Research on Pedestrian Detection Algorithm Based on Deep Learning

Ying Wang, Ying Tian

Abstract—**Pedestrian detection model based on deep learning has been widely used in various fields. However, its capabilities are limited when operating in complex environments due to high false alarm rates and low detection accuracy. To address these limitations, this paper presents an improved model called YOLOv5s-FCC. This enhanced model retains the distinctive features of the original model while improving several aspects such as the loss function, feature extraction, feature output, and anchor frames. The model's performance is evaluated on the CrowdHuman and WiderPerson datasets, which consist of high pedestrian density and significant obstacles, making detection challenging. To overcome these challenges, the datasets are first re-clustered using the K-Means++ clustering technique to obtain optimal anchor frames. Additionally, the Focal-EIOU loss function is employed to accelerate convergence speed and improve regression results. A coordinate convolution layer is fused before the output of different-scale features to provide more informative content. Lastly, a new pyramid pool layer of CSPCSPPF space is used to extract feature information, which reduces the repetition of gradient information and realizes more accurate detection. Comparative experiments using testing sets from both datasets demonstrate an improvement in Mean Average Precision of 2.2% and 1.8%, respectively, along with a reduction in the miss rate of 4.6% and 3.7%.**

Index Terms—**Pedestrian detection; Object detection; YOLO; Focal-EIOU loss; Coordinate convolution**

I. Introduction

DEEP learning has become a fundamental technique in computer vision, thanks to its ability to independently learn data features independently and achieve significant advancements in general object detection [1]. Pedestrian detection, which is a complex aspect of object detection, has received significant attention from researchers and is widely applied in intelligent transportation, autonomous driving, intelligent security, robotics, and other fields. Pedestrian detection serves as the technical foundation for pedestrian re-recognition, as well as human behavior and posture analysis [2]. Although this technology has achieved commendable results, detecting pedestrians in complex environments remains challenging, with substantial room for improvement in terms of both detection accuracy and speed.

Pedestrian detection can be simply understood as locating all pedestrian objects in a given image or video sequence [3]. Compared to other object detection tasks, it is more challenging due to the high uncertainty associated with pedestrian posture, clothing and behaviors. The detection performance often falls short for small pedestrians in the distance, as well as in situations involving intra-class and inter-class occlusion. This paper aims to enhance an existing deep learning-based object detection model to improve the accuracy of pedestrian detection in complex environments. The improvements encompass optimizing predefined anchor frames, enhancing feature extraction and the loss function, and refining feature output.

The object detection model based on deep learning can be roughly divided into one-stage and two-stage algorithms. In the two-stage object detection algorithm, the process includes two stages of feature extraction. First, the candidate region of the target is extracted, and then CNN performs feature extraction. A classic example of such a model is R-CNN [4], which has been significantly improved through the development of two-stage algorithms such as Fast R-CNN [5], Faster R-CNN [6], and Mask R-CNN [7]. On the other hand, in one-stage algorithms such as SSD [8] and the YOLO series [9]-[12], the step of extracting the candidate target region is omitted, and a single detection is used to obtain the category probability and position coordinates. Each of these two types of algorithms has its own advantages and drawbacks. Specifically, one-stage algorithms achieve higher detection speed but lower accuracy compared to two-stage algorithms.

Deep learning-based research has emerged as the main technique for pedestrian detection, leading researchers worldwide to devise many improved algorithms in this area. These innovations have paved the way for subsequent research in pedestrian detection. Chu et al. [13] presented a straightforward yet powerful proposal-based object detector designed to detect densely populated instances. This approach incorporates multi-instance prediction and integrates cutting-edge technologies, including EMD loss, SET NMS, and refinement modules. Lee et al. [14] introduced the innovative Multi-View Target Transformation (MVTT) method, specifically designed to tackle the distortion issues inherent in multi-view aggregation. By encoding all target features and constraining the region of interest for the projected features, MVTT effectively overcomes these challenges. Another significant contribution comes from Liu et al.'s proposal of the Visual Language Semantic Self-Monitoring Context Software Pedestrian Detection (VLPD) method [15]. This method stands out as it explicitly models the semantic context without relying on any additional cues, showcasing its innovation in pedestrian detection techniques. Furthermore, Liu et al. [16] introduced

Manuscript received May 29, 2023; revised October 7, 2023.

Ying Wang is a postgraduate student of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: wangying_yy0609@163.com).

Ying Tian is a Professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (corresponding author to provide phone: +8613898015263; e-mail: astianying@126.com).

a module named Asymptotic Positioning Fitting (ALF). This module, though simple, proves to be highly effective. ALF involves stacking multiple predictors and enhances detection results by evolving the default anchor frame of the Single Shot MultiBox Detector (SSD) directly and incrementally.

This study adopts version 6.0 of YOLOv5s [17] as the basic framework, with subsequent modifications aimed at augmenting the precision of pedestrian detection in complex environments. The modification approach involves the re-clustering of pre-defined anchor frames to derive detection anchor frames that are better suited to the specific dataset used in this investigation. To enhance the model's precision in locating targets, a coordinate convolution structure layer is introduced before the multi-scale feature output. Furthermore, the Focal-EIoU loss function is introduced to discern between high-quality and low-quality anchor frames based on gradient information, effectively suppressing the latter. Finally, the SPPF structure is replaced by the CSPCSPPF module, enabling the extraction of features at a deeper hierarchical level.

II. Related Algorithms

YOLOv5 is a prominent one-stage object detection algorithm that combines several superior deep learning object detection frameworks and continues to be refined. Although there is no official publication, it is widely recognized in academic circles. The YOLOv5 model can be categorized into four different architectures(S, M, L, and X) based on the size of the model, with the model's size and parameters increasing progressively [18]. The key distinction between these models is the location of the feature extraction module and the number of convolution kernels [19]. In this study, the YOLOv5s model, which is the smallest, fastest and easiest to train, is used as the basis for model improvement. YOLOv5s consists of four main components: Input, Backbone, Neck, and Head. Input conducts mosaic data enhancement on the input data, The Backbone module is responsible for image feature extraction and includes CBS (Conv+BN+SiLU), C3, SPPF, and other modules. C3 is a feature extraction module based on convolution operations. The C3 module in this model consists of two structures that differ in their linkage mode. SPPF is a spatial pyramid pooling module for maximum pooling operation, which connects feature maps of different scales through the convolution layer to fuse feature knowledge. In the meantime, Neck integrates various feature data tiers to boost the model's stability and accuracy. The prediction component of the network is the Head, where each prediction head outputs object detection results at different scales.

III. Improvements

The performance of pedestrian detection in complex environments is suboptimal. However, occlusion, density, and other factors affect pedestrian detection, and urgently need to be solved to mitigate problems such as low accuracy and high false detection rate. This paper proposes a solution to the aforementioned problems by enhancing the YOLOv5s network model, as demonstrated in the illustrated structure diagram of the improved YOLOv5s-FCC model Fig. 1.

The YOLOv5s network model outputs prediction frames using preset anchor frames and updates network parameters. The datasets used in this paper exhibit high pedestrian density, which requires re-clustering using the K-Means++ algorithm to obtain a more suitable preset anchor frame for outputting the prediction frames. To improve spatial perception and address defects in the YOLOv5s loss function, coordinates are added to the coordinate convolution prior to integration with the feature output of different scales. The Focal-EIoU loss function is also introduced to distinguish between high and low quality anchor frames, allowing the regression process can focus on high quality frames and suppress low quality ones. The CSPCSPPF module replaces the structure of SPPF and extracts features at a deeper level. The enhanced YOLOv5s-FCC model greatly increases detection precision and lowers the missed rate, according to experimental data.

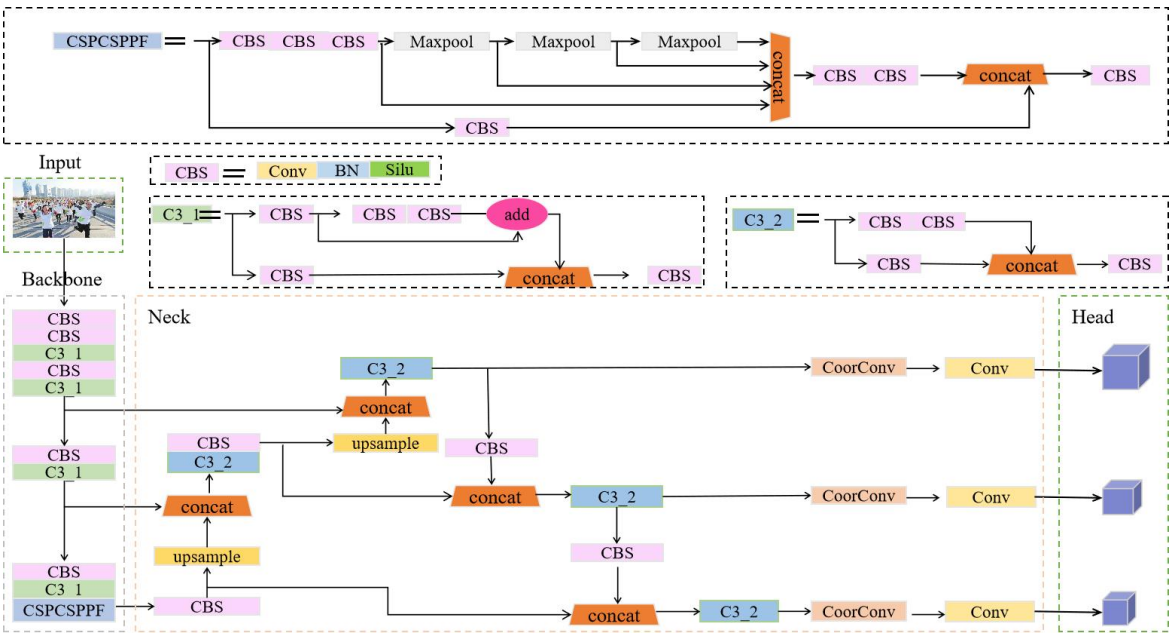


Fig. 1. YOLOv5s-FCC structure diagram

A. K-Means++ clustering anchor algorithm

The YOLOv5s model utilizes the K-Means clustering algorithm to derive 9 predefined anchor frames with varying scales. These anchor frames are then scored evaluated based on their intersection ratio following feature fusion. K-Means is an unsupervised learning algorithm specifically designed for grouping similar objects into clusters. It rapidly converges and effectively groups large datasets based on their distance similarity. The algorithm identifies K clusters and calculates the centre point of each cluster as the average of the contained values. However, the arbitrary choice of initial centroids can significantly impact convergence.

To address this issue and to improve the adaptability of the predefined anchor frames, this study proposes employing the K-Means++ clustering anchor algorithm to re-clustering the dense pedestrian datasets employed in the experiment. K-Means++ is an improved version of the K-Means clustering algorithm that addresses the problem of centroid selection. According to research, the quality of clustering results is directly proportional to the distances between clustering centres. When selecting n ($0 < n < K$), the K-Means++ algorithm favors choosing points that are farther away from the current n centres to increase the distance between clustering centres. This approach leads to better and higher quality results.

The following are the steps involved in the K-Means++ algorithm:

(1) The first initial clustering centre is selected randomly from a sample point in the data set χ.

(2) The symbol $D(x)$ denotes the distance between the present cluster centre and the sample point. Meanwhile, the symbol $P(x)$ calculates the possibility of choosing each sample point as the next cluster centre, with the maximum value being selected.

$$P(x) = \frac{D(x)^2}{\sum_{x \in \chi} D(x)^2} \tag{1}$$

(3) Repeat step (2) until k clustering centres are selected.

B. CoordConv Layer

The convolution layer is a fundamental element of deep learning network models. Traditional convolution offers advantages such as fewer parameters, high computational efficiency, and translation invariance, making it particularly suitable for classification tasks. However, it has limitations in capturing spatial information. To overcome this limitation, a coordinate convolution [20] was developed by extending the traditional convolution layer and introducing a coordinate channel next to the feature map. This additional channel allows the convolution process to incorporate spatial position information from the feature map.

In this paper, the coordinate convolution is integrated before the feature output of three different scales of the baseline network, and the position information is introduced into the features of different channels before the feature output to further determine the target spatial position, which improves the accuracy of target detection. The structure diagram of the coordinate convolution is shown in Fig. 2.

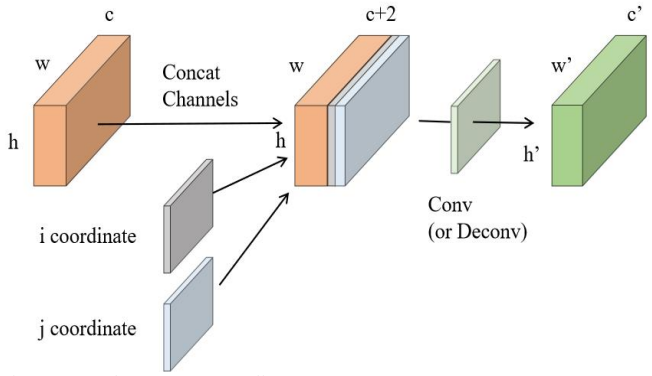


Fig. 2. CoordConv structure diagram

Fig. 2 shows that coordinate convolution involves appending two coordinate channels behind the feature map and then performing a traditional convolution operation. One of these channels representd the i-coordinate, while the other represents the j-coordinate. A convolution without modifying these channels would be an ordinary traditional convolution, preserving the characteristics of the original operation. However, if the weight of these coordinates is non-zero, the network can learn coordinate information. Thus, for different task requirements, coordinate convolution could teach the network to become either translation invariant or translation dependent.

C. Focal-EIOU Loss

In object detection, the regression of the bounding box by the loss function is the crucial step to determine the target. To address this issue, a wide range of IOU-based loss functions have been proposed, including GIOU Loss [21], DIOU Loss, CIOU Loss [22], and others. However, the previously mentioned loss functions often face issues such as slow convergence and imprecise regression results. A method called EIOU Loss [23] is proposed, which decomposes the aspect ratio based on CIOU, evaluates the disparities between the overlapping area, centre point, and aspect ratio in the regression of the bounding box, based on three geometric conditions. By taking the height and width into account in the loss function, the gap between the predicted and actual target bounding boxes is minimized, resulting in faster convergence and more accurate positioning. Focal Loss is introduced to reduce the impact of irrelevant anchor frames and improve the effectiveness of optimal anchor frames in model optimization, Focal Loss is introduced. Focal EIOU Loss function is proposed to improve the performance of EIOU. The formula is as follows:

$$\begin{aligned} L_{EIOU} &= L_{IOU} + L_{dis} + L_{asp} \\ &= 1 - IOU + \frac{\rho^2(b, b^{gt})}{(w^c)^2 + (h^c)^2} \\ &\quad + \frac{\rho^2(w, w^{gt})}{(w^c)^2} + \frac{\rho^2(h, h^{gt})}{(h^c)^2} \end{aligned} \tag{2}$$

$$L_{Focal-EIOU} = IOU^{\gamma} L_{EIOU} \tag{3}$$

EIOU loss is made up of three components: IOU loss, distance loss, and aspect loss. In the formula above, the variable w^c represents the width of the minimum bounding rectangle between the predicted and actual bounding boxes, and h^c represents the height. The variable ρ corresponds to the

Euclidean distance between two points. The parameter b denotes the centre of the bounding box, while the parameter γ adjusts the degree of outlier suppression. Experimental results indicate that a setting of γ=0.5 gives optimal performance. By applying the Focal-EIOU loss function, the imbalance in the training samples can be mitigated, leading to improved detection accuracy.

D. CSPCSPPF

The YOLO series of object detection networks utilize spatial pyramid pooling (SPP) [24] in feature extraction to generate multiple receptive fields through maximum pooling, making the algorithm adaptable to images with varying resolutions. In YOLOv5, the SPPF module increase extraction speed by appending a convolutional layer to SPP. The CSP module [25] addresses the challenge of heavy computation in network optimisation resulting from the duplicity of gradient information from the standpoint of network structure design. The CSP module initially splits the feature dimension of the base layer into two halves without duplicating gradient data, speeding up computation and increasing detection accuracy.

Through the combination of CSP structure and SPPF, CSPCSPPF structure can obtain four different receptive fields through the maximum pooling of four respective scales, which can better handle different object information and distinguish large and small targets. Then, after processing with the spatial pyramid's pooling layer, it passes to the CSP module, which divides the features to enable the network model to learn more information. In this paper, the CSPCSPPF module is incorporated into the feature extraction to replace the SPPF module, resulting in better detection accuracy in the baseline model.

IV. Experiments

A. Experiment Environments

In this experiment, we used a DELL R730 server, a single NVIDIA GTX1080Ti graphics card, and CentOS7 as the operating system. The development ecosystem comprises Python3.7 in PyCharm 2019.3.3 and the PyTorch1.7.1 deep learning framework. We use YOLOv5s as the baseline network, with a batch size set of 16, an initial learning rate assigned to 0.01, a momentum factor of 0.937, and a weight attenuation coefficient of 0.0005. We train the model for a total of 200 epochs. Sentence structures have been modified to improve clarity and readability.

B. Datasets

In the present study employs the CrowdHuman [26] and WiderPerson [27] datasets for the detection experiment. Due to the unavailability of test set labels, the necessary processing is performed on the dataset information, including reference pixels and label categories. Subsequently, the training set and verification set data would be divide into subsets consisting of 80% training set, 10% verification set, and 10% test set. The CrowdHuman dataset is a large dataset with a significant amount of data. The training and verification sets contain an excessive number of samples, with an average of 20 pedestrians per image and varying degrees and types of occlusion. Furthermore, each human example has three distinct types of labels, namely head, visible area, and whole body. The present study uses 19,368 publicly labelled images, proportionally divided into 15,494 training sets, 1,938 verification sets, and 1,936 test sets. The WiderPerson dataset, it is a benchmark dataset for pedestrian detection covering a wide range of scenes, beyond just traffic scenarios. The study divided 9,000 publicly labelled images into subsets of 80% training set, 10% verification set, and 10% test set, resulting in 7,200 training sets, 900 verification sets, and 900 test sets.

C. Evaluation Indicators

Commonly utilized metrics in object detection encompass recall, precision, and mAP. Recall signifies the proportion of accurately identified positive samples, whereas precision denotes the ratio of positive samples among all identified samples. To calculate mAP, it is essential to compute the average precision, corresponding to the area beneath the PR curve. The mAP is determined by the mean value of AP across all categories. In addition, within pedestrian detection, a frequently utilized evaluation index is the miss rate MR, where a lower MR indicates superior performance.

The following formula is used to calculate Recall:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

The following formula is used to calculate Precision:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

The following formula is used to calculate mAP:

$$mAP = \int_0^1 p(R)dR \tag{6}$$

The following formula is used to calculate MR:

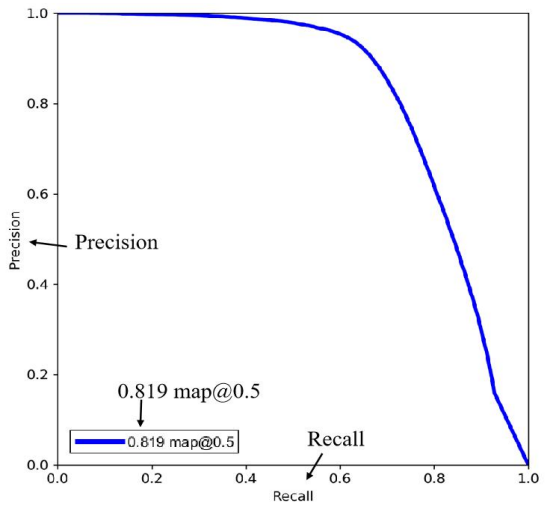
$$MR = \frac{FN}{FN + TP} \tag{7}$$

D. Model Contrast Experiments

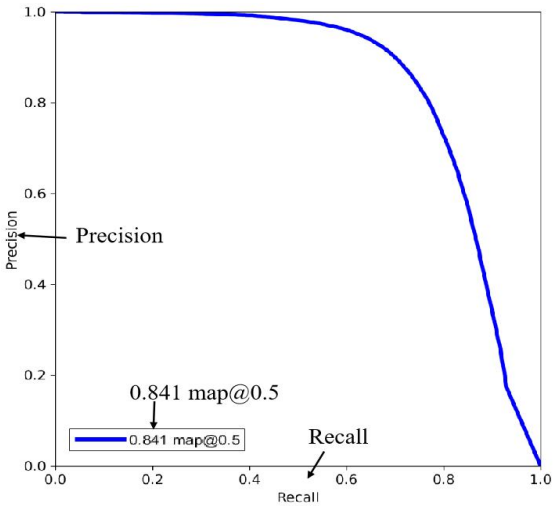
In this study, the performance of four detection networks was assessed through a comparison of experimental results obtained from the CrowdHuman and WiderPerson datasets. Table I displays the mean average precision (mAP) and miss rate (MR) for each network. Specifically, this paper introduces an improved network, YOLOv5s-FCC, which is compared to three other networks: Faster-RCNN (a two-stage model), YOLOv4-tiny (an one-stage lightweight model), and YOLOv5s (a baseline network).

TABLE I
The experimental results of each networks on CrowdHuman and WiderPerson dataset

Model Name	CrowdHuman		WiderPerson	
	mAP (%)	MR (%)	mAP (%)	MR (%)
Faster-RCNN	81.2	55.8	64.3	71.4
YOLOv4-tiny	75.7	44.6	51.6	49.5
YOLOv5s	81.9	31.4	70.1	40.4
YOLOv5s-FCC	84.1	26.8	71.9	36.7

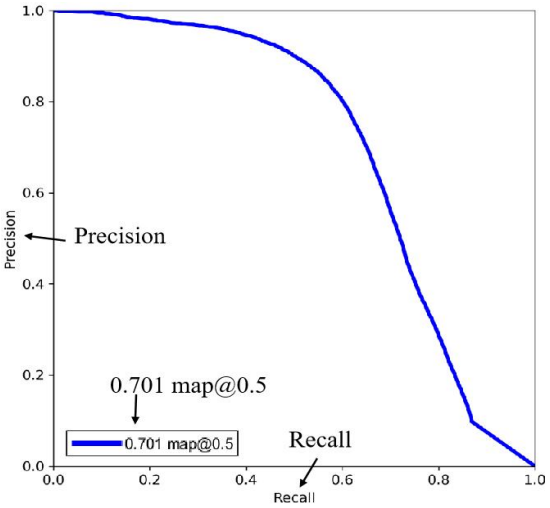


(a) The PR curve of CrowdHuman dataset(before)

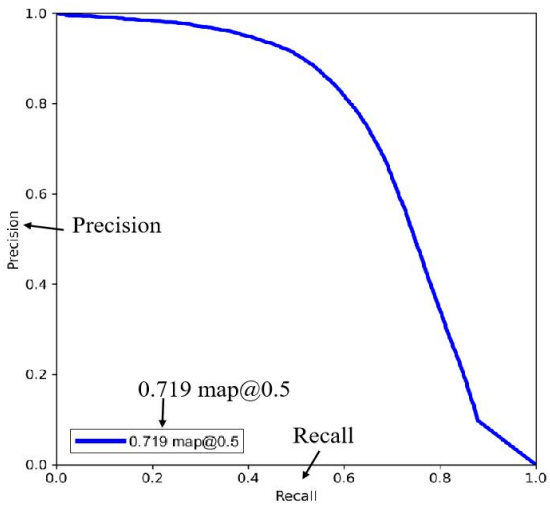


(b) The PR curve of CrowdHuman dataset(after)

Fig. 3. The PR curve of the CrowdHuman dataset



(a) The PR curve of WiderPerson dataset(before)



(b) The PR curve of WiderPerson dataset(after)

Fig. 4. The PR curve of the WiderPerson dataset

E. P-R Curves Comparison.

To more intuitively express the contrast effect between the improved network and the baseline network more intuitively, we trained the two models and plotted their PR curves for comparative analysis. Fig. 3 is a comparison plot of PR curves on the CrowdHuman pedestrian detection dataset, where (a) is the PR curve of the original YOLOv5s model and (b) is the PR curve of the improved YOLOv5s-FCC model. Fig. 4 is a comparison diagram of the PR curves on the WiderPerson pedestrian detection dataset, where (a) is the PR curve of the original YOLOv5s model and (b) is the PR curve of the improved YOLOv5s-FCC model

The figure above shows the PR curves of two data sets under the baseline model and the improved model. The mAP value, at the intersection of each curve with the y-axis and x-axis when IOU is equal to 0.5, can be used to assess the effectiveness of the model. The area enclosed by the PR curve and the horizontal and vertical coordinates can represent the performance of the model. The results show that the PR curve area of YOLOv5s-FCC is larger, which indicates that the model has better performance. The increase in mAP is consistent across both datasets.

F. Ablation Experiment

To explore the universality of the improved algorithm in this paper, we conducted the following five groups of experiments on CrowdHUman and WiderPerson.pedestrian detection datasets.

Group 1: The YOLOv5s baseline network is trained and verified on both datasets.

Group 2: Both datasets are re-clustered using K-Means++. On this basis, the YOLOv5s model is trained and verified.

Group 3: Coordinate convolution is applied to the third group of enhancements to train and verify the enhancement effect before outputting three different feature scales.

Group 4: Implemented the Focal-EIOU Loss as the model's loss function in Group 3 and validated its effectiveness.

Group 5: Used CSPCSPPF as the pooling layer of the spatial pyramid based on Group 4, and verified the improvement effect through training.

TABLE II
RESULT OF ABLATION EXPERIMENTS OF CROWDHUMAN DATASET

Group	Epoch	mAP（%）	MR（%）
Group 1	200	81.9	31.4
Group 2	200	82.5	30.9
Group 3	200	82.8	29.9
Group 4	200	84.0	27.3
Group 5	200	84.1	26.8

TABLE III
RESULT OF ABLATION EXPERIMENTS OF WIDERPERSON DATASET

Group	Epoch	mAP（%）	MR（%）
Group 1	200	70.1	40.4
Group 2	200	70.5	39.6
Group 3	200	70.8	39.1
Group 4	200	71.9	38.2
Group 5	200	71.9	36.7

Table II shows a 2.2% increase in mAP and a 4.6% decrease in MR on the CrowdHuman dataset, while Table III shows a 1.8% increase in mAP and a 3.7% decrease in MR on the WiderPerson dataset. These results demonstrate that with the same epoch training, the YOLOv5s-FCC model improves mAP and decreases MR compared to the original YOLOv5s model on both datasets. Although mAP was not improved in Group 5 compared to Group 4, the 1.5% decrease in MR still has the potential to optimise the model. Therefore, the YOLOv5s-FCC model remains valid.

G. *Compare of Test Results*

Images with different lighting conditions, different densities, different pedestrian postures and different definitions are carefully selected to evaluate and validate the effectiveness and universality of the enhanced algorithm. The result of YOLOv5s reasoning is shown in the left image, and the result of YOLOv5s-FCC is shown in the right image, and the mark is distinguished in the images. Fig. 5 shows the reasoning results obtained from the CrowdHuman dataset, while Fig. 6 shows the results obtained from the WiderPerson dataset.

The results in Fig. 5 show an increase in the detection confidence for the target pedestrian, as well as the detection of previously blocked or missed pedestrians due to different postures. The results in Fig. 6 show the detection confidence for pedestrians in very crowded conditions. The consistency of the inference results between the two detection datasets can increase confidence and reduce missed pedestrian detection. This proves the effectiveness of YOLOv5s-FCC.

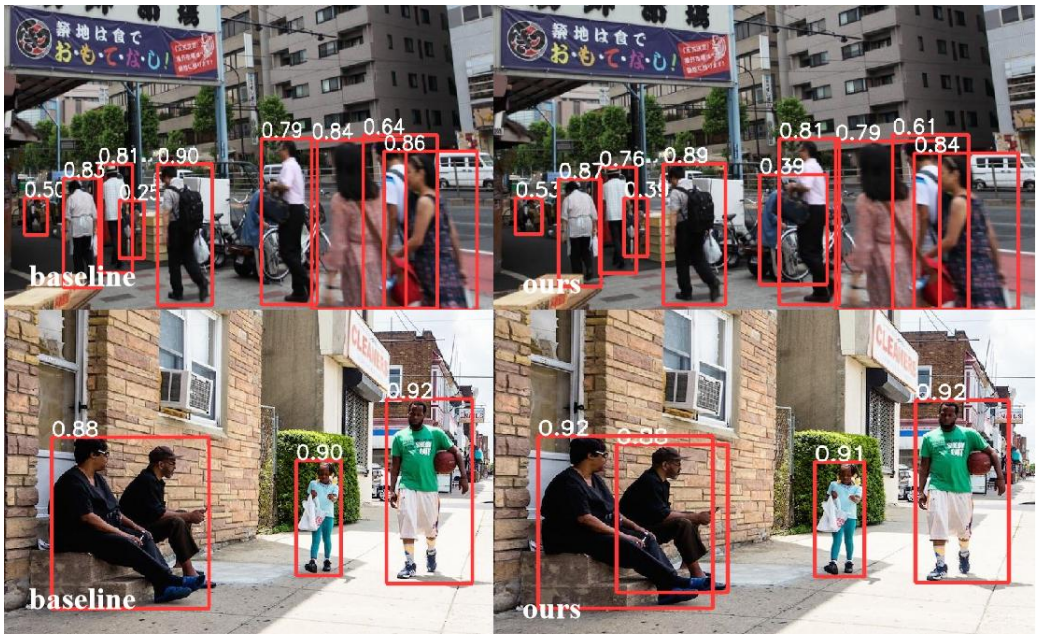


Fig. 5. Inference result on CrowdHuman dataset

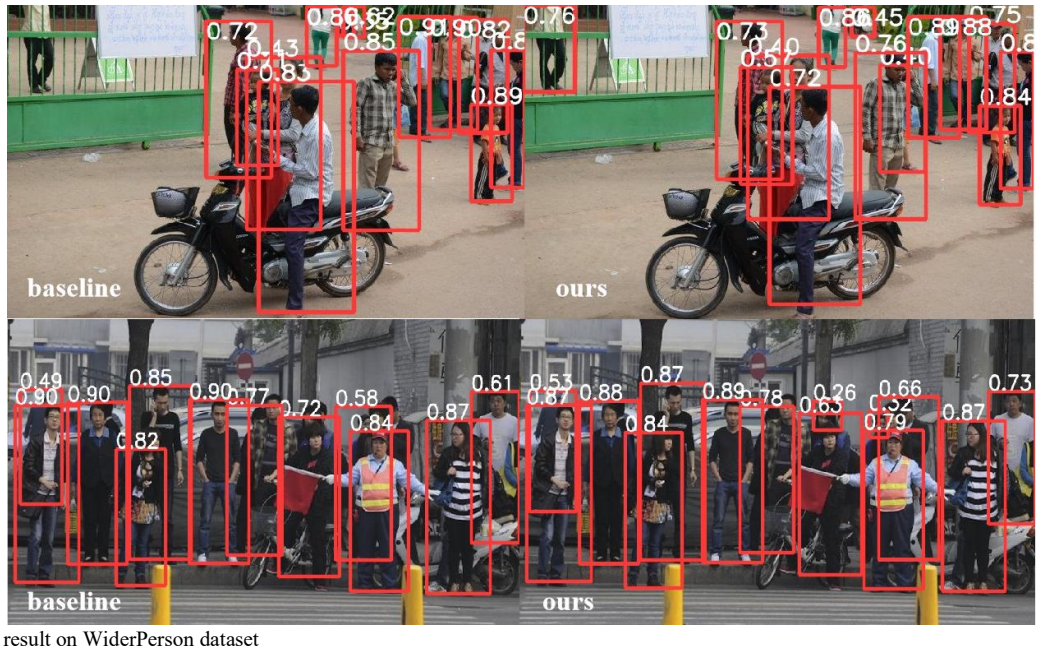


Fig. 6. Inference result on WiderPerson dataset

V. Conclusion

Based on the improvement of the YOLOv5s model, this paper proposes a human detection model called YOLOv5s-FCC. This model aims to balance the samples and obtain the target coordinate information. First, the datasets are processed by re-clustering the preset anchor frame. Then, coordinate convolution is added before the detection output to capture target position information. The positive and negative sample information is balanced using Focal-EIOU Loss. Finally, the improved feature pool pyramid is fused to improve the detection precision and reduce the miss rate in complex environments. Finally, the enhanced feature pool pyramid is integrated to improve detection accuracy and reduce miss rates in complex environments. The experimental results illustrate performance improvements in various datasets, highlighting the potential for broader application. Future research could focus on improving the speed of the pedestrian detection model to enable more efficient real-time detection.

References

[1] Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). "Deep learning for generic object detection: A survey." *International journal of computer vision*, 128, pp. 261-318.

[2] Guo, Z., Liao, W., Xiao, Y., Veelaert, P., & Philips, W. (2021). "Weak segmentation supervised deep neural networks for pedestrian detection." *Pattern Recognition*, 119, 108063.

[3] Nguyen, D. T., Li, W., & Ogunbona, P. O. (2016). "Human detection from images and videos: A survey." *Pattern Recognition*, 51, pp. 148-175.

[4] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation." 2014 *IEEE Conference on Computer Vision and Pattern Recognition, Columbus*, OH, USA, 2014, pp. 580-587.

[5] R. Girshick, "Fast R-CNN." 2015 *IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1440-1448.

[6] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[7] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). "Mask r-cnn." In *Proceedings of the IEEE international conference on computer vision*, pp. 2961-2969.

[8] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). "SSD: Single shot multibox detector." In *Computer Vision–ECCV 2016: 14th European Conference*, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 21-37.

[9] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection." 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779-788.

[10] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger." 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 6517-6525.

[11] Redmon, J., & Farhadi, A. (2018). "Yolov3: An incremental improvement." *arXiv preprint arXiv:1804.02767*.

[12] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). "Yolov4: Optimal speed and accuracy of object detection." *arXiv preprint arXiv:2004.10934*.

[13] X. Chu, A. Zheng, X. Zhang and J. Sun, "Detection in Crowded Scenes: One Proposal, Multiple Predictions." 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 12211-12220.

[14] Lee, W. Y., Jovanov, L., & Philips, W. (2023). "Multi-view Target Transformation for Pedestrian Detection." In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 90-99.

[15] Liu, M., Jiang, J., Zhu, C., & Yin, X. C. (2023). "VLPD: Context-Aware Pedestrian Detection via Vision-Language Semantic Self-Supervision." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6662-6671.

[16] Liu, W., Liao, S., Hu, W., Liang, X., & Chen, X. (2018). "Learning efficient single-stage pedestrian detectors by asymptotic localization fitting." In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 618-634.

[17] Ultralytics. yolov5 (online). Available : https://github.com/ultralytics/yolov5 (accessed on 18 May 2020).

[18] N. Yang, and J. Zhao, "Dangerous Driving Behavior Recognition Based on Improved YoloV5 and Openpose," *IAENG International Journal of Computer Science*, vol. 49, no.4, pp1112-1122, 2022.

[19] Yan, B., Fan, P., Lei, X., Liu, Z., & Yang, F. (2021). "A real-time apple targets detection method for picking robot based on improved YOLOv5." *Remote Sensing*, 13(9), 1619.

[20] Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., & Yosinski, J. (2018). "An intriguing failing of convolutional neural networks and the coordconv solution." *Advances in neural information processing systems*, 31. pp. 9628-9639.

[21] Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). "Generalized intersection over union: A metric and a loss for bounding box regression." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658-666.

[22] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2020, April). "Distance-IoU loss: Faster and better learning for bounding box regression." In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34, No. 07, pp. 12993-13000.

[23] Zhang, Y. F., Ren, W., Zhang, Z., Jia, Z., Wang, L., & Tan, T. (2022). "Focal and efficient IOU loss for accurate bounding box regression." *Neurocomputing*, 506, pp. 146-157.

[24] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015.

[25] Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., & Yeh, I. H. (2020). "CSPNet: A new backbone that can enhance learning capability of CNN." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. pp. 390-391.

[26] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd." *arXiv preprint arXiv:1805.00123, 2018*.

[27] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li and G. Guo, "WiderPerson: A Diverse Dataset for Dense Pedestrian Detection in the Wild." In *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 380-393, Feb. 2020.