electronics

MDPI

Article

An Unsafe Behavior Detection Method Based on Improved YOLO Framework

Binbin Chen [1], Xiuhui Wang [1,*], Qifu Bao [2,*], Bo Jia [2], Xuesheng Li [2] and Yaru Wang [2]

[1] Department of Computer Science and Technology, China Jiliang University, Hangzhou 310018, China; P20030854004@cjlu.edu.cn
[2] Key Laboratory of Safety Engineering and Technology Research of Zhejiang Province, Hangzhou 310027, China; bjia@cjlu.uu.me (B.J.); xsli@cjlu.uu.me (X.L.); yrwang@cjlu.uu.me (Y.W.)
* Correspondence: wangxiuhui@cjlu.edu.cn (X.W.); qfbao@cjlu.uu.me (Q.B.)

Abstract: In industrial production, accidents caused by the unsafe behavior of operators often bring serious economic losses. Therefore, how to use artificial intelligence technology to monitor the unsafe behavior of operators in a production area in real time has become a research topic of great concern. Based on the YOLOv5 framework, this paper proposes an improved YOLO network to detect unsafe behaviors such as not wearing safety helmets and smoking in industrial places. First, the proposed network uses a novel adaptive self-attention embedding (ASAE) model to improve the backbone network and reduce the loss of context information in the high-level feature map by reducing the number of feature channels. Second, a new weighted feature pyramid network (WFPN) module is used to replace the original enhanced feature-extraction network PANet to alleviate the loss of feature information caused by too many network layers. Finally, the experimental results on the self-constructed behavior dataset show that the proposed framework has higher detection accuracy than traditional methods. The average detection accuracy of smoking increased by 3.3%, and the average detection accuracy of not wearing a helmet increased by 3.1%.

Keywords: behavior detection; YOLO; ASAE; WFPN

Citation: Chen, B.; Wang, X.; Bao, Q.; Jia, B.; Li, X.; Wang, Y. An Unsafe Behavior Detection Method Based on Improved YOLO Framework. *Electronics* **2022**, *11*, 1912. https://doi.org/10.3390/electronics11121912

Academic Editor: Marcin Witczak

Received: 19 April 2022
Accepted: 17 June 2022
Published: 20 June 2022

1. Introduction

Frequent accidents and casualties reflect the problem of the untimely discovery of accidents in current industrial production management. In particular, accidents occur from time to time due to operators' incorrect wearing of safety helmets, smoking, and other violations. Therefore, it is of great practical significance to apply target detection and behavior recognition technology to the safety supervision of industrial production and construction, to effectively protect the safety of operators and reduce economic losses caused by accidents.

Moving target detection and behavior recognition [1,2] is one of the basic tasks in the field of computer vision, and it is also the core task of video surveillance. However, target detection is still one of the most challenging directions in the field of computer vision because of the diversity of target posture, the irregularity of action, the complexity of scene, the resolution of camera and the transformation of lighting conditions [3,4]. Existing target-detection algorithms can be divided into traditional target-detection algorithms and target-detection algorithms based on deep learning. The processing process of traditional target-detection algorithm generally includes the following four steps: (1) image preprocessing, including image loading, image noise reduction, image color space conversion and image morphological operation; (2) features (shape, texture and color) in the sliding window being extracted by moving the sliding window with a specified size in the image to be detected, (3) the AdaBoost classifier [5] and SVM classifier [6] being used to judge whether there is a target of interest in the sliding window, and (4) overlapping sliding windows containing the same target being merged to obtain the bounding box of the final target. However,

because the traditional target-detection algorithm detects the target object through hand-crafted features, it leads to the problem of the low accuracy of target detection and poor generalization of the model [7,8].

On the other hand, a target-detection algorithm based on deep learning is a promising research direction. In 2014, Ross Girshick et al. proposed the R-CNN network model [9] and introduced deep learning into target detection for the first time, which is a milestone in the research direction of target detection. Subsequently, Ross Girshick [10] proposed the Fast R-CNN model, which means that the target-detection algorithm based on deep learning has developed into an end-to-end network model. Recently, an improved R-CNN model [11], named OCR-R-CNN, is presented for elevator button recognition, which consists of a region-based convolutional neural network (R-CNN)-based button detector and an attention-RNN-based character recognizer. Aiming at recognizing insulators and detecting faults timely and accurately, an insulator recognition and fault-detection model was proposed in [12], which is based on the faster region convolutional neural network (R-CNN) and feature pyramid networks (FPNs). The main difference between R-CNN and traditional target-detection algorithms is the use of deep neural network to extract image features, and their common point is that they use AdaBoost and SVM classifiers to classify them. Target-detection algorithms based on deep learning can be further divided into region-proposal-based models and region-proposal-free models. Typical representatives of the model based on region proposal include R-CNN, SPP_Ne [13], Fast R-CNN, Faster R-CNN [14], etc. Region-proposal-free models, such as SSD [15] and YOLO [16], are different from R-CNN series algorithms, which directly carry out intensive sampling at different positions within the target picture, and then use a convolutional neural network to directly extract features for classification and regression in one step. In practical application, the above two models have their own advantages and disadvantages in the detection task. Region-proposal-free models represented by SSD and Yolo series have a faster detection speed, but their accuracy is not as good as that represented by the Fast R-CNN series. In addition, in the process of target feature extraction, the continuous down-sampling operation of the region-proposal-free model will cause the loss of detailed features, resulting in incomplete effective information during detection, making it unable to achieve the effect of deep learning and mainstream detection, and difficult to meet the needs of detection and recognition. Moreover, the R-CNN-based methods cannot make full use of the context information of local objects in the whole picture after transforming the detection problem into the classification problem of local areas of the picture.

In view of the above problems, this paper proposes an improved YOLO network to detect unsafe behaviors such as not wearing safety helmets and smoking in industrial places. The main contributions of our work include the following aspects:

- A novel adaptive self-attention embedding (ASAE) model. By reducing the number of feature channels, the AAE model can improve the backbone network and reduce the loss of context information in the high-level feature map.
- A new weighted feature pyramid network (WFPN) model. The proposed WFPN model is used to replace the original enhanced feature-extraction network PANet to alleviate the loss of feature information caused by too many network layers.
- More accurate boundary box detection. In the process of boundary box detection, through the optimization of loss function, the regression accuracy of the boundary box is improved, and the cases of missed detection and false detection are reduced.

2. Methodology

As shown in Figure 1, we propose an unsafe behavior recognition network based on YOLO framework, called YOLO-AW. The input images are 300 × 300 in size, which contain the marked smoking or helmet-wearing behavior information. The backbone network of the proposed network adopts CSPDarknet53 [17] +ASAM, in which the three residual network blocks in CSPDarknet53 are shown as CSP1_1, CSP1_2 and csp1_3 in Figure 1. Because the multiple convolution operations in the residual module may lead to

the disappearance of target details and edge information, and then affect the target location, we add an adaptive ASAE model to the second residual module, which reduces the loss of context information in the high-level feature map caused by the reduction of feature channels and improves the sensitivity of the network to unsafe behavior detection. Second, a new feature-fusion module WFPN is added to the network neck to obtain more efficient semantic feature information about the upper and lower layers, to improve the accuracy of the target-detection algorithm. Third, using Alpha-IoU [18] to replace the original GIoU index, the high or low-IOU target loss and target gradient are adaptively reweighted to improve the regression accuracy of the bounding box. In addition, DIoU-NMS is used to replace the traditional NMS. By considering the overlapping area and the center distance of the two frames, the error suppression often caused by occlusion is reduced. No matter of whether it is a clean or noisy environment, no additional parameters are introduced, and the training time is not increased. Finally, the proposed network obtains the detection output of three different scales.

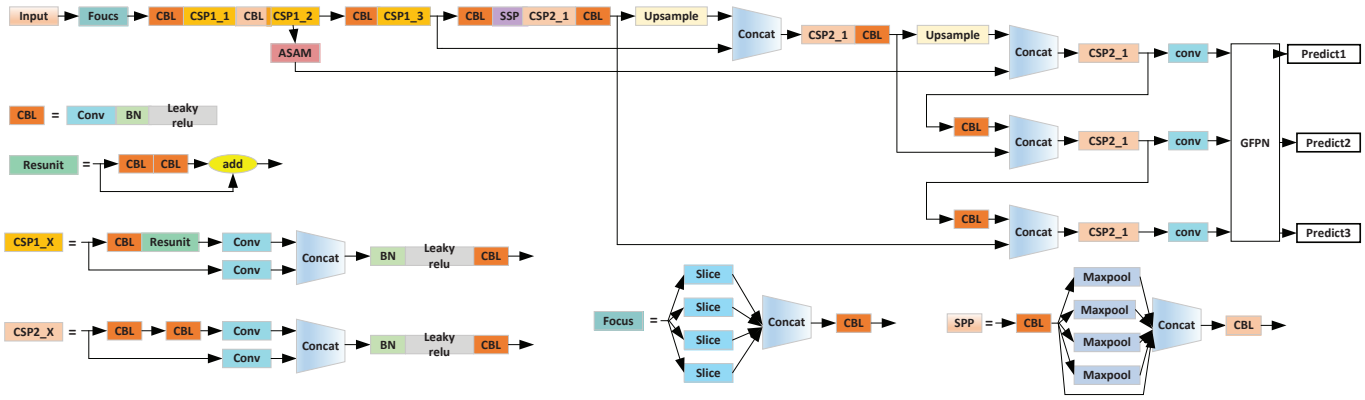

Figure 1. Architecture of the proposed YOLO-AW network.

2.1. Adaptive Self-Attention Embedding Module

To better integrate more high-quality upper and lower semantic feature information in each feature map, it is necessary to further improve the ability of network feature extraction. Inspired by ResNet [19] and self-attention [20], this paper constructs an adaptive ASAE module. The operation of the ASAE module can be divided into two steps. First, multiple context features of different scales are obtained through an adaptive average pool layer, and the targets in the dataset are adaptively converted to the specified output size. Then, one spatial weight map is generated for each feature map through the spatial attention mechanism. Through this weight graph, the context features are fused to generate a new feature graph containing multi-scale context information. To prevent the loss of features in the process of propagation, we combine the new feature map with the adaptive feature map and the original high-level feature map, and finally obtain a new feature map through feature fusion.

The specific structure of the ASAE module is shown in Figure 2. First, the input data passes through the input layer $C3$ and the adaptive pool layer to obtain three context features with different scales ($\beta 1 \times H \times W$, $\beta 2 \times H \times W$, $\beta 3 \times H \times W$), which are tagged as $C31$, $C32$ and $C33$, respectively. Then, each context feature is a 1×1 convolution to obtain the same channel dimension 256. Finally, the convolution results are sampled up to $H \times W$ scale by bilinear interpolation for subsequent feature fusion.

Given the input feature map $C3$, its length, width and number of channels are H, W and C, respectively. After adaptive pooling and up-sampling operations, three new feature maps, $C31$, $C32$ and $C33$, are obtained, respectively.

$$C31 = ups_1(Avg_1(C3)) \tag{1}$$

$$C32 = ups_2(Avg_2(C3)) \tag{2}$$

$$C33 = ups_3(Avg_3(C3)) \tag{3}$$

Then, the context features $C31$ and $C32$ are fused through the Concat layer, and the obtained feature map passes through 3×3 convolution layer, ReLU activation layer, 1×1 convolution layer and Sigmoid activation layer in turn to generate corresponding spatial weights for each feature mapping as:

$$w_i = \rho(conv_1(\gamma(conv_3(cat(C31, C32))))) \tag{4}$$

where $cat(.)$ is the channel fusion function, $conv_3(.)$ is the 3×3 convolution operation, $\gamma(.)$ is the RELU activation function, $conv_1(.)$ is the 1×1 convolution operation, and $\rho(.)$ is the Sigmoid activation function.

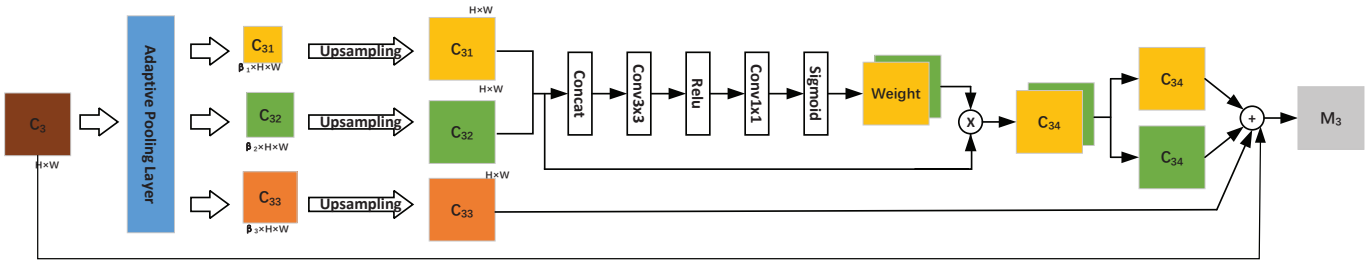

Figure 2. Architecture of the proposed ASAE module.

Next, $C35$ is obtained by multiplying the generated weight map and the feature map after merging channels, and then $C35$ is separated according to channels and combined with the context feature $C33$ and the input feature map $C3$ in turn. This operation makes the final feature map have rich multi-scale context information, and alleviates the information loss caused by the reduction of the number of channels to a certain extent. The calculation process is as follows:

$$C35 = C35 \bigotimes w_i \tag{5}$$

$$M3 = C33 + C35[:C] + C35[C:2C] + C3 \tag{6}$$

where $C35[:C]$ and $C35[C:2C]$ are the feature maps obtained by separating the obtained feature map $C35$.

2.2. Weighted Feature Pyramid Network Module

In practical application scenarios, due to different collection angles and objects of concern, the violations to be detected are also different, resulting in the different importance of different scale feature maps. Therefore, this paper improves PANet [21] to obtain a weighted feature-fusion structure WFPN. The improved structure can better extract significant features to improve the accuracy of a target-detection algorithm.

As shown in Figure 3, the neck of the original YOLOv5 network uses PANet for multi-scale fusion and outputs feature images. Based on FPN, PANet adds a bottom-up enhancement structure, and changes the original one-way fusion to two-way fusion, which consists of two parts: a top-down integration path and bottom-up integration path. As shown in Figure 4, the bottom-up fusion method uses the nearest-neighbor method to up-sample the feature map M twice, and adds the feature map M to the feature map N of the previous layer after 1×1 convolution. The PANet network combines the strong semantic information of the high-level feature map with the positioning information of the low-level feature map, and uses the accurate low-level positioning signal to enhance the whole feature level, to shorten the information path between the low-level and top-level features. However, in the fusion process, PANet directly adds the features of different levels without considering their unequal contribution to the final output.

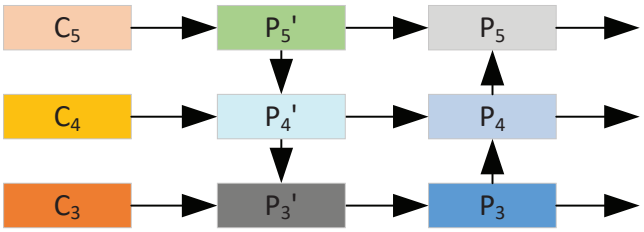

Figure 3. Architecture of the PANet network.

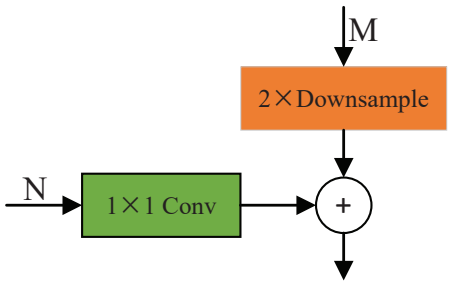

Figure 4. The bottom-up fusion method.

To make the network more efficient and enable the network to learn the weights of different input features, the WFPN module proposed in this paper adopts the idea of two-way fusion, constructs top-down and bottom-up two-way channels to obtain multi-scale information, and unifies the feature resolution scale through up-sampling and down-sampling when fusing between different scales, so as to achieve more effective feature fusion. At the same time, the nodes with only one input edge and output edge in the PANet network are removed, and when the input and output nodes are at the same level, an additional edge is added using the residual method, and the weighted fusion method is used to fuse the feature layers of different resolutions, which can fuse more features without increasing the computational overhead. Through the multi-scale feature fusion of cross level connection and same-level jump connection, the module makes full use of the characteristics of violations in industrial sites of different scales to improve the accuracy of information processing.

As shown in Figure 5, first, the features of C3 to C5 layers of the backbone network are transferred to WFPN as the selected multi-scale feature layer, and a convolution layer is used to extract significant features and reduce the feature dimension. Second, the top-down feature fusion is carried out by up-sampling, and the high-level features are fused with the low-level feature information through up-sampling. Thirdly, the fused features are analyzed by a 3×3 convolution operation to eliminate the aliasing effect. Finally, the top-down feature fusion is completed through maximum pooling, and the obtained features are fused with high-level features to obtain three output features, P3, P4 and P5, with different scales.

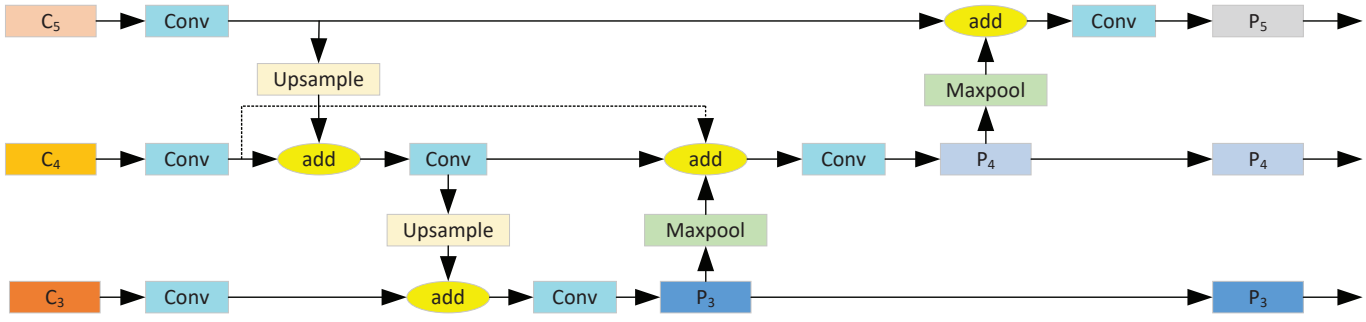

Figure 5. Architecture of the proposed WFPN module.

2.3. Boundary Box Detection

The classification results of the YOLO-AW network proposed in this paper include three aspects: the prediction category, confidence, and location of each prediction box. Therefore, when constructing the loss function, it is necessary to evaluate the above components, calculate the category error, confidence error and position error of the prediction results, respectively, and obtain the overall loss function by weighting. The *loss* function of the prediction box is defined as:

$$loss = loss_{location} + loss_{conf} + loss_{class} \tag{7}$$

where $loss_{location}$ is the loss of location, $loss_{conf}$ is the loss of confidence, and $loss_{class}$ is the loss of category. In YOLOv5, the loss location is calculated by GIoU loss function, but the aspect ratio of the bounding box is not considered in the regression process. When the two boxes intersect, the convergence is slow in the horizontal and vertical directions.

To solve the above problems, this paper uses α-IoU to replace the original GIoU and uses parameter α to adaptively readjust the loss and gradient of IoU target, so that the detector has greater flexibility in realizing different levels of bounding box regression accuracy. α-IoU is defined as follows:

$$\iota_{\alpha-IoU} = \frac{1 - IoU^{\alpha}}{\alpha}, \alpha > 0 \tag{8}$$

In addition, in traditional NMS, the IoU index is often used to suppress redundant prediction-bounding box detection. Because occlusion often produces error suppression, overlapping areas need to be considered. Therefore, the YOLO-AW network proposed in this paper takes DIoU as the criterion of NMS, and considers the overlapping area and the center distance of the two frames at the same time. The set of prediction boxes is obtained through the network model, then the scores of all boxes are sorted, the box with the highest score is selected, the box with the highest score and the remaining boxes in the set are subjected to non-maximum suppression operation based on DIoU, and the box with the highest score is continuously selected, and finally the detection result is obtained.

3. Experiments

The development environment of the experiment in this section is Anaconda 1.7.2 plus Pytorch 1.2.0. The experimental platform is Intel® Xeon® CPU E5-2630 v4 (10 core, 2.4 GHz), equipped with two NIVIDA® GeForce® RTX 2080 Ti graphics cards (11 GB video memory) and Centos® 7.9.2009 operating system. In the experiment, the initial learning rate is set to 1×10^{-3}, the Adam optimizer is used, the epoch is set to 200, and the batch size is set to 4.

3.1. Datasets and Evaluation Criteria

Since no public datasets are available, we established a dataset containing smoking and helmet-wearing behavior for algorithm testing. Most of these image data come from the video surveillance of the industrial scene, and the other part comes from video and movies. Among them, the image data from the industrial site is collected by the surveillance camera and normalized to the size of 300×300. The production of the dataset included four steps: data collection and sorting, data preprocessing, data filtering, and dataset annotation. In this work, 1000 pieces of data were created manually and expanded to 4000 through data enhancement. The dataset is divided into a training set and test set in the ratio of 9:1, and samples of different resolutions and sizes are averaged to ensure that images of different sizes can be fully trained and tested. In the segmentation of training samples and test samples, we use the cross-validation method to ensure the fairness and effectiveness of the test. The experimental dataset has been submitted as an attachment to the manuscript.

In addition, this paper uses the commonly used evaluation criteria *AP* (average accuracy), *mAP* (mean of average accuracy) and *Recall* to quantitatively evaluate the effect of the target-detection network, which is defined as follows:

$$precison = \frac{TP}{TP + FP} \tag{9}$$

$$recall = \frac{TP}{TP + FN} \tag{10}$$

$$AP = \int_0^1 precison(recall) \tag{11}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{12}$$

where TP is the positive sample with correct classification, FP is the positive sample with wrong classification, N is the number of types of target objects, and AP_i is the AP value of the class i.

3.2. Comparative Experiments

In this section, the proposed YOLO-AW method is compared with other mainstream target-detection methods, and the experimental results are quantitatively analyzed using AP, mAP and Recall criteria, as shown in Table 1. To test whether the performance improvement is significant, we also conducted a paired *t*-test and provided the statistical results in Table 2.

Table 1. Comparison of recognition results of different methods (IoU = 0.5).

Method	AP (Safety Helmet)	AP (Smoking)	mAP	Recall
Faster-R-CNN	0.697	0.667	0.682	0.583
YOLOv4	0.75	0.686	0.718	0.697
YOLOv5	0.751	0.719	0.735	0.638
YOLO-AW	0.782	0.752	0.767	0.689

Table 2. Paired *t*-test results of different methods.

Method	Mean	Standard Deviation	Standard Error of Difference
Faster-R-CNN	0.6573	0.0510	0.0255
YOLOv4	0.7128	0.0282	0.0141
YOLOv5	0.7108	0.0502	0.0251
YOLO-AW	0.7475	0.0409	0.0204

It can be seen from Tables 1 and 2 that the AP of the method proposed in this paper increased by 3.3% in smoking detection, 3.1% in helmet-wearing detection, 3.2% in map and 5.1% in recall. This is mainly because the proposed network uses a new ASAE model to reduce the loss of context information in the high-level feature map, and replaces the original enhanced feature-extraction network PANet by a new WFPN network to alleviate the loss of feature information caused by too many network layers.

Furthermore, to verify the improvement effect of the loss function in this paper, we conducted comparative experiments on different types of IoU loss functions, and the results are shown in Figure 6 and Table 3. It can be seen that compared with the traditional IoU algorithm, the α-DIoU (α = 2) loss function introduced in this paper is superior to the existing IoU loss functions, which improves the late training (e.g., after 150 epochs) by increasing the gradient of high-IoU objects, but has little negative impact on the early training (e.g., after the first 100 epochs). This feature helps to reduce the gradient of low-IoU objects in the early training stage, i.e., it stabilizes the model training when the early gradient is large, and provides stronger robustness for small datasets and noise boxes.

Table 3. Comparison of AP results corresponding to different IoU (α = 2).

Cases	GIoU	α-DIoU	α-CIoU	α-GIoU
Safety Helmet	0.751	0.777	0.738	0.771
Smoking	0.719	0.707	0.69	0.679
All	0.735	0.742	0.714	0.725

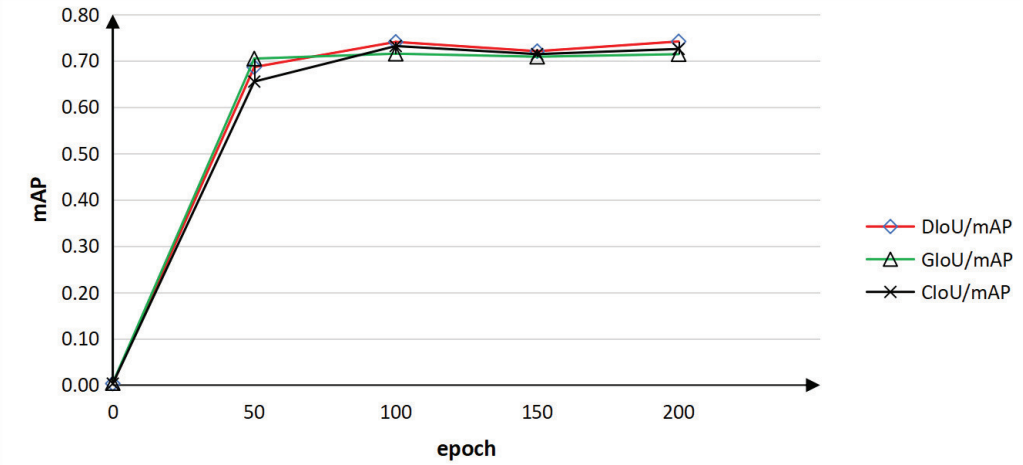

Figure 6. The mAP results of different α-IoU across 200 training stages.

In addition, aiming at the small target and serious occlusion in the process of smoking behavior detection, we designed a group of experiments to compare and test the DIoU-NMS algorithm in the network, as shown in Table 4. Compared with the traditional NMS algorithm, the DIoU-NMS algorithm presented in this paper does not involve unknown parameters to be updated. Therefore, the model after the improved NMS algorithm can be tested directly without training. It can be seen from Table 4 that the improved DIoU-NMS algorithm improves the mAP by 1.1%, and the detection accuracy of unsafe behaviors in industrial site is better than the traditional YOLOv5 algorithm.

Table 4. Comparison of AP results of two different NMS algorithms.

Cases	NMS	DIoU-NMS
Safety Helmet	0.751	0.751
Smoking	0.719	0.74
All	0.735	0.746

3.3. Ablation Experiments

The ablation experiment designed in this section is used to compare and analyze the contributions of the YOLO-AW proposed network in this paper in terms of the four improvement points. The results are shown in Table 5.

Table 5. Contribution comparison of each module.

ASAE	WFPN	α-DIoU	DIoU_NMS	mAP
✓				0.752
	✓			0.75
		✓		0.742
			✓	0.746
✓	✓			0.758
✓	✓	✓		0.763
✓	✓	✓	✓	0.767

It can be seen from Table 5 that the average accuracy of the new network after adding the ASAE model is improved by 1.7%, while the weighted feature-fusion module WFPN can also improve the average accuracy of the network by 1.5%. On the other hand, the introduction of α-IoU loss increases the average precision of the new model by 0.7%, and the addition of DIoU_NMS further improved the detection accuracy by 1.1%. Moreover, Table 5 also demonstrates that the average accuracy of the final model can reach 76.7%, which is 3.2% higher than that of the original YOLOv5.

4. Conclusions

Based on the YOLOv5 network, this paper proposes a YOLO-AW recognition network for detecting smoking and helmet-wearing behavior. The proposed network fuses the feature map output from the backbone network through a new ASAE module, and uses a new feature-fusion structure WFPN, Alpha-IoU and DIoU-NMS to improve the detection effect of unsafe behavior. Finally, the experimental results on the self-developed behavior dataset show that the improved algorithm has higher detection accuracy than the traditional algorithm. When IoU is 0.5, the Recall rate of this method is 68.9%, mAP is 76.7%, AP for smoking detection is 75.2%, and AP for helmet-wearing detection is 78.2%. Compared with other mainstream recognition networks, the proposed network can detect the unsafe behavior of industrial field operators more effectively, and has strong application value.

Author Contributions: Conceptualization, X.W., X.L., Y.W. and B.J.; methodology, X.W. and B.C.; software, Y.W. and B.C.; validation, Y.W. and X.L.; formal analysis, X.W. and B.J.; investigation, B.J., Y.W. and Q.B.; resources, Y.W. and Q.B.; data curation, Y.W. and B.C.; writing—draft B.C.; writing—X.W. and B.C.; visualization, X.L.; supervision, Q.B.; project administration, Q.B. and X.W.; funding acquisition, Q.B., Y.W. and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Key Laboratory of Safety Engineering and Technology Research of Zhejiang Province grant number No. 202001, Key Research and Development Projects in Zhejiang Province grant number No. 2021C03151 and Natural Science Foundation of Zhejiang Province grant number No. LY20F020018.

Acknowledgments: Thank Shangyu XinHeCheng Chemical Co., Ltd. for providing the experimental data in this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shen, C.; Chen, Y.; Yang, G.; Guan, X. Toward Hand-Dominated Activity Recognition Systems with Wristband-Interaction Behavior Analysis. *IEEE Trans. Syst. Man Cybern. Syst.* **2020**, *50*, 2501–2511. [CrossRef]
2. Kacem, A.; Daoudi, M.; Amor, B.B.; Berretti, S.; Alvarez-Paiva, J.C. A Novel Geometric Framework on Gram Matrix Trajectories for Human Behavior Understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 1–14. [CrossRef] [PubMed]
3. Uzair, M.; Brinkworth, R.S.; Finn, A. Bio-Inspired Video Enhancement for Small Moving Target Detection. *IEEE Trans. Image Process.* **2021**, *30*, 1232–1244. [CrossRef] [PubMed]
4. Liu, Y.; Wang, X.; Yan, K. Hand gesture recognition based on concentric circular scan lines and weighted K-nearest neighbor algorithm. *Multimed. Tools Appl.* **2018**, *77*, 209–223. [CrossRef]
5. Lu, H.; Gao, H.; Ye, M.; Wang, X. A Hybrid Ensemble Algorithm Combining AdaBoost and Genetic Algorithm for Cancer Classification with Gene Expression Data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 863–870. [CrossRef] [PubMed]
6. Yang, Y.; Wang, J.; Yang, Y. Improving SVM classifier with prior knowledge in microcalcification detection1. In Proceedings of the 2012 19th IEEE International Conference on Image Processing, Orlando, FL, USA, 30 September–3 October 2012; pp. 2837–2840. [CrossRef]
7. Li, B.; Jiang, W.; Gu, J. Research on Target Detection algorithm based on Deep Learning Technology. In Proceedings of the 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA), Shenyang, China, 22–24 January 2021; pp. 137–142. [CrossRef]
8. Liu, R.; Yu, Z.; Mo, D.; Cai, Y. An Improved Faster-RCNN Algorithm for Object Detection in Remote Sensing Images. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020; pp. 7188–7192. [CrossRef]

9. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. [CrossRef]
10. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
11. Zhu, D.; Fang, Y.; Min, Z.; Ho, D.; Meng, M.Q.H. OCR-RCNN: An Accurate and Efficient Framework for Elevator Button Recognition. *IEEE Trans. Ind. Electron.* **2022**, *69*, 582–591. [CrossRef]
12. Zhao, W.; Xu, M.; Cheng, X.; Zhao, Z. An Insulator in Transmission Lines Recognition and Fault Detection Model Based on Improved Faster RCNN. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–8. [CrossRef]
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
15. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2016; Volume 9905 LNCS, pp. 21–37.
16. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
17. Feng, D.; Liang, M.; Wang, G. Improved YOLOv4 Based on Dilated Convolution and Focal Loss. In Proceedings of the 2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), Dalian, China, 27–28 August 2021; pp. 966–971. [CrossRef]
18. He, J.; Erfani, S.; Ma, X.; Bailey, J.; Chi, Y.; Hua, X.S. Alpha-IoU: A Family of Power Intersection over Union Losses for Bounding Box Regression. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 20230–20242.
19. Jia, X.; Mai, X.; Cui, Y.; Yuan, Y.; Xing, X.; Seo, H.; Xing, L.; Meng, M.Q.H. Automatic Polyp Recognition in Colonoscopy Images Using Deep Learning and Two-Stage Pyramidal Feature Prediction. *IEEE Trans. Autom. Sci. Eng.* **2020**, *17*, 1570–1584. [CrossRef]
20. Shanthamallu, U.S.; Thiagarajan, J.J.; Song, H.; Spanias, A. GrAMME: Semisupervised Learning Using Multilayered Graph Attention Models. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 3977–3988. [CrossRef] [PubMed]
21. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10778–10787. [CrossRef]