Study on Small Target Detection of Construction Site Helmet Based on YOLOv5

Xingyan Xia[1] and Junyong Zhai[2(✉)]

[1] School of Software, Southeast University, Suzhou 215123, China
[2] School of Automation, Southeast University, Nanjing 210096, China
jyzhai@seu.edu.cn

Abstract. Recently, the helmet wear detection method based on computer vision has become an important means of construction units to implement management. Improving the detection accuracy and detection speed of helmet wear identification is the critical challenge for applications. On the construction sites, the camera's location is high and far from workers. As detection targets in the small and medium range, workers and helmets require a smaller sensory field for detection. This paper aims to improve the network structure of YOLOv5 by adding a set of convolutional layers to the backbone network and performing feature fusion with shallow residual network. In each branch, helmet detection is performed according to the front and rear semantic information to form a deep fusion fast safety helmet detection model. In terms of loss function, this paper also makes improvements accordingly, replacing GIoU Loss with the better CIoU Loss, and using the improved model algorithm to detect the state of workers wearing helmets, the accuracy can reach 91.6%, and the mAP reaches 93.2%.

Keywords: Helmet detection · YOLOv5 · Small target detection

1 Introduction

Recently, intelligent surveillance has been used as one of the primary fields of computer vision engineering. With the development of computer hardware, GPUs have been widely used for parallel computing. The increase in computational speed has made it a reality to train large deep neural networks. Many scholars have proposed a set of deep learning-based target detection algorithms, such as R-CNN [1], Fast R-CNN [2], and Faster R-CNN [3]. The literature [4] proposed YOLOv1 to further improve the detection rate. The work [5] proposed YOLOv3 detection algorithm, which reached 45 f/s on video detection. The literature [6] proposed YOLOv4 detection algorithm in 2020. Then, the YOLOv5 detection algorithm was proposed as a new detection algorithm, which has the advantages of high detection judgment accuracy, flexibility and speed, and provides pre-trained models of different sizes [7]. In the YOLOv5s model with a small size of data, the fastest object detection in the video can reach 140 f/s.

Y. Jia et al. (Eds.): CISC 2022, LNEE 951, pp. 18–27, 2022.
https://doi.org/10.1007/978-981-19-6226-4_3

The construction industry is an essential pillar of our national economy, and many construction workers work long hours on construction sites. At present, the identification of helmet wearing on the construction site is mainly based on manual inspection, which leads to a series of problems, such as high supervision costs, subjective interference, and inability to monitor the whole process. By deploying video monitoring equipment on the construction site and using machine vision-related methods to automatically identify helmets, it is possible to reduce supervision costs while improving regulatory information level. The work [8] combined multi-scale training, increasing the number of anchor points, and Online Hard Example Mining (OHEM) [9] on Faster R-CNN to make helmet-wearing detection more accurate. The work [10] incorporated dense blocks into the YOLOv2 target detection algorithm and compressed the model to one-tenth of the original size using MobileNets [11] lightweight network structure to make the helmet detection faster. The above algorithms have their own advantages in detection accuracy and detection speed, but cannot effectively guarantee the accuracy and efficiency of the detection of small targets detection. Based on the YOLOv2 network architecture, the work [12] proposed an advanced small target detection algorithm. By constructing a feature pyramid network (FPN) [13], the image description problem of helmet wearing detection on construction sites is solved. However, the detection speed is relatively slow and there are some limitations in real-time detection, which does not meet the demand of the actual construction environment. The work [14] proposed a helmet detection method in a complex environments under far-field video detection conditions and generated a dataset that satisfies various features of construction site environments, and used the Faster R-CNN algorithm to achieve helmet wearing small target detection in far-field situations. However, the results showed that its detection speed is less than 5 f/s, which does not meet the needs of the actual construction environment.

From the above analysis, it is clear that the helmet detection problem is one of the hotspots researched by scholars. It is still a challenge to improve the recognition speed of small targets and match the demands of real-time detection while maintaining high precision. To end this, this paper proposes an advanced method. A set of convolutional layers dedicated to detecting small targets is added to the shallow position of the backbone network by fusing it with deep features, which makes the perceptual field of the network smaller for the targets and can take better care of small targets. In addition, this paper also improves the original loss function to improve the detection accuracy of the network. The main contribution of this paper is to propose a helmet detection algorithm that can replace manual work to achieve stable detection of helmet wearing status in complex environments. Finally, the experimental results show that the improved method effectively improves the detection accuracy of small helmet targets and the detection speed is almost unchanged, with the detection accuracy reaching 91.6% and mAP reaching 93.2%.

2 Main Results

2.1 Overview of YOLOv5

YOLOv5 is a one-stage object detection algorithm for object detection, improved from YOLOv1-YOLOv4 [4–6,15]. Unlike Faster-RCNN [3], YOLOv5 network is obtained directly from the coordinates of the bounding box and the probability of each class through regression, which makes the object detection model much faster and better for real-time detection. The network structure of YOLOv5 can be separated into four parts: input, backbone, neck, and prediction layers. The structure of the YOLOv5 network model is shown in Fig. 1.

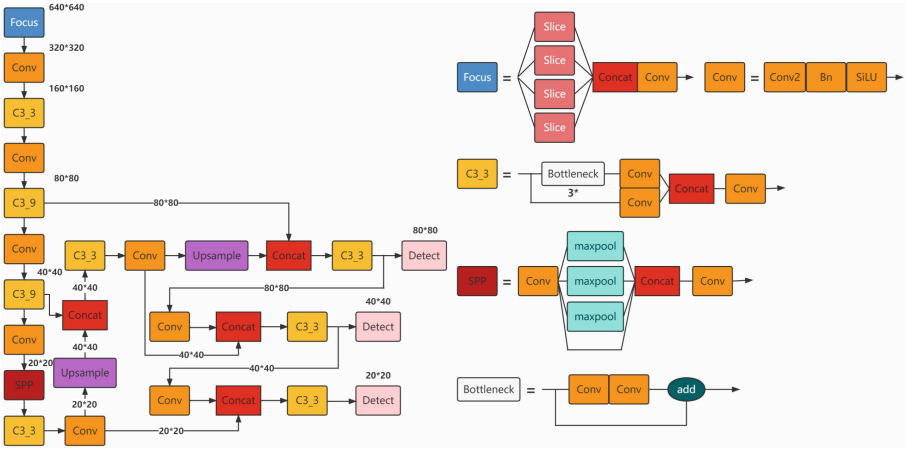


Fig. 1. YOLOv5 network structure

YOLOv5 adopts the Mosaic data enhancement method on the input side, randomly calling four images and stacking them by random size and distribution to increase the precision of small target recognition, and can calculate four images at the same time to reduce the memory consumption. In addition, YOLOv5 also adds adaptive anchor box calculation and adaptive image resizing to optimize input on the Input side. The Focus and CSPnet structures in Backbone are primarily used to improve the learning performance of the entire convolutional neural network while drastically lowering computation; in Neck, a combination of FPN and PAN is used, combining the regular FPN layer with the bottom-up feature pyramid to fuse the extracted semantic features with the location features, and the backbone layer with the model acquires richer feature information by fusing the features with the detection layer. YOLOv5 network gives 4 versions of network pre-training models: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x in descending order of network model depth and feature map width. The lightweight YOLOv5s model is widely used in commercial scenarios because of its small data and small memory occupation. Combined with the demand for

real-time detection at construction sites, the pre-trained model chosen in this paper is YOLOv5s after comparison tests.

2.2 Improved YOLOv5 Algorithm

2.2.1 Small Target Prediction Based on YOLOv5 Network

One of the reasons why YOLOv5 does not work well for small target detection is that tiny target samples have a small sample size, and the downsampling multiplier of YOLOv5 is relatively large. As a result, learning the feature information of small targets is challenging for deeper feature maps. Therefore, we propose to add a small target detection layer to stitch the shallow feature maps with deeper feature maps for detection. For example, the network's input is 640 * 640, and YOLOv5 is downsampled 5 times, so the final feature maps of 20 * 20, 40 * 40, and 80 * 80 will be generated.

The detection mechanism of the YOLOv5 network is to detect each small square by using three anchor frames with different scale sets, and the training process is to make the anchor frames close to Ground Truth. For different resolutions, the corresponding anchor frame scales are different, which is why different resolutions can detect objects at different scales.

The biggest of the three feature maps created by the original YOLOv5 network, 80 * 80, is responsible for recognizing tiny targets while correlating to the original picture of 640 * 640; each frame of the feature map has a perceptual field of $640/80 = 8 * 8$ in size. The original network will struggle to learn the target's feature information if the target's width or height in the source picture is less than 8 pixels.

2.2.2 Improved Multi-scale Prediction Network

The network structure of YOLOv5 has been somewhat altered in order to increase the detection capabilities of the original network for tiny targets, and the revised YOLOv5 network structure is presented in Fig. 2. A small target detection layer has been built specifically to increase small target identification accuracy.

The newly added small target detection layer is mainly composed of several operational layers: after layer 17, feature maps continue to be up-sampled and other processing is done to make them more detailed. Layer 20 is used to fuse the acquired feature map of size 160 * 160 with the feature map of layer 2 in Concat to produce a larger map for small target detection. The small target detection is added in the 31st layer, and the improved entire network uses four layers [21, 24, 27, 30] for target detection.

In this paper, four prediction heads with sizes of 20 * 20, 40 * 40, 80 * 80, and 160 * 160 are designed, corresponding to the detection of feature maps at four scales: large, medium, small, and ultra-small, respectively. Compared with the original network, there is an additional output of 160 * 160. From the introduction above, we can see that among the four feature maps, the feature map with the enormous scale 160 * 160 has a minor perceptual field, and corresponding to

the original image of 640 * 640, the perceptual field of each frame of the feature map is 640/160 = 4 * 4 size, so it is responsible for detecting small target objects. If the target's width or height in the source image is around 4 pixels, the network can detect it. Compared to the original 8 * 8, the smallest range said to be detectable; the improved network can detect objects 4 times smaller.

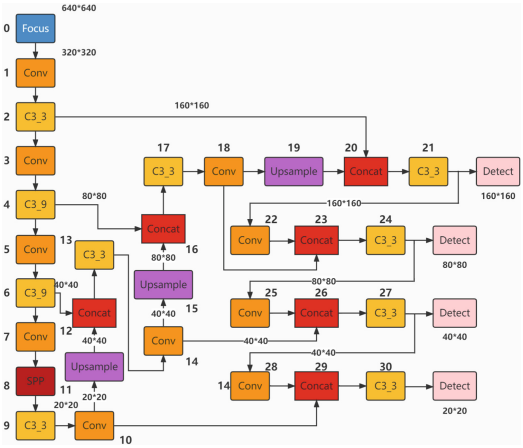


Fig. 2. Improved YOLOv5 network structure

With the addition of detection layers, there is a good improvement for small target detection. And since the pre-trained model used in this paper is YOLOv5s with the smallest amount of data, it is characterized by its fast speed. So the effect of real-time detection can be achieved.

2.2.3 CIoU_Loss

The loss function is crucial for target detection systems. The loss function can often influence the detection effect of the target detection system a lot, so the study of the loss function has been a hot topic in the field of target detection. The loss function of the YOLOv5 network consists of 3 components:

$$loss = loss_{class} + loss_{location} + loss_{confidence} \tag{1}$$

where, $loss_{class}$ is the classification loss, $loss_{location}$ is the location loss and $loss_{confidence}$ is the confidence loss.

The original YOLOv5 network uses GIoU_Loss as the calculation of $loss_{location}$. However, GIoU also has its drawbacks: when two prediction boxes are of the same height and width and at the same level, GIoU degenerates to IoU; when two boxes intersect, convergence is slow in both horizontal and vertical directions; in addition, the network regression is not accurate enough.

Here we introduce the CIoU_Loss function for calculation. CIoU_Loss considers the distance between the object and the anchor, the cross-over rate, the scale

and the punishment term to make the anticipated casing more predictable with the real edge, thus making the bounding box regression more stable and solving the problems of divergence and oscillation in the process of training when only the IOU function is used. The CIoU and CIoU_Loss are calculated as follows:

$$CIoU = IoU - (\frac{\rho^2(b, b^{gt})}{c^2} + \alpha\upsilon) \tag{2}$$

$$CIoU\_Loss = 1 - CIoU \tag{3}$$

where α and υ are calculated as follows.

$$\upsilon = \frac{4}{\pi^2}(\arctan\frac{\omega^{gt}}{h^{gt}} - \arctan\frac{\omega}{h})^2 \tag{4}$$

$$\alpha = \frac{\upsilon}{(1 - IoU) + \upsilon} \tag{5}$$

where b denotes the center point of the projection, b^{gt} indicates the center point of the formal box. $\rho^2(b, b^{gt})$ denotes the Euclidean distance between the prediction frame's centroid and the Ground Truth, c denotes the distance between the diagonal of the minimum closed area containing the predicted and actual boxes. h^{gt}, ω^{gt}, h, ω correspond to the height and width of the Ground Truth and the prediction box, respectively.

3 Experiments

3.1 Experimental Environment and Experimental Data

The experiments are based on the Linux development platform, using Ubuntu 20.04.1 operating system, and the deep learning framework used is Pytorch 1.9.0 + CUDA 10.2. The GPU used is NVIDIA GeForce RTX 2080Ti.

The helmet dataset used in this paper consists of two parts: one is from the online open source helmet dataset SHWD, and the other is the images intercepted by the cameras installed at construction sites. The SHWD dataset consists of real images of different construction scenes and is annotated by professional annotators, with high sample representativeness and authenticity. However, most of this dataset are large targets, which are easy to detect; therefore, some construction site images containing small detection targets are added to allow the model to fully learn the characteristics of small target samples and apply them smoothly under the actual construction site environment.

The final dataset has 4209 images, of which 3367(80%) are in the training set, and 842(20%) are in the test set. Before using the SHWD dataset, each image and its annotation should be preprocessed by modifying the dataset annotation XML file into a text format file required for YOLOv5 training.

3.2 Analysis of the Results

The detection results are shown in Fig. 3. It is clear that the modified algorithm performs effectively in a variety of detection settings. Whether it is a large target located slightly close to the lens or a tiny target far away from the lens, the trained model can accurately identify targets wearing helmets and those not wearing helmets.

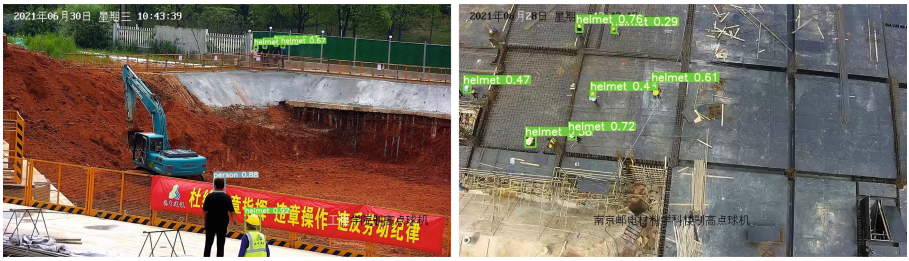


Fig. 3. Improved YOLOv5 network detection results

There are many commonly used performance metrics in target detection algorithm testing, and the main ones used in this paper are the following: Precision(P), Recall(R), Average Precision(AP), and mean Average Precision(mAP). The significance of each evaluation metric is as follows: TP is the positive samples correctly classified in the algorithm; FP is the positive samples incorrectly classified; FN is the negative samples incorrectly classified; N is the number of pictures, and NC is the target type.

Precision(P): measures the percentage of correct predictions among all outcomes with positive predictions:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

Recall(R): measures the percentage of all positive samples predicted correctly:

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

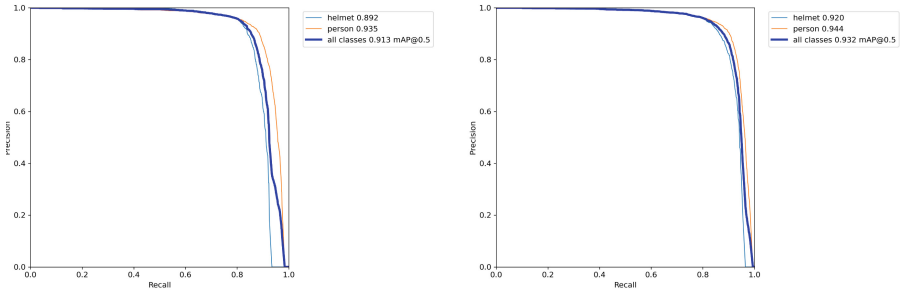
Average Precision(AP): the mean value of the accuracy of the prediction sample:

$$AP = \frac{\sum Precision}{N} \tag{8}$$

mean Average Precision(mAP): the average of the AP for all categories:

$$mAP = \frac{\sum AP}{NC} \tag{9}$$

This paper conducts comparison experiments using the YOLOv5 algorithm model and the improved YOLOv5 algorithm model. In the target detection algorithm, mAP is usually used to assess the detection accuracy of the model. As mentioned above, mAP is the average of the AP for all categories, and the AP value is the area of the P-R curve enclosed by the precision and recall. The P-R curves of the network before and after the improvement are shown in Fig. 4.



(a) YOLOv5 network (b) improved YOLOv5 network

Fig. 4. P-R curve of the network

From the above two figures, it is clear that the AP value of helmet target detection, AP value of unhelmeted target detection and mAP value increase by 2.8%, 0.9% and 1.9% respectively for the improved network compared with the original network. On the detection speed, the improved network model FPS can reach about 31 f/s, which can meet the requirements of real-time detection very well.

Table 1 reflects the performance comparison between the common target detection algorithms of today [2,3,5] and the proposed algorithms on the same dataset.

Table 1. Comparison of different detection methods on the same dataset.

Algorithm	AP/%		mAP/%	FPS f/s
	Helmet	Person		
Faster RCNN	74.5	20.1	47.3	26.3
YOLOv3	87.4	91.4	88.2	33.2
YOLOv5	89.2	93.5	91.3	34.2
Ours	92.0	94.4	93.2	30.6

Small target object picture data is picked for comparative studies in order to validate the efficiency of the improved approach for small target identification, and the experimental results are displayed in Fig. 5.

Fig. 5. Comparison of small target detection effect before and after improvement

In Fig. 5, the top row of images shows the detection results obtained by applying the improved YOLOv5 algorithm, and the bottom row shows the detection results obtained by using the original YOLOv5 network. The accuracy of the improved YOLOv5 algorithm for small target object detection is clearly higher than that of the original algorithm, indicating that the improved algorithm extracts richer detailed features and improves the phenomenon of missed and false detection, as well as the low detection rate of the original algorithm in detecting small targets, as shown by the experimental results.

4 Conclusion

This research presents a YOLOv5 network-based helmet identification technique. The detection performance of tiny helmet targets is enhanced by adding a specifically designed small target detection layer to the network and altering the original localization loss function. The improved model makes full use of the shallow feature information and incorporates the high-level semantic information to enhance the feature expression capability of the path aggregation network, which effectively improves the small target detection accuracy based on satisfying the real-time detection. The algorithm is tested on dataset such as SHWD and the performance was compared with many current mainstream object detection algorithms. The results show that the algorithm in this paper can achieve high precision and good detection effect of helmet detection while ensuring the detection speed. In conclusion, the method suggested in this research achieves small target identification of construction site helmets, which is critical for the intelligent development of construction site safety protection.

References

1. Girshick, R., Donahue, J., Darrell, T., et al.: Region-based convolutional networks for accurate object detection and segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **38**(1), 142–158 (2015). https://doi.org/10.1109/TPAMI.2015.2437384
2. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448. IEEE (2015). https://doi.org/10.1109/ICCV.2015.169
3. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031
4. Redmon, J., Divvala, S., Girshick, R., et al. You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788. IEEE (2016). https://doi.org/10.48550/arXiv.1506.02640
5. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018). https://doi.org/10.48550/arXiv.1804.02767
6. Bochkovskiy, A., Wang, C., Liao, H.: YOLOv4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020). https://doi.org/10.48550/arXiv.2004.10934
7. Kaspereulaers, M., Hahn, N., Berger, S., et al.: Short communication: detecting heavy goods vehicles in rest areas in winter conditions using YOLOv5. Algorithms **14**(4), 114 (2021). https://doi.org/10.3390/a14040114
8. Xv, S.: Research on helmet wearing detection based on improved faster R-CNN. Comput. Appl. Res. **37**(3), 901–905 (2020). https://doi.org/10.19734/j.issn.1001-3695.2018.07.0667
9. Chu, J., Leng, L.: Object detection based on multi-layer convolutional feature fusion and online hard example mining. IEEE Access **6**, 19959–19967 (2018). https://doi.org/10.1109/ACCESS.2018.2815149
10. Fang, M., Sun, T., Shao, Z.: Fast helmet wear detection based on improved YOLOv2. Opt. Precis. Eng. **27**(5), 1196–1205 (2019). https://doi.org/10.3788/OPE.20192705.1196
11. Howard, A.G., Zhu, M., Chen, B., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications (2017). https://doi.org/10.48550/arXiv.1704.04861
12. Zhang, M., Li, J.: Remote sensing image object detection technology based on improved YOLOv2 algorithm. Comput. Sci. **47**(6A), 176–180 (2020). https://doi.org/10.11896/jsjkx.191100206
13. Lin, T., Dollar, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.106
14. Fang, Q., Li, H.: Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. Autom. Constr. **85**, 1–9 (2018). https://doi.org/10.1016/j.autcon.2017.09.018
15. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger, pp. 6517–6525. IEEE (2017). https://doi.org/10.1109/CVPR.2017.690