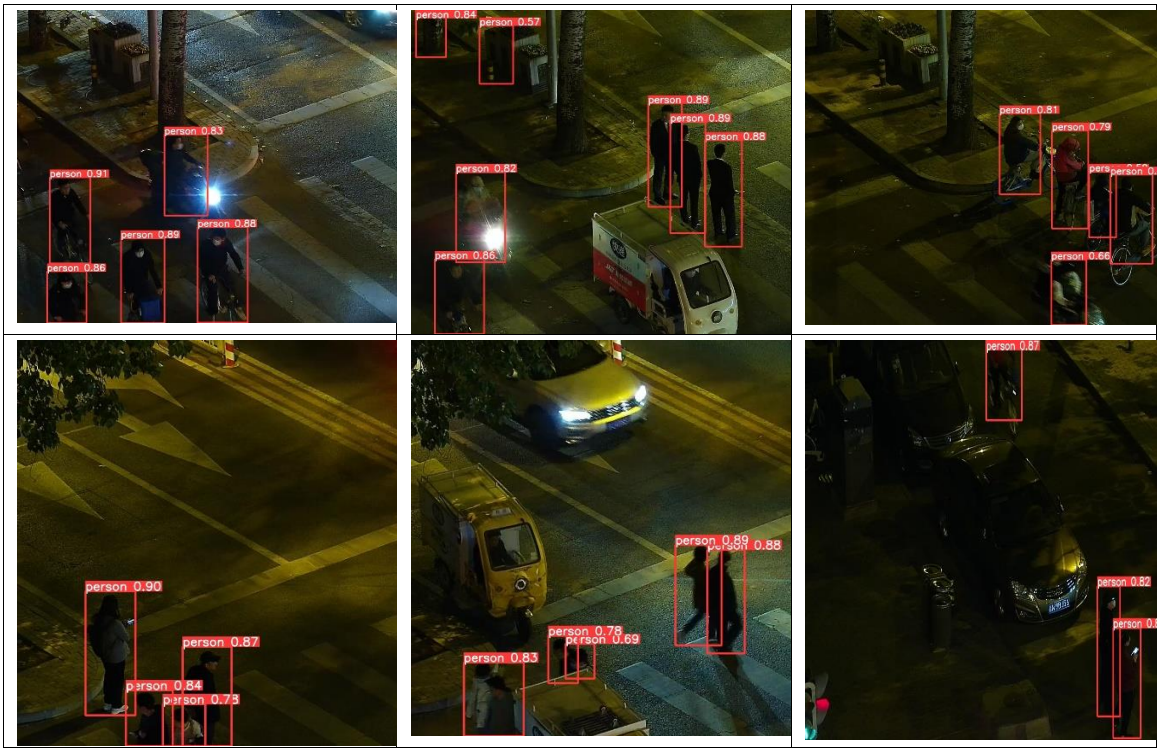
We tested YOLOv5 and YOLOv3 on the our data set. We recorded the differences between the models on the four aspects of Params,Precision,Recall and F1-score. The YOLOv5 completely surpasses YOLOv3 in four aspects. the three performance evaluation indicators of the model, Precision, Recall, F1-score, mAP@.5, and mAP@.5:.95 have been improved. The results are shown in Table 3, and in Figure 7.

Tabel 3. Experimental Result

Models	Params	Precision /%	Recall /%	F1-Score	mAP@.5	mAP@.5:.95
YOLOv3	6,257,3300	88.1	86.3	87.1	89.1	45.5
YOLOv5	7,012,822	93.7	89.6	90.9	94.9	55.8



ISSN: 2664-9640



5. Conclusion

In this paper, the target detection algorithm of YOLOv5 is used in pedestrian detection at night, which has a significant improvement in performance and detection accuracy compared to YOLOv3. In this field, it has certain practical value.

References

[1] Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.

[2] Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.

[3] Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv 2015, arXiv:1506.01497. [CrossRef] [PubMed]

[4] He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969

[5] . Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. arXiv 2016, arXiv:1605.06409.

[6] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In European Conference on Computer Vision; Springer: Cham, The Netherlands, 2016; pp. 21–37.

[7] Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.

[8] Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.

[9] Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. arXiv 2018, arXiv:1804.02767. 10. Bochkovskiy, A.;

ISSN: 2664-9640

[10] Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. arXiv 2020, arXiv:2004.10934.

[11] Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.

[12] Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.