Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 2, 2023

A modified YOLOv5 helmet detection algorithm based on Swin Transformer

Zhongling Liu, Yi Luo, Chanyi Liu *

Sichuan University of Science & Engineering, Yibin 633000, China
* **Corresponding author**: Chanyi Liu (Email: 1084400802@qq.com)

Abstract: For the current stage of helmet detection in complex environments with low accuracy, missed detection and not easy to manage wearing, this paper proposes a YOLOv5 face helmet detection algorithm based on Swin Transformer improvement from the overall semantics of the image. In this paper, experiments are conducted using a self-built dataset to further enhance the performance of the model and improve the accuracy of face helmet detection through Mosaic data enhancement, label smoothing processing, adaptive weighted features combined with Wconcat module and the application of C3TR and C3STR modules to fuse multi-scale information, enhance the feature extraction capability of the network, and improve the generalization and robustness of the model with a self-built dataset . Experiments show that the improved YOLOv5 face helmet detection algorithm mAP based on Swin Transformer improves 5.7% compared with Faster RCNN, 6.1% compared with YOLOV3, 5.3% compared with YOLOV4, and 1.6% compared with the original algorithm. It performs well in helmet face detection tasks in complex environments, achieving real-time detection and higher accuracy, while reducing missed detections.

Keywords: Safety helmet detection; YOLOv5; Swin Transformer.

1. Introduction

In recent years, the phenomenon that construction site operators do not wear helmets and cause safety problems occurs frequently [1]. The safety awareness of the workers is weak, the management personnel do not do their duty of supervision and supervision in time, and the situation of not wearing helmets as required is a huge safety hazard in construction. Therefore, the detection of the helmet wearing situation of the workers and the identification of those who are not wearing helmets play an important role in the safety management afterwards [2]. Target detection is an important topic in computer vision. With the rapid development of computer vision, deep learning methods for target detection are mainly divided into two technical paths: One stage detection algorithm and Two stage detection algorithm. 2014 Ross B. Girshick of University of California, Berkeley proposed the algorithm of R-CNN [3] to combine CNN with the field of target detection, which opened the era of two stage. In 2015, Ross Girshick of Microsoft Research proposed an improved Fast R-CNN [4] algorithm, which borrowed the structure of SPP-Net algorithm to accomplish a high-efficiency classification task based on the original research, and improved the training speed and accuracy by using multi-task error against the shortcoming of R-CNN training into multiple steps. 2015, Shaoqing Ren and Kai-Ming He of Microsoft Research In 2015, Shaoqing Ren, Kaiming He, and Ross Girshick of Microsoft Research used a Region Proposal Networks (RPN) to compute posals, and the Faster R-CNN [5] algorithm was proposed by RPN + fast R-CNN. 2017 Tsung-Yi Lin of Facebook et al. proposed the FPN algorithm, which truly incorporates the whole process of object detection into a neural network.

The two-stage approach represented by R-CNN algorithm faster almost achieves the optimal effect, but encounters a bottleneck in speed, and there is great room for improvement in real-time performance.

The YOLO algorithm proposed by Joseph Redmon et al. at the University of Washington in 2015 inherited the OverFeat algorithm, a regression-based one-stage method called YOLO (You Only Look Once), which abstracts detection as a regression problem. After improvements by Joseph Redmon et al. the YOLOv2 and YOLO9000[6] algorithms were presented at CVPR 2017, enabling end-to-end training. in 2018 Joseph Redmon et al. proposed Darknet-53 again, using three feature pyramids of different sizes to fuse features (Feature Map) making YOLOV3 [7] to achieve better classification performance. In April 2020, YOLOv4 [8] using CSP (Cross Stage Partial) Darknet-53 as Backbone and replacing the feature pyramid network with PAnet (Path Aggregation Network) was proposed by Alexey Bochkovskiy et al. proposed. In June of the same year, the YOLOv5 [9] algorithm with improvements such as Focus structure and Mosaic data augmentation has greatly improved its performance in terms of speed and accuracy, and so far YOLOv5 has been released in V6.0 with many Necks.

Recent years, more and more network models have been proposed and promoted to various detection tasks, among which the YOLO series algorithms have many advantages such as simple structure, high versatility, fast speed, applicable to multiple scenes, low background misidentification rate, etc. With the continuous iterative update of the algorithm, the algorithm performance and possibilities are being explored. Ding [10] et al. proposed a safety helmet detection method based on improved YOLOv3. For the problems of weak real-time performance and low detection accuracy in industrial site helmet wear detection, K-Means++ clustering algorithm is used to optimize the acquisition of helmet wear prior frame and introduce residual module in the network prediction stage to further improve the average accuracy. Jing [11] et al. proposed a helmet wear detection algorithm based on the improved YOLOv4 based on three feature map outputs Yue [12] et al. proposed an improved YOLOv5 based helmet wear detection algorithm based on a 128×128 feature map output with 8-fold sampling to 4-fold fusion of more small target features, to address the drawback that YOLOv5 cannot get effective features by weight focusing, using the attention module and retaining more prediction frames using the Soft-NMS algorithm to improve the accuracy of prediction frames. facility cost problem proposed a safety helmet detection method based on lightweight deep learning model using a lightweight improved version of Tiny-YOLOv3, LT-YOLO, which greatly reduces the complexity of the model.

In summary, the helmet detection wearing problem has been widely studied, however, although the above YOLO helmet detection algorithm has achieved performance optimization in different aspects, it is still difficult to maintain high accuracy in the complex background of the construction environment with overlapping crowds of multiple small and medium-sized targets. The reason is that the complex construction environment, target occlusion, light and other factors are very easy to affect the detection results, and the helmet target in the original size of the image is relatively low to cause target miss detection and background false detection.

Inspired by the design ideas of Transformer [14], which has achieved great success in CV field in 2020 (e.g., ViT for image classification, DETR model for target detection), Transformer has achieved optimal results in detection, classification and segmentation in successive fields, with the features of strong data adaptation ability and more robust global information extraction.

To address the problems of low accuracy of helmet detection in complex environments and the difficulty of construction parties to manage helmet wearing, this paper proposes a YOLOv5 face helmet detection algorithm based on the improved Swin Transformer from the overall semantics of images and collects 3000 images of faces and helmets in various environments from the web to build a helmet face detection dataset. While accurately distinguishing helmet-wearing and non-helmet-wearing people, the network's feature extraction capability is enhanced by fusing multi-scale information based on Transformer Block and sliding window mechanism, and the accuracy of face helmet detection is improved. Experiments show that the improved YOLOv5 face helmet detection algorithm mAP based on Swin Transformer improves 5.7% compared to Faster RCNN, 6.1% compared to YOLOV3, 5.3% compared to YOLOV4, and 1.6% compared to the original algorithm. It performs well in helmet face detection tasks in complex environments and achieving real-time detection with higher accuracy.

2. Improved YOLOv5 algorithms

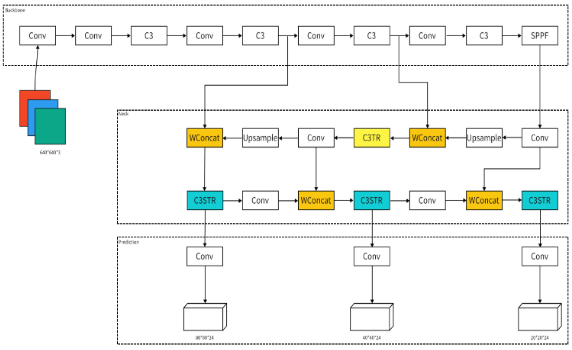

Figure 1. The improved YOLOv5 network

The Neck introduces C3TR and C3STR modules and location codes into the original network of YOLOv5, and replaces the Bottleneck in the original C3 module with Transformer Block and Swin Transformer Block modules. The retained C3 module aims to streamline the network structure of YOLOv5, reduce the computation and reduce the reasoning time of the model. On this basis, C3TR and C3STR have stronger contextual learning ability. Considering that C3TR has a strong ability to capture global information, but the number of parameters is large, it needs to calculate the self-attention of each region, and the speed is slow, while C3STR has a relatively weak ability to capture global information, but fewer parameters than C3TR, window migration can make up for the limitations of C3TR in reducing the global field of sensitivity and capturing more comprehensive features. Figure 8 shows the specific replacement method. The performance of the model is effectively improved in the complex environment where the target helmet occupies a relatively small part in the original drawing.

3. Experimental Results and Analysis

In order to verify the detection performance of the YOLOv5 algorithm combined with transformer, this paper compares it with the mainstream target detection algorithms that have emerged in recent years in the same data set and environment. The experimental results are shown in Table 1

Since there are many small and medium-sized targets in the helmet wearing detection task, Faster R-CNN as the classical representative of two-stage target detection model is not conducive to the detection of small and multi-scale objects because its feature map is only a single layer and the two stages of RPN and RCNN are divided, and it takes the longest time with the lowest accuracy. yolov5 has added more tricks, and the YOLOv5 V6.0 version algorithm used in this paper has higher accuracy before improvement, but the Recall is lower, which tends to affect the correct recognition of detection targets in complex situations. Combined with the transformer structure of YOLOv5-tr although Precision slightly decreased, but Recall and mAP@0.5 increased by 4.9% and 1.6%, respectively. In meeting the requirements of real-time detection to obtain better detection results.

Table 1. Three Scheme comparing

Model	P	R	mAP@.5
Faster R-CNN	0.940	0.891	0.925
YOLOv3	0.967	0.868	0.921
YOLOv4	0.944	0.880	0.929
YOLOv5	0.976	0.929	0.966
YOLOv5-tr	0.968	0.978	0.982

Since there are many small and medium-sized targets in the helmet wearing detection task, Faster R-CNN as the classical representative of two-stage target detection model is not conducive to the detection of small and multi-scale objects because its feature map is only a single layer and the two stages of RPN and RCNN are divided, and it takes the longest time with the lowest accuracy. yolov5 has added more tricks, and the YOLOv5 V6.0 version algorithm used in this paper has higher accuracy before improvement, but the Recall is lower, which tends to affect the correct recognition of detection targets in complex situations. Combined with the transformer structure of YOLOv5-tr although Precision slightly decreased, but Recalland mAP@0.5 increased by 4.9% and 1.6%, respectively. In meeting the requirements of real-time detection to obtain better detection results.

4. Conclusion

In this paper, we propose the YOLOv5-tr algorithm combined with transformer module on the basis of YOLOv5, and conduct experiments using the home-built dataset to increase the generalization and robustness of the model through Mosaic data enhancement, label smoothing processing, adaptive weighting features combined with Wconcat module and the application of C3TR and C3STR.Experiments show that the improved YOLOv5-tr has high accuracy for multiple occlusion detection of small and medium-sized targets in complex situations, and reduces the leakage detection while satisfying real-time detection.

Transformer undoubtedly has superior performance and powerful potential compared with convolutional neural networks. The limitation of the improvement based on transformer in this paper is that the accuracy is higher while the model is theoretically more computationally intensive, and the hardware deployment conditions are more demanding. In future work, it is expected to discuss the impact of modules in different locations on the detection results and computational effort, and to establish an identification system to determine the information related to people not wearing helmets. It is hoped that better algorithms can be borrowed for future experiments and exploration.

References

[1] Shi Hui. Research on deep learning based construction site helmet wearing detection algorithm [D]. Wuhan Institute of Technology,2019.DOI:10.27381/d.cnki.gwlgu.2019.001638.

[2] Yu Bo. Safety helmet detection based on intelligent video surveillance [D]. Hebei University of Technology, 2011.

[3] Gkioxari G , Hariharan B , Girshick R , et al. R-CNNs for Pose Estimation and Action Detection[J]. Computer ence, 2014.J.-M. Chang, W.-T. Hsiao, J.-L. Chen, H.-C. Chao, Mobile relay stations navigation-based self-optimization handover mechanism in WiMAX Networks, in: Proc. 2009 International Conference on Ubiquitous Information Technologies & Applications, 2009.B. Smith, “An approach to graphs of linear forms (Unpublished work style),” unpublished.

[4] Girshick R. Fast R-CNN[C]// International Conference on Computer Vision. IEEE Computer Society, 2015. J.G. Wilson, F.C. Fraser (Eds.), Handbook of Teratology, vols. 1-4, Plenum Press, New York, 1977-1978.

[5] Ren S , He K , Girshick R , et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[C]// NIPS. 2016. W. Strunk Jr., E.B. White, The Elements of Style, third ed., MacMillan, New York, 1979 (Chapter 4).

[6] Redmon J , Farhadi A . YOLO9000: Better, Faster, Stronger[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017:6517-6525. Cancer Research UK, Cancer statistics reports for the UK2003 (accessed 13.03.03).

http://www.cancerresearchuk.org/aboutcancer/statistics/cancerstatsreport/

[7] Redmon J, Farhadi A.YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018.M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.

[8] Bochkovskiy A , Wang C Y , Liao H . YOLOv4: Optimal Speed and Accuracy of Object Detection[J]. 2020.

[9] Wang J , Chen Y , Gao M , et al. Improved YOLOv5 network for real-time multi-scale traffic sign detection[J]. 2021.

[10] Ding W-L, Fei Sh-Min. Research on helmet detection method based on improved YOLOv3 [J]. Electronic Testing, 2022.

[11] Jin Yufang, Wu Xiang, Dong Hui, et al. Improved helmet wearing detection algorithm based on YOLO v4 [J]. Computer Science, 2021.

[12] Yue H, Huang H, Lin MH, Gao M, Li Y, Chen L. Safety helmet wearing detection based on improved YOLOv5 [J]. Computers and Modernization,2022(06):104-108+126.

[13] Qin ZH, Lei M, Song WG, Zhang W. Safety helmet detection method based on lightweight deep learning model[J]. Science Technology and Engineering,2022,22(14):5659-5665.

[14] Vaswani A , Shazeer N , Parmar N , et al. Attention Is All You Need[C]// arXiv. arXiv, 2017.

[15] Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows[J]. 2021.