Detection of construction worker safety protection equipment based on fast YOLO

Xin Li*[ac], Norriza Hussin[a], Nor Alina Ismail[b]
[a]Faculty of Engineering, Built Environment & Information Technology, SEGi University, Selangor Darul Ehsan, 547810, Malaysia
[b] Faculty of Data Science and Computing, University Malaysia Kelantan, Kota Bharu, Kelantan, 16100, Malaysia
[c] Department of Research and Development, Hebei Voin Environmental Protection Technology Co., Ltd., Hebei, 054000, China
*SUKD2101034@segi4u.my

ABSTRACT

Enterprises are often concerned about disasters, particularly accidents involving personnel. Therefore, there is ongoing research dedicated to safeguarding the safety of employees within enterprises. In recent years, with the continuous development of machine vision technology, industry and academia have been working to adopt machine vision methods to address safety hazards in workers' production. Machine vision applications are still being researched for specific sectors and lack generalization. Algorithms commonly face issues of high computational complexity and demanding hardware requirements. This paper adopts the lightweight YOLOV5 as the baseline algorithm and enhances its accuracy using a receptive field attention mechanism. SeNet is introduced to improve the generalization of object detection, and IDetect Head is employed to increase the efficiency of the detection head. Ultimately, the algorithm's accuracy is enhanced by 3.7%, and mAP50 is increased by 3.0%. This algorithm can be deployed to Internet of Things (IoT) machine vision terminals, reducing deployment costs and improving monitoring efficiency.

Keywords: Personal Safety, Machine Vision, YOLOV5, IDetect Head

1. INTRODUCTION

With the acceleration of industrialization, the issue of industrial safety has been widely discussed by all sectors of society[1]. Workers' personal protective equipment becomes particularly important, especially in high-risk environments, such as construction sites, chemical plants, and mines. These include, but are not limited to, safety helmets[2], safety glasses, protective clothing, and safety shoes. Relying solely on humans for discrimination and supervision has great limitations, such as human or intentional omissions.

Object detection is an important research field in computer vision, with wide applications in various object detection and industrial flaw detection fields[3]. In the current era of continuous development of convolutional neural networks, deep convolutional neural networks have applications in fields such as human detection, facial recognition, and vehicle detection.

Applying object detection technology to automatic detection and recognition of worker protective equipment is a popular method[4]. Through many studies, it has been found that deep network learning models have high applicability to helmets, seat belts, and other personal protective equipment. However, these studies are based on identifying label categories within a single industry.

As mentioned above, the main focus of this study is to design a deep learning-based object detection algorithm for workers' personal protective equipment. Through this algorithm, enterprises can accurately identify whether workers' personal protective equipment in various industrial enterprises is complete in real time. This equipment includes more comprehensive targets such as helmets, protective goggles, seat belts, safety shoes, etc[5]. The algorithm's detection speed and accuracy will be enhanced by improving and optimizing the existing technology, providing technical support for construction site safety management, and reducing the risk of safety accidents.

International Conference on Automation Control, Algorithm, and Intelligent Bionics (ACAIB 2024), edited by Jing Na, Shuping He, Proc. of SPIE Vol. 13259, 1325944
 · 0277-786X · doi: 10.1117/12.3039641

2. RELATED WORKS

In object detection, the evolution from Convolutional Neural Networks (CNNs) to attention mechanisms has enhanced the focus on features, improving recognition accuracy. To reduce computational load and complexity, integrating convolutional neural networks and attention mechanisms has become mainstream.

2.1 Convolutional neural network

Object detection algorithms have developed rapidly, from convolutional neural networks to attention mechanisms and the fusion of neural networks and attention mechanisms. At the beginning of CNN, object detection mainly consisted of one-stage and two-stage object detection algorithms such as Faster R-CNN[6], YOLO[7], SSD[8], etc. Among them, the fastest-developing algorithm is the first stage, especially the YOLO family, which is currently the most widely used method.

The Yolo family has evolved from the first version to the ninth version. Each version adopts the latest algorithm research results at that time. YOLOV5 is the most widely used version in the industry, adopting CSP DarkNet54 as the basic framework, using base anchor for boundary calculation, and using SIoU loss function for loss calculation[9]. The optimizer includes SGD AdmaW, etc. The detection accuracy has been increased by further reducing the number of network layers. YOLOv5 is currently one of the most popular and widely used object detection algorithms, with good performance in industrial safety, autonomous driving, and other application scenarios.

2.2 Attention mechanism

Attention mechanisms in object detection can be mainly categorized into spatial attention mechanisms, channel attention mechanisms, multi-scale attention mechanisms, and receptive field attention mechanisms. Among them, Spatial attention[10] mechanisms primarily focus on capturing the important information of different positions in the image. Channel attention mechanisms[11] aim to highlight the importance of varying channel features. Multi-scale attention mechanisms focus on the critical target information at different scales. Receptive field attention mechanisms[12] aim to study the adaptive adjustment of receptive field size to improve the detection accuracy for targets of different sizes.

2.3 Feature Pyramid Network

The feature pyramid is mainly used in computer vision to solve multi-scale image feature extraction[13], which uses multiple downsampling methods to extract features from image data. By using a feature pyramid network, the recognition accuracy of targets of different sizes in object detection algorithms can be improved, the ability to handle target diversity can be enhanced, and the application range of the algorithm can be expanded.

3. PROPOSED METHOD

3.1 Improved backbone model

The algorithmic model framework of this study is based on improvements made to YOLOv5, where we use RFCAConv layers[12] instead of Conv layers for convolution in the backbone network while still employing C3 for feature fusion. In the head part, we have upgraded the Concat module to a BiFPN[14] feature pyramid structure, enhancing the ability to recognize backgrounds. We introduced the SeNetV2 module between Backbone and Head, enhancing channel attention's influence on feature values. In the detection head, IDetect has been used instead of Detect, improving the accuracy of the detection head and its capability to detect small objects. The structural diagram is shown in Figure 1.

A convolutional neural network is a fast visual processing algorithm, and the attention mechanism is a method to improve the accuracy of local features, including spatial attention, channel attention, etc., in attention classification. The method enhances spatial attention limitations through receptive field attention. RFA can be designed as a convolutional calculation method to replace standard convolution. RFAConv learns attention through the interaction of receptive field feature information, ultimately improving network performance. AvgPool is used to aggregate receptive field feature information to balance computation and cost. Finally, softmax is used to emphasize the importance of each feature, and the calculation of RFA is represented as:

$$\begin{aligned} F &= Soft\max\left(g^{1\times 1}\left(AvgPool\left(X\right)\right)\right)\times \mathrm{Re}LU\left(Norm\left(g^{k\times k}\left(X\right)\right)\right) \\ &= A_{rf}\times F_{rf} \end{aligned} \tag{1}$$

Where $g^{i\times i}$A represents a grouped convolution of size$i\times i$, K represents the size of the convolution kernel, Norm represents normalization, X represents input feature mapping, and F is obtained by multiplying the attention map A_{rf} with the transformed receptive-field spatial feature F_{rf}.

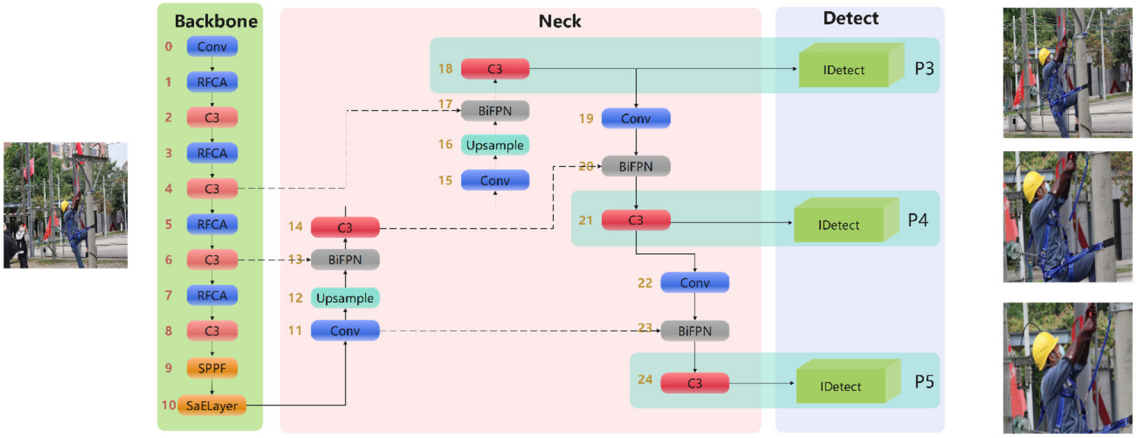

Figure 1 Improved YOLOV5 network structure.

Perform group convolution on the target to obtain multiple sets of feature values, conduct Norm processing and calculate ReLU loss function accuracy on each group, interact with the feature values, and average the values in the H and W directions of the interacting data. Finally, connect and convolve these feature values.Separate the feature values and perform convolution and activation separately. Perform weight processing on each grouped feature value, and finally perform convolutional output.As shown in Figure 2.

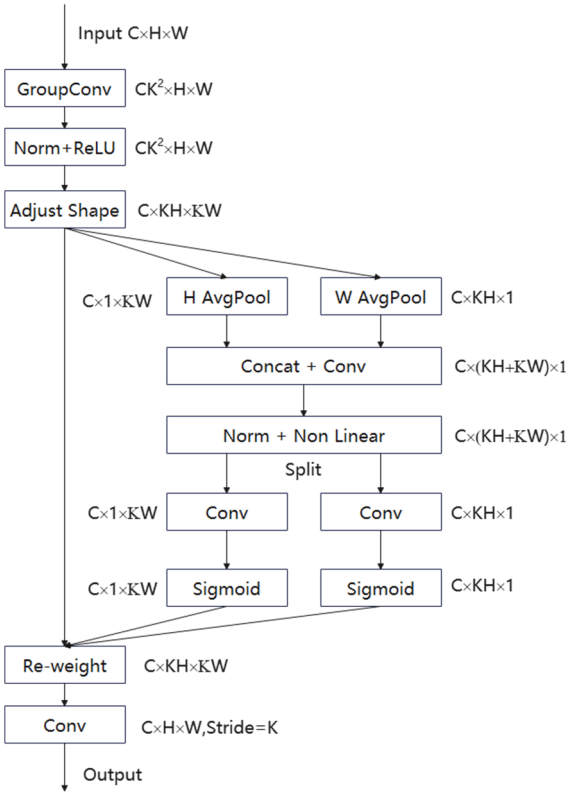

Figure 2 Structure of RFCAConv.

3.2 Improving attention model

SENet has increased the channel representation ability through squeezing and excitation operations, significantly improving the feature value expression ability. The SENet module has added an aggregation module, resulting in the SaENet module. SENet has increased the channel representation ability through squeezing and excitation operations, significantly improving the feature value expression ability. The SENet module has added an aggregation module, resulting in the SaENet module. The SaENeT module is similar to the ResNeXt network for multi-branch aggregation. The structural diagram is shown in Figure 3. SaENet can capture more complex spatial features through multiple convolution operations, enhancing its ability to detect targets in responsible environments.

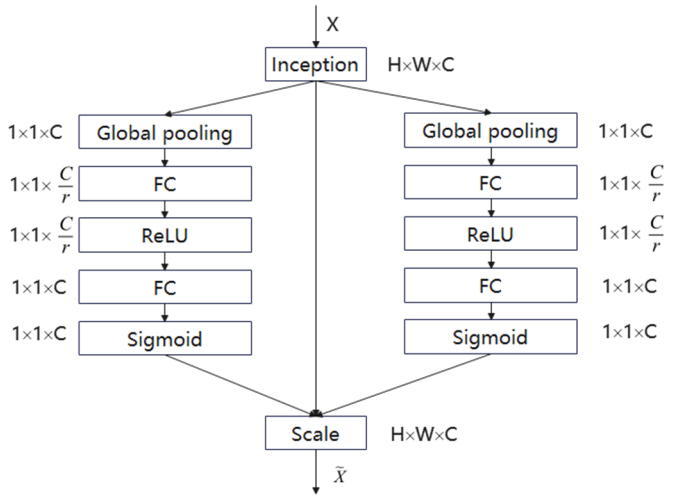

Figure 3 SaENet architecture.

The function expression of SaENet is as follows:

$$SaENet = x + F\left(x \cdot E_x\left(\sum Sq(x)\right)\right) \tag{2}$$

3.3 Improving neck model

The feature pyramid is an intermediate component that connects the backbone and neck sections, playing a crucial role in multi-scale feature fusion for addressing target detection at various resolutions. BiFPN integrates bidirectional cross-scale connections and rapid normalization for fusing features.The structural diagram is shown in Figure 4.

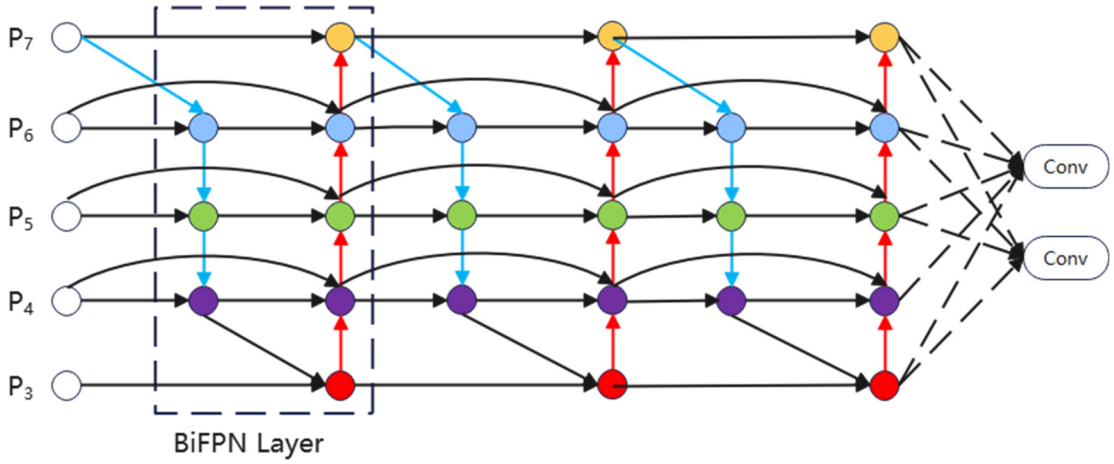

Figure 4 The Application of BiFPN in Models.

Taking the sixth layer feature fusion as an example, the formula is as follows.

$$P_6^{td} = Conv\left(\frac{w_1 \cdot P_6^{in} + w_2 \cdot \mathrm{Re}size(P_7^{in})}{w_1 + w_2 + \epsilon}\right)$$
$$P_6^{out} = Conv\left(\frac{w_1' \cdot P_6^{in} + w_2' \cdot P_6^{td} + w_3' \cdot \mathrm{Re}size(P_5^{out})}{w_1' + w_2' + w_3' + \epsilon}\right) \quad (3)$$

3.4 Introducing a new detection head module

The IDetect detection head of YOLOv7[15] was introduced in YOLOv5. The IDetect detection head is a high-speed and high-precision detection method that can perform multi-target localization in real-time object detection and grid images and predict each grid's target category and position, achieving object detection tasks.

4. EXPERIMENT AND ANALYSIS

The dataset used in this experiment combines an internally acquired dataset and a publicly available dataset, divided into 70% for training, 20% for validation, and 10% for testing. This dataset includes 23 categories, including Ear protection, Eyeglasses, Face_shield, Goggles, Hazmat_suits, Head, Helmet, Mask, Person, Reflective_vest, Respirator, Safety_harness, Unprotected_body, Unprotected_ear, Unprotected_eye, Unprotected_face, Unprotected_hand, Unprotected_head, Work_coverall, Work_trousers, Worke_gloves, Worke_jacket, Worke_shoes. We want to establish an industry-wide employee safety protection device testing system that applies to all enterprises, not only the construction industry, power industry, high-altitude operations, etc. An example dataset is shown in Figure 5.

Figure 5 Dataset examples.

The experiment adopts the Windows 11 operating system, with an Intel Xeon E3-1225 processor, 32GB of memory, and an Nvidia Geforce 3090 graphics card. The algorithm training uses the AdamW optimizer, with an initial learning rate set to 0.001, and employs a cosine scheduling strategy for learning rate adjustment. The weight is set to 0.01, the training batch size is 56, and the training is conducted for 300 epochs. Data augmentation is not used in the last ten epochs. We compared commonly used object detection algorithms and obtained results, as shown in Table 1.

Table 1. Margins and print area specifications.

Model	P/%	R/%	mAP50/%	mAP95/%	GFLOPS	Size/MB	Par/MB
YOLOV5S	69.5	59.8	58.9	33.9	15.9	14.6	7.07
EfficientNet	65.6	53.5	55.2	29.0	2.7	2.6	1.10
MobileNetV3	62.0	49.7	51.5	25.7	2.6	3.2	1.41
ShuffutNet	57.0	43.0	44.4	20.3	1.9	2.1	0.87
FasterNet	68.2	55.0	56.6	29.9	11.9	11.6	5.60
Ours	73.2	59.6	61.9	34.5	17.1	15.0	7.35

From Table 1, it can be seen that our algorithm has significant advantages in accuracy, recall, and other performance aspects. Compared with the baseline algorithm, there is no significant increase in parameter count, computational complexity, and model weight. To compare the effects of RFCAConv, SaENet, IDetect, and other modules on the algorithm, corresponding ablation experiments were conducted, and the results are shown in Table 2.

Table 2 Ablation experiments of our model.

Baseline	RFCAConv	SaENet	BiFPN	IDetect	P/%	mAP50/%	GFLOPS	Size/MB
√					69.5	58.9	15.9	14.6
√	√				70.5	61.1	16.4	14.7
√		√			70.4	58.8	16.0	14.7
√			√		72.1	60.4	16.6	14.9
√				√	72.5	60.9	15.9	14.4
√		√	√		69.8	59.3	16.6	15.0
√		√		√	71.1	60.9	16.6	14.5
√		√	√	√	71.4	61.1	16.6	14.8
√			√	√	70.8	58.0	16.6	14.7
√	√	√			70.3	61.2	16.5	14.8
√	√		√		72.2	59.6	17.1	15.0
√	√			√	72.9	60.0	16.4	14.5
√	√		√	√	71.7	59.8	17.1	14.8
√	√	√		√	71.7	60.8	16.5	14.7
√	√	√	√		71.8	61.8	17.1	15.1
√	√	√		√	72.8	61.8	16.5	14.7
√	√	√	√	√	73.2	61.9	17.1	15.0

Compared to the baseline YOLOV5S, the method proposed in this paper has increased the precision by 3.7% and improved the mAP50 by 3.0%. As seen in Figure 6, the results demonstrate that the method can achieve a comparatively high level of object detection outcomes.

Figure 6 Results of worker safety equipment target detection.

5. CONCLUSIONS

The training model improved by our algorithm has certain advantages in detecting worker safety equipment in complex backgrounds, especially in achieving an accuracy improvement of 3.7% without changing the baseline algorithm's model weight and computational complexity, and a 3.0% improvement in mAP50 detection. By judging the results, we can conclude that this algorithm can play its role in this field and contribute to reducing employee accidents.

In algorithm development, there are urgent problems that need solving. For instance, employees' work and tools are still integrated.In addition to detection, equipment, tools, limbs, and movements should also be related so that the algorithm can provide emergency avoidance reminders for employees' work steps. Our subsequent research will focus on improving the relevant issues.

ACKNOWLEDGMENT

This work is funded by Science Research Project of Hebei Education Department under Project,the number is QN2024166.

REFERENCES

[1] Minsoo Park, Dai Quoc Tran, Jinyeong Bak, Seunghee Park. Small and overlapping worker detection at construction sites. Automation in Construction. 2023;151:104856. doi:10.1016/j.autcon.2023.104856

[2] Yu-Ci Chen, Wen-June Wang. Safety Helmet Wearing Detection System Based on a Two-Stage Network Model. In: 2023 5th International Conference on Computer Communication and the Internet (ICCCI). IEEE; 2023:122-126. doi:10.1109/ICCCI59363.2023.10210093

[3] Zhihui Xie, Min Fu, Xuefeng Liu. Detection of Fittings Based on the Dynamic Graph CNN and U-Net Embedded with Bi-Level Routing Attention. Electronics. 2023;12(22):4611. doi:10.3390/electronics12224611

[4] Lyuchao Liao, Linsen Luo, Jinya Su, Zhu Xiao, Fumin Zou, Yuyuan Lin. Eagle-YOLO: An Eagle-Inspired YOLO for Object Detection in Unmanned Aerial Vehicles Scenarios. Mathematics. 2023;11(9):2093. doi:10.3390/math11092093

[5] Hao Hu, Guo,Fang, Zhao Liu. Object Detection Based on Improved YOLOX-S Model in Construction Sites. Journal of Frontiers of Computer Science and Technology. 2023;17(5):1089-1101.

[6] Girshick R. Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV). IEEE; 2015:1440-1448. doi:10.1109/ICCV.2015.169

[7] Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016:779-788. doi:10.1109/CVPR.2016.91

[8] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector. In: Vol 9905. ; 2016:21-37. doi:10.1007/978-3-319-46448-0_2

[9] Mohammad Hossein Hamzenejadi, Hadis Mohseni. Fine-tuned YOLOv5 for real-time vehicle detection in UAV imagery: Architectural improvements and performance boost. Expert Systems with Applications. 2023;231:120845. doi:10.1016/j.eswa.2023.120845

[10] Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, et al. CBAM: Convolutional Block Attention Module. In: Computer Vision – ECCV 2018. Vol 11211. Lecture Notes in Computer Science. Springer International Publishing; 2018:3-19. doi:10.1007/978-3-030-01234-2_1

[11] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, Qinghua Hu. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. Published online April 7, 2020. Accessed March 29, 2024. http://arxiv.org/abs/1910.03151

[12] Zhang X, Liu C, Yang D, et al. RFAConv: Innovating Spatial Attention and Standard Convolutional Operation. Published online March 28, 2024. Accessed April 10, 2024. http://arxiv.org/abs/2304.03198

[13] Jishen Peng, Wenkun Shi, Haiming He, Liye Song. Improved Yolov5S normative identification algorithm of power place operators. In: 2022 2nd International Conference on Electrical Engineering and Control Science (IC2ECS). IEEE; 2022:942-950. doi:10.1109/IC2ECS57645.2022.10088065

[14] Tan M, Pang R, Le QV. EfficientDet: Scalable and Efficient Object Detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2020:10778-10787. doi:10.1109/CVPR42600.2020.01079

[15] Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2023:7464-7475. doi:10.1109/CVPR52729.2023.00721