BDF-YOLOV5: Improved YOLOV5 Based on Bi-directional Fusion Network for Dense Pedestrian Detection

Yuhui Xu and Ruian Liu

Abstract Pedestrian detection is a challenging task in the field of computer vision and plays a crucial role in downstream tasks, such as video surveillance and autonomous driving.Despite significant progress over the past two decades, scale variance and occlusion remain prominent issues. To address these problems, we propose BDF-YOLOv5 in this paper. Based on YOLOv5, we replace the original FPN with the BDF network structure. Furthermore, to further improve our BDF-YOLOv5, we additionally improved the loss function for bounding box regression and proposed weighted-CIOU. Extensive experimental results on the Crowdhuman dataset demonstrate the feasibility of our method. Compared to the baseline model (YOLOv5), BDF-YOLOv5 achieves an improvement of approximately 4.0%.

Keywords BDF-YOLOv5 · Pedestrian detection · YOLOv5 · Weighted-CIOU

1 Introduction

Human-centric computer vision tasks, such as pedestrian detection, face recognition, pose estimation, assisted driving, and intelligent robotics, have made great strides in the past decade. Among these tasks, pedestrian detection is one of the most basic and widely applicable. In addition to its important role in some application scenarios, such as video surveillance and traffic safety, pedestrian detection is also a foundational task for several other visual tasks. This paper aims to improve the performance of pedestrian detection and provide convenience for the aforementioned application scenarios.

In recent years, deep convolutional neural network-based object detection algorithms have been widely applied and made significant progress in natural scene object detection tasks, such as two-stage detectors including R-CNN [1], Faster R-CNN [2], SPPNet [3], and single-stage detectors such as SSD [4] and RetinaNet [5]. However,

Y. Xu · R. Liu (✉)
College of Electronic Information and Communication Engineering,
Tianjin Normal University, Tianjin 300387, China
e-mail: ruianliu@sina.com

W. Wang et al. (eds.), *Communications, Signal Processing, and Systems*, Lecture Notes in Electrical Engineering 1032, https://doi.org/10.1007/978-981-99-7505-1_52

Fig. 1 Illustrates with intuitive examples the three main issues in dense pedestrian detection tasks: occlusion, scale variation, and pedestrian diversity

most previous object detectors are designed for natural scene images. When directly applying these models to pedestrian object detection tasks in dense crowds, there are three main issues, as visually illustrated in Fig. 1. First, pedestrians are too dense, resulting in occlusions between individuals. Second, pedestrian detection scenes are often macroscopic images, leading to significant changes in target scale. Finally, the varied postures of pedestrians also pose challenges for detection.

In object detection tasks, the YOLO series [6–9] plays an important role in one-stage detectors due to its fast and efficient characteristics. In this paper, we propose an improved model, BDF-YOLOv5, based on YOLOv5 [10], to address the aforementioned three issues. The main contributions of this study are as follows:

- The backbone network still adopts the original version of CSPDarknet53, while we propose a novel feature network called BDF (Bi-Directional Fusion Network) for the neck part. By bi-directionally fusing high- and low-level semantic information of feature maps, BDF can effectively retain feature information and alleviate scale variation, thus improving the accuracy of multi-scale object detection.
- We propose a novel (BBR) bounding box regression loss function,Weighted CIOU, which balances the contribution of high-quality and low-quality regression samples. This effectively reduces false positives and mitigates occlusion issues in dense crowd scenes.
- For feature fusion, we abandoned the commonly used C3 module and instead adopted CSPNet.

2 Related Work

2.1 Pedestrian Detection Pipeline

Most pedestrian detectors, including hand-crafted feature-based methods and current mainstream deep learning models, consist of three main components: proposal generation, feature extraction, and classification and regression. Traditional hand-crafted methods such as Histograms of Oriented Gradients (HOG) analyze the local gradients of an image to capture the texture and shape information, and then send the extracted HOG features to downstream classifiers (*e.g.* SVM or AdaBoost [11]) for pedestrian detection.

With the rapid development of Convolutional Neural Networks (CNN), the research field of general object detection has also made great progress. The R-CNN series of algorithms have transformed pedestrian detection into an object detection problem, and by adopting a combination of region proposal extraction and classifiers, have gradually improved the detection accuracy. ALF, based on Single Shot MultiBox Detector (SSD) [4], stacks multiple predictors together and inherits the high efficiency of single-shot detectors, learning better detection from default anchor boxes. To improve the detection performance for occluded pedestrians, MGAN utilizes additional information from the visible boundary boxes as the guidance for attention masks [12].

2.2 Loss Functions for BBR

The bounding box regression loss function is critical to determining the performance of target localization. YOLOv1 [7] uses mean squared error to reduce the influence of large targets on the bounding box. YOLOv3 [8] proposed creating a penalty term to reduce the competitiveness of large boxes. To address the gradient disappearance problem of IOU loss, GIOU [13] incorporates the minimum bounding box into the penalty term, while DIOU [14] considers the distance between targets and anchors. CIOU [14] adds a scale and aspect ratio penalty to DIOU. SIOU [15] additionally considers angle cost,which has a faster convergence rate and better performance.

2.3 Object Detector Architecture

Based on convolutional neural networks, object detectors can generally be divided into two types: (1) one-stage detector, such as YOLOv5 [10] and SSD [4]; (2) two-stage detector, such as R-CNN [2] and CenterNet2 [16]. However, both one-stage and two-stage detectors consist of three parts, namely, backbone for feature extraction, neck for feature processing and fusion, and head for class and bounding box prediction. Next, we will introduce the neck structure in detail.

Neck. The design of neck is to better utilize the features extracted by the backbone. It performs further processing and reasonable utilization on the features maps extracted at different stages. The earliest neck structure only adopted up-sampling and down-sampling modules, which is simple and does not require feature aggregation operations. Currently, the commonly used neck structures include: FPN [17] that aggregates features from top-down paths, PANet [18] that adds additional bottom-up paths based on FPN, BiFPN [19] that repeatedly performs top-down and bottom-up operations while merging single-node inputs, and NAS-FPN [20] that uses neural architecture search to automatically design a feature network.

3 Proposed Method

3.1 About YOLOV5

Among the YOLOv5 series, four models have been proposed. They are YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. YOLOv5s is the smallest model with only 7.2 M parameters, which is very suitable for deployment on mobile devices. In our approach, we consider using the S model for real-time performance, although the X model may show better performance. Typically, YOLOv5 adopts the architecture of CSPDarknet53 and the SPP layer as the backbone, PANet as the neck, and YOLO detection head.

3.2 BDF-YOLOV5

BDF. The structure of BDF-YOLOv5 is shown in Fig. 2, due to the limitations of PANet, which only facilitates a single cross-layer connection and fails to fully exploit the exchange of high-level and low-level semantic information. Furthermore, on datasets focused on dense pedestrian detection, instances of the targets exhibit significant scale variations and include a considerable number of small objects. To overcome these challenges, we have modified the neck portion of YOLOV5 with the proposed BDF Net for improved performance.This method employs a bidirectional fusion strategy that combines bottom-up and top-down network pathways, enabling bidirectional cross-layer information exchange in intermediate feature layers. Although this incurs additional computational burden, the model accuracy is significantly improved. (P2 and P5 layers are located at the lowest and highest levels, respectively, and only support unidirectional input).

Fusion Block. As shown in Fig. 3, we abandoned the original 3×3 fusion module C3 and instead adopted CSPNet in the feature fusion block. In the residual part, we used ResConv, which first passes through a 3×3 convolutional layer, a 1×1 convolutional

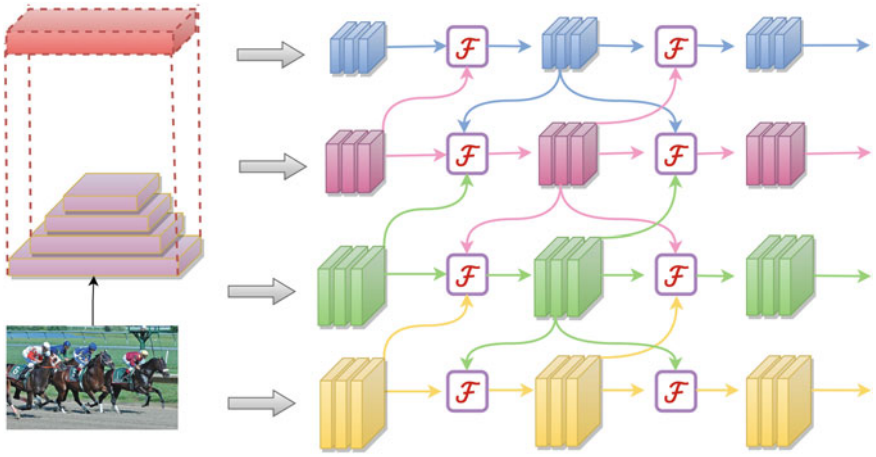

Fig. 2 The architecture of the BDF-YOLOv5. The BDF network as the neck to refine and fuse high-level semantic and low-level spatial features. In addition, a new fusion method is adopted

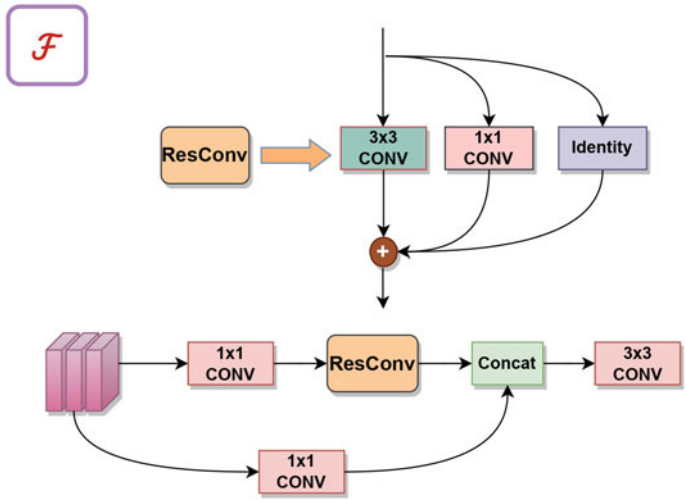

Fig. 3 The architecture of the fusion block

layer, and an identity layer, and finally adds them together directly along the channel dimension to achieve cross-layer connections.

Weighted CIOU. Most of the previous loss functions (GIOU, DIOU, CIOU) ignored the imbalance issue in Bounding Box Regression (BBR). As mentioned before, there inevitably exist a large number of low-quality samples (i.e., anchor boxes with little overlap with the target box) during the training process, which contribute the majority and thus reduce the sensitivity to high-quality regression samples during

optimization. Therefore, it is necessary to strengthen the contribution of high-quality regression boxes. In this paper, we added a weighting factor IOU^{γ} on the basis of the CIOU loss function,as the Eq. 1 shown. By judging the quality of regression boxes based on overlapping degree, we optimized the imbalance issue.

$$\phi_{CIOU} = IOU^{\gamma} \times (1 - IOU + \frac{\rho^2(b, b_{gt})}{C^2} + \alpha\upsilon). \tag{1}$$

The variables used in the Eq. 1 are defined as follows: b, b_{gt} denotes the center points of the predicted box and ground truth box, ρ denotes the Euclidean distance between the two center points, C represents the diagonal distance of the smallest closed area that can contain both boxes, υ is the similarity between the aspect ratios of the predicted box and ground truth box, and α represents the weight coefficient.

4 Experiments

4.1 Experimental Setting

The experiments were conducted in Python version 3.8 and all the models were trained and tested on an NVIDIA RTX3090 GPU. To ensure the reliability of the experiments, no pre-trained weights were used during training. The CrowdHuman [21] dataset was utilized in our experiments, which is a benchmark dataset to better evaluate detectors in crowd scenarios. The CrowdHuman dataset is large, richly-annotated, and contains high diversity.

4.2 Comparison Experiment

To demonstrate the advantages of the proposed method in this paper, we compared it with the original YOLOv5, YOLOv3, YOLOv5-lite, and SDD in terms of map0.5, map0.5:0.95, and detection speed (FPS). The results are shown in Table 1.

4.3 Ablation Studies

To further substantiate the effectiveness of our proposed improvements, we conducted several ablation experiments. Table 2 lists the effects of individual components as well as their combined impact, providing a comprehensive analysis of our network‘s performance.

Effect of the loss function. After optimizing the loss function, it was observed that the detection speed decreased from the original 107 FPS to 95 FPS, which is attributed to the introduction of additional computational steps. However, there was a remarkable improvement of 1.6 in mAP0.5, justifying the trade-off in terms of detection speed.

Effect of the BDF. By adopting our proposed BDF network, the number of layers of the original YOLOv5s has increased from 157 to the current 218, and the parameter count has also increased by 2.0 M.The sacrifice of model complexity has resulted in higher precision in detection results, indicating that our network plays a certain role in addressing the scale variance problem in objects detection.

Effect of model ensemble We integrated all our proposed improvements into a single model, as shown in Table 2. The final version achieved an improvement of 3.9 in testing performance, while maintaining a real-time detection speed of 86 FPS. These results demonstrate that our approach is capable of meeting the requirements for real-time detection applications.

5 Conclusion

In this paper, we proposed an advanced dense pedestrian detector based on YOLOv5. Throughout the experimentation process, we explored numerous feature extraction techniques, including but not limited to network architecture modifications. The results were satisfactory, with an improvement of nearly 4.0 over the original network.

Table 1 The comparison of the performance in crowdhuman dataaset

Methods	mAP0.5 (%)	mAP0.5:0.95 (%)	FPS
SSD	72.7	42.6	39
YOLOv3	83.6	50.6	51
YOLOv5-Lite	67.4	28.2	**138**
YOLOv5s	80.5	47.2	111
BDF-YOLOv5	**84.4**	**51.5**	86

The bold values are the best network model for the comparison of indicators in the ablation experiment

Table 2 Ablation Study:the impact of each added component

Methods	Parameters	mAP0.5 (%)	FPS
YOLOv5s	7.2M	80.5	111
YOLOv5 + weight-CIOU	7.2M	82.1 (↑ 1.6)	107
BDF-YOLOv5 + weight-CIOU	9.2M	83.6 (↑ 1.5)	95
BDF-YOLOv5 + weight-CIOU + fusion block	9.5M	84.4 (↑ 0.8)	86

We hope that this report will be helpful to the field of dense pedestrian detection. In the future, we plan to explore domain adaptation issues and further improve the performance of our proposed detector.

Acknowledgements Fund Project This work was supported by Tianjin Normal University Graduate Research Innovation Project Funding (2023KYCX005Y).

References

1. Girshick RB, Donahue J, Darrell T, Malik J (2013) Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE conference on computer vision and pattern recognition, pp 580–587
2. Girshick RB (2015) Fast r-cnn. In: 2015 IEEE international conference on computer vision (ICCV), pp 1440-1448
3. He K, Zhang X, Ren S, Sun J (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37:1904–1916
4. Liu W, Anguelov D, Erhan D, Szegedy C, Reed SE, Fu C-Y, Berg AC (2015) Ssd: single shot multibox detector. In: European conference on computer vision
5. Lin T-Y, Goyal P, Girshick RB, He K, Dollár P (2017) Focal loss for dense object detection. In: 2017 IEEE international conference on computer vision (ICCV), pp 2999–3007
6. Redmon J, Farhadi A (2016) Yolo9000: better, faster, stronger. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 6517–6525
7. Redmon J, Divvala SK, Girshick RB, Farhadi A (2015) You only look once: unified, real-time object detection. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 779–788
8. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. ArXiv, vol. abs/1804.02767
9. Bochkovskiy A, Wang C-Y, Liao H-YM (2020) Yolov4: optimal speed and accuracy of object detection. ArXiv, vol. abs/2004.10934
10. Jocher G, Stoken A, Borovec J, NanoCode012, Chaurasia A, TaoXie, Changyu L, Laughing AV, tkianai, yxNONG, Hogan A, lorenzomammana, AlexWang1900, Hajek J, Diaconu L, Marc, Kwon Y, oleg, wanghaoyang0106, Defretin Y, Lohia A, ml5ah, Milanko B, Fineran B, Khromov D, Yiwei D, Durgesh D, Ingham F (Apr. 2021) Ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations. [Online]. Available: https://doi.org/10.5281/zenodo.4679653
11. Lin Z, Pei W, Chen F, Zhang D, Lu G (2021) Pedestrian detection by exemplar-guided contrastive learning. IEEE Trans Image Process 32:2003–2016
12. Hasan I, Liao S, Li J, Akram SU, Shao L (2022) Pedestrian detection: domain generalization, cnns, transformers and beyond. ArXiv, vol abs/2201.03176
13. Rezatofighi SH, Tsoi N, Gwak J, Sadeghian A, Reid ID, Savarese S (2019) Generalized intersection over union: a metric and a loss for bounding box regression. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 658–666
14. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D (2019) Distance-iou loss: faster and better learning for bounding box regression. In: AAAI conference on artificial intelligence
15. Gevorgyan Z (2022) Siou loss: more powerful learning for bounding box regression. ArXiv, vol abs/2205.12740
16. Zhou X, Koltun V, Krähenbühl P (2021) Probabilistic two-stage detection. ArXiv, vol abs/2103.07461
17. Lin T-Y, Dollár P, Girshick RB, He K, Hariharan B, Belongie SJ (2016) Feature pyramid networks for object detection. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 936–944

18. Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path aggregation network for instance segmentation. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 8759–8768
19. Tan M, Pang R, Le QV (2019) Efficientdet: scalable and efficient object detection. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10 778–10 787
20. Ghiasi G, Lin T-Y, Pang R, Le QV (2019) Nas-fpn: learning scalable feature pyramid architecture for object detection. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 7029–7038
21. Shao S, Zhao Z, Li B, Xiao T, Yu G, Zhang X, Sun J (2018) Crowdhuman: a benchmark for detecting human in a crowd. ArXiv, vol abs/1805.00123