ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Safety Helmet Wearing Detection Algorithm based on Improved YOLOv5s in Complex Environment

Taoying Hu[1,*] *and Jiusheng Zhou*[2]

[1] School of Computer and Information Engineering, Maanshan Teacher's College, Maanshan 243041, Anhui, China

[2] Equipment management department,Special Steel Company of Masteel Co., Ltd., China Baowu Group, Maanshan 243000, Anhui, China

SUMMARY: *Safety helmet wearing detection has become particularly important in work sites such as construction, steel and mining. However, the complex environment of the work sites, with numerous equipment, dense personnel, and insufficient lighting, poses a challenge to safety helmet detection. To solve these issues, a target detection algorithm based on improved YOLOv5s is proposed. The BiFormer attention mechanism is added to the neck layer of YOLOv5s to reduce the computational burden by utilizing the Bi-Level Routing Attention (BRA) mechanism, which implements dynamic querying through sparse matrices and focuses on the key information. By introducing the BiFormer attention mechanism, the model can capture key features in images, especially under low-light and high-reflection conditions, enhancing the recognition capability of safety helmet features. Additionally, the introduction of Wise-IoU optimizes the performance evaluation of the model for detecting targets of varying sizes and complexities through a weighted intersection-over-union approach. Experimental results show that the improved YOLOv5s enhances detection accuracy and speed, particularly excelling in the detection of small and overlapping targets. Compared to the YOLOv5s model, the improved model achieved an improvement of 4.4% in accuracy, 2.3% in recall, and 14FPS in detection speed by. Additionally, it can be seen from the experimental transformation curves that mAP@0.5 and mAP@0.5:0.95 have increased significantly. Tests in real-world scenarios validate the practicality and robustness of the algorithm.*

KEYWORDS: *YOLOv5s, Safety Helmet Wearing Detection, BiFormer, Wise-IoU.*

1 Introduction

A safety helmet in construction, steel, mining and other industries is the most effective and direct protection tool; detecting whether the operator is wearing it can effectively reduce the external impact of the material on the production staff brought about by the injury [1], so as to protect the stable development of enterprises [2]. In addition, the realization of automatic helmet wearing detection can not only reduce safety hazards at work sites by recording and analyzing violations, but also has an important role in promoting the improvement of the information management level of safety production.

Traditional safety helmet detection methods, such as manual inspection or video surveillance, suffer from regulatory blind spots and inefficiency. Realizing the automation, real-time and intelligence of safety helmet wearing and detection has become possible with

*luck_hty@163.com
https://doi.org/10.65102/is20261181
Published: 10 June, 2026

the rapid development of computer vision and deep learning technology[3]. The following is the evolution of target detection algorithms from traditional detection methods to two-stage and single-stage: before the maturity of deep learning techniques around 2000, target detection relied mainly on hand-designed features, such as LBP features [4], Haar features [5], and HOG edge feature detection [6].

However, these methods have limited generalization ability in complex environments and are computationally inefficient.In 2013, Girshick et al.[7] first introduced the deep learning-based R-CNN target detection algorithm, a region-based convolutional neural network approach, which is relatively computationally expensive despite its significant progress in detection accuracy. Subsequently, Girshick et al. proposed Fast R-CNN [8] and Faster R-CNN [9] algorithms in 2015 and 2016.By optimizing the R-CNN, these algorithms not only improved the detection speed, but also enhanced the performance. Although two-stage target detection algorithms excel in prediction accuracy, they still have room for improvement in detection speed. Since 2016, the development of single-stage target detection algorithms has brought a major breakthrough in the field, and the creation of the You Only Look Once (YOLO)[10] algorithm transforms the target detection task into a regression problem, where the category and location of the target can be predicted simultaneously through a single forward propagation process. This method improves the detection speed. As the YOLO series of algorithms continue to be iterated and updated, each optimization improves the accuracy and speed of detection to varying degrees.In 2016, Liu et al. introduced the Single Shot MultiBox Detector (SSD) algorithm[11], which effectively has improved the recognition capability for small objects by detecting feature maps of different scales from the VGG network. This algorithm is another single-stage target detection algorithm with excellent performance that is comparable to that of YOLO series.

However, in the complex environments of workplaces, such as buildings, steel, industrial, and mining, traditional target detection models often face many challenges. For example, equipment reflections and shadows may affect the clarity of the image, dense crowds of people may result in overlapping targets, and insufficient light may degrade the image quality, all of which can negatively affect the accuracy and efficiency of target detection.In response to these complex environmental characteristics, a scheme for helmet wear detection based on an improved YOLOv5s model is designed using a large-scale steel production site as the research background. Considering the lightweight and fast detection characteristics of YOLOv5s, it is chosen as the base model. Subsequently, the model is improved as follows:

1) The BiFormer attention mechanism is integrated into YOLOv5s to enhance the model's ability to capture key features.

2) The bounding box loss function is optimized, and a weighted IoU computation method, Wise-IoU, replaces the CIoU loss function of the original model to evaluate the detection performance more accurately.

2 Improved YOLOv5

2.1 YOLOv5 Model

YOLOv5 was released by the Ultralytics team and this version builds on the improvements made in the YOLOv4 series. A variety of different model sizes is provided to suit different industrial application scenarios, while the code structure has been simplified to make it easier to use and deploy.YOLOv5 introduces a series of lightweight models in versions S, M, L, and X, which are differentiated according to the complexity of the model and the computational resources required.

The YOLOv5 model is consists of an image input stage (input), a backbone network (backbone), a connection network (neck), and output head (head).

1) Input: YOLOv5 enhances the model performance by Mosaic data augmentation to mix four images to increase sample diversity, adaptive anchor box calculation to optimize bounding box predictions, and adaptive image scaling technology to ensure the model's detection accuracy for targets of various sizes. These innovative methods significantly improve the ability of the model to recognize and locate targets of different sizes.

2) Backbone: YOLOv5 uses CSPDarknet53 as its backbone network, which is an optimized version of the Darknet network. By introducing the Cross Stage Partial Network (CSPNet) architecture, it reduces computational complexity while still maintaining efficient feature extraction capabilities. This improvement makes YOLOv5 more efficient in processing images, while ensuring the accuracy of object detection.

3) Neck: YOLOv5 integrates the Spatial Pyramid Pooling with Focus (SPPF) module in the model's neck. This design combines the Spatial Pyramid Pooling (SPP) technology with the Focus module. Such a combination not only enhances the model's ability to capture multi-scale features but also improves the effectiveness of feature fusion, thereby optimizing the detection performance for targets of different sizes[12].

4) Head: The output end of YOLOv5 uses the CIoU Loss as the loss function, which includes bounding box regression loss, confidence loss, and classification loss. These loss functions work together in the model's training process, helping to predict the target's position more accurately.

2.2 Biformer Attention Mechanism

Incorporating an attention mechanism allows the detection model to focus on important information and fully absorb important information[13]. By introducing the BiFormer attention mechanism, the model is can capture the key features of the safety helmet features more efficiently, especially under low-light and high-reflection conditions. However, the introduction of the attention mechanism inevitably brings about the problem of high computational cost and memory occupation of the model when dealing with large-scale data. Many researchers have committed to performing optimization work in this area, such as lightweight structure design. BiFormer attention mechanism[14] provides a new paradigm for visual transformers that effectively reduces computational complexity and memory usage.The core innovation of BiFormer is its Bi-level Routing Attention (BRA) mechanism. BiFormer not only effectively reduces the computational burden but also improves the efficiency of computation and the performance of the model by allowing each query to focus on the semantically most relevant key-value pairs through a dynamic sparse attention pattern.

Biformer model innovation lies in the implementation of a two-tier routing attention mechanism (BRA). This attention mechanism is divided into two layers: coarse-grained area-level filtering and fine-grained Token-to-Token attention computation [15]. Specifically, it consists of three steps:

Region partitioning and input projection. The input feature map $X \in R^{H\times W\times C}$, is partitioned into a non-overlapping regions $S \times S$, where each region contains feature vectors $\frac{HW}{S\times S}$. Linear transformation is performed to obtain the tensor of the query (query, Q), key (key, K) and value (value, V), and the linear mapping is calculated as follows:

$$Q = X^r W^q, K = X^r W^k, V = X^r W^v. \quad (1)$$

where are the projection weights of Q, K, and V, respectively.

Regional-to-regional routing constructs a directed graph at the regional level to determine the focus of other regions $Q^r, K^r \in R^{S^2 \times C}$, each region should focus on. Retaining the most important connections $I^r = \text{topkIndex}(A^r)$ through a top-k operation, forming a routing index matrix [16].

1. Token-to-Token attention, which utilizes a route index matrix to apply fine-grained attention. Initial filtering of irrelevant key-value pairs is achieved by constructing and pruning a directed graph of inter-region correlations at region level. Token-to-Token attention computation is performed in the selected routing regions to ensure that each query focuses on the semantically most relevant regions. Using the indexing matrix I^r, BiFormer is able to aggregate relevant key and value tensors, aggregating the key and value tensors is calculated as follows.

$$K^g = gather(K, I^r), V^g = gather(V, I^r). \tag{2}$$

Token-to-Token attention is then performed on these aggregated K-V pairs which is computed as follows.

$$O = Attention(Q, K^g, V^g) + LCE(V) \tag{3}$$

The structural of BiFormer design is based on BRA, and a four-level pyramid structure is adopted, containing multiple stages, each of which uses the BiFormer model for feature transformation as shown in Figure 1. This structural design not only optimizes the propagation and extraction of features, but also achieves flexible processing of inputs with different resolutions by adjusting the top-k parameter of different stages, which enhances the adaptability of the model to complex visual tasks. In stage i, the input spatial resolution is reduced using overlapping patch embedding (i=1) or using the patch merging (i=2,3,4) module, while the number of channels is increased.

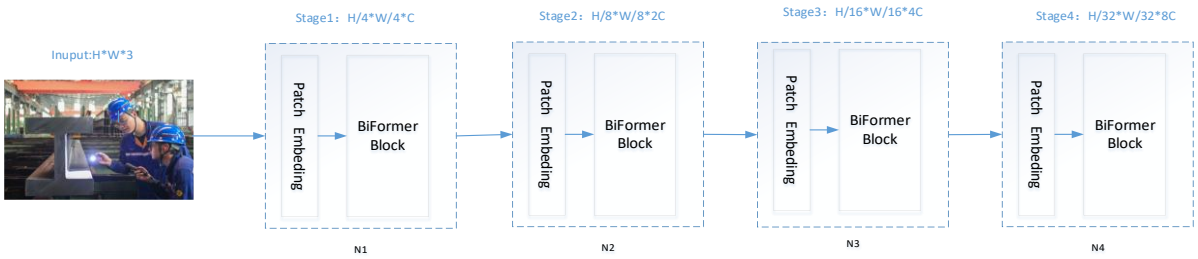

Figure 1: BiFormer structure

By integrating the BiFormer behind the C3 module in front of the head in the neck of the YOLOv5 structure, each BiFormer module is similar to a small transformer. Embedding the BiFormer attention mechanism improves the detection accuracy without decreasing the detection efficiency of the model by generating a large amount of computation. The network structure of the improved YOLOv5s is shown in Figure 2.

2.3 WISE-IOU

The Wise-IoU (WIoU) loss function is introduced to optimize the original Complete-IoU(CIoU) function of YOLOv5. WIoU takes into account the dynamic focusing mechanism, which enables it to perform well in complex application scenarios.Dynamic focusing mechanism if WIoU can adapt to the uneven distribution of targets, dynamically adjust the focus of the loss function, and optimize the model's learning process. The WIoU can help the model predict the bounding box more accurately, especially for small targets in the case of target detection tasks with different sizes. In addition, WIoU reduces the harmful gradient owing to low-quality annotations by

reasonably distributing the gradient gain and improving the model's performance in low-light scenarios. The WIoU can also effectively deal with the mutual interference between targets and improve the detection accuracy in a densely populated environment.

The WIoU loss function currently contains three versions, WIoUv1, WIoUv2 and WIoUv3, each of which has different characteristics and application scenarios.WIoUv1 is the initial version, which is constructed based on the distance attention mechanism, focusing on reducing the regression loss weight of high-quality anchor frames and reducing the attention to the distance when the anchor frames overlap well with the target frames[17].WIoUv1 is designed to weaken the penalty of the geometric factors when the anchor frames overlap well with the target frames, and to improve the generalization ability of the model. WIoUv1 is calculated as follows:

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \tag{4}$$

$$\mathcal{R}_{WIoU} = exp\left(\frac{(x-x_{gt})^2+(y-y_{gt})^2}{(W_g^2+H_g^2)^*}\right) \tag{5}$$

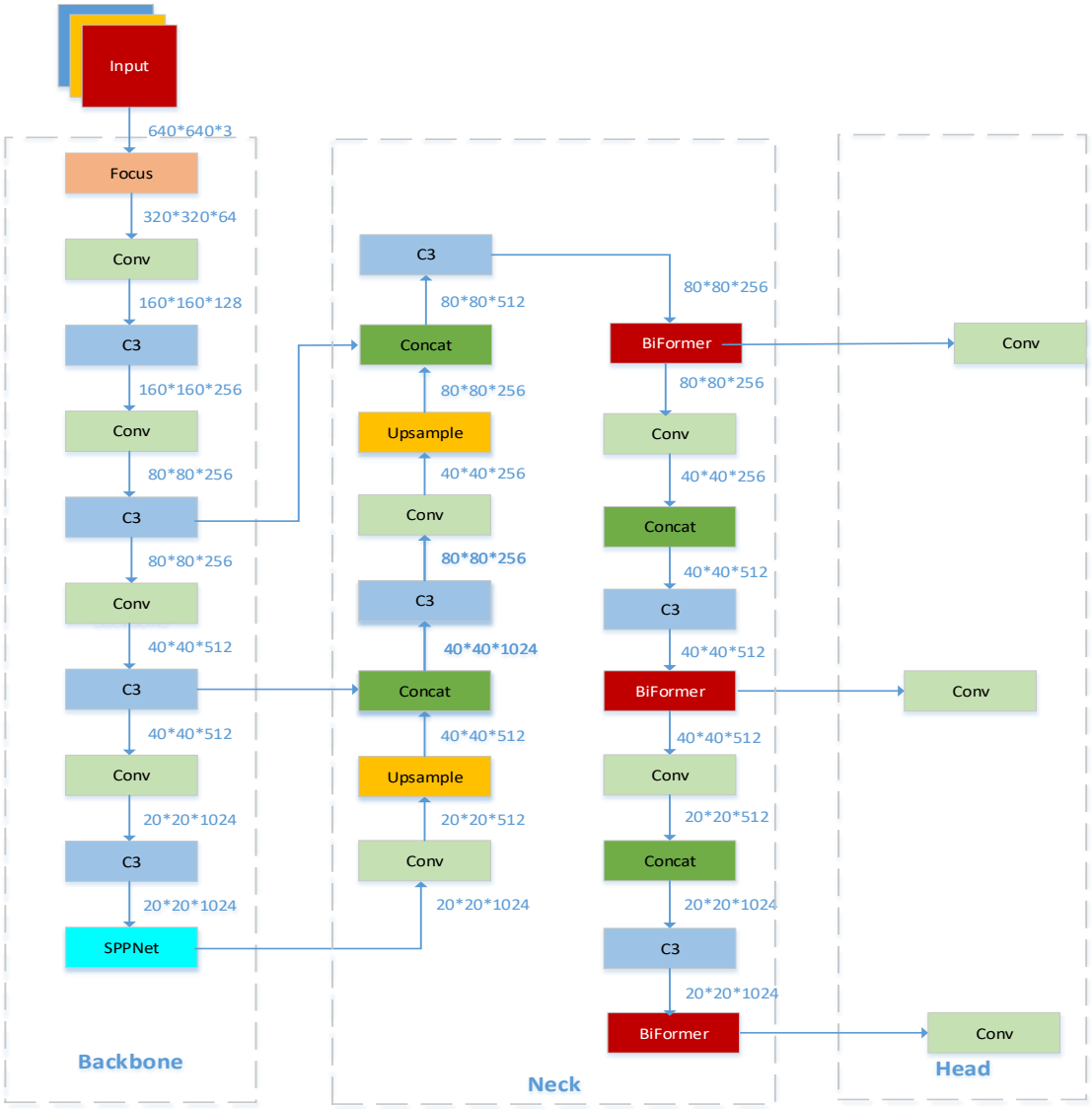

Figure 2: Improved YOLOv5 network structure

where Wg and Hg denote the width and height of the smallest closed box that can cover both the predicted and actual boxes. To prevent the WIoU from generating gradients during the training process that may prevent the model from converging, the computation of Wg and Hg is specifically handled to separate it from the main computational graph, where the superscript * denotes this operation [18]. This treatment effectively removes factors that may negatively affect model training convergence.

WIoUv2 introduces a monotonic focusing mechanism, similar to Focal loss, which reduces the impact of simple samples on the loss value through a monotonically increasing focusing factor, making the model more focused on difficult samples.WIoUv2 solves the problem of slow convergence in the late stages of training by maintaining a high gradient gain through a dynamically updated normalization factor.The WIoUv2 is caculated as follows:

$$\mathcal{L}_{WIoUv2} = \left(\frac{\mathcal{L}^{*}_{IoVt}}{\bar{\mathcal{L}}_{IoU}}\right)^{'} \mathcal{L}_{WIoUv1}, \gamma > 0 \tag{6}$$

where L_{IoU} is the IoU loss. L_{WIoUv1} is the loss of WIoUv1, and γ is a hyper-parameter that controls the strength of monotonic focusing. The monotonic focusing coefficient $L^{*}_{IoU}, \bar{\mathcal{L}}_{IoU}$ normalized by a factor that is an exponential sliding mean with momentum $m$and $\left(\frac{\mathcal{L}^{*}_{IoUU}}{\bar{\mathcal{L}}_{IoU}}\right)^{\gamma} \in [0,1]$ are the gradient gain r.

WIoUv3 is a state-of-the-art version of WIoU that employs a dynamic non-monotonic focusing mechanism that uses outliers to assess the quality of anchor frames and provides a judicious gradient gain assignment strategy [19].WIoUv3 reduces the harmful gradients generated by low-quality samples while decreasing the competitiveness of high-quality anchor frames, allowing the model to focus more on average-quality anchor frames, thereby improving the overall performance of the detector.WIoUv3 constructs a nonmonotonic focusing coefficient using outliers as shown in Equation 7.

$$\beta = \frac{\mathcal{L}^{*}_{IoU}}{\bar{\mathcal{L}}_{IoU}} \in [0, +\infty) \tag{7}$$

Then β is applied to WIoUv1 to construct the WIoUv3 as shown in Equation 8.

$$\mathcal{L}_{WIoUv3} = r\mathcal{L}_{WIoUv1}, r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{8}$$

where$\alpha$and $\delta$are hyper-parameters that control the morphology of the nonmonotonic focusing coefficients.

In practical applications, which the version of the WIoU to choose depends on the specific task and dataset characteristics. Considering the complexity of the practical application scenarios. It is more appropriate to cite WIoUv3 for this issue because it can effectively reduce the negative impact of low-quality samples on model training.

3 Experimnets

3.1 Experimental Environment and Dataset

The hardware system used for the experiment is an Intel(R) Core(TM) i7-10750H CPU and an NVIDIA GeForce RTX3060 graphics card. The software system is the Windows 10 operating system and CUDA version 12.1. Experimental environments using Anaconda installation and configuration of deep learning environment. The first installation of Anaconda3 and then use the

Anaconda Installers to install python and pytorch. The version are Python3.9 and Pytorch2.3 respectively. Pycharm was used as the developmental environment for the experiments.

The dataset used in the experiment is based on the publicly available PASCAL Visual Object Classes (VOC) dataset, specifically the VOC2007 version, which includes 20 classes. This experiment focuses on the "person" class, selecting 5,209 images as foundational data. To enhance sample diversity and improve the accuracy of detection usage in steel production environments, an additional 774 images from various real scenarios in a large steel plant were collected through on-site photography, web scraping of promotional sites, and frame extraction from surveillance videos. During image collection, factors such as lighting conditions, shooting angles and distances, obstruction by objects, and image clarity were considered to ensure the dataset diversity. The supplementary dataset was annotated using image labeling tools to create a custom dataset. This experiment utilized LabelImg for manual annotation of the collected images. The labels are primarily divided into two categories: "hat" and "person," with those wearing safety helmet labeled as "hat" and those not wearing them labeled as "person."

The labeled file is generated in VOC (XML) format, and then the script is used to convert the label file in XML format to YOLO (TXT) format file.The XML file records the labeled image folder, filename, path, source, format, category of the target object, location and other information. The YOLO corresponding TXT file records the (class_id, x, y, w, h) of each tag, with the parameters of the class, coordinates of the center point, and width and height, respectively [20].

According to the actual demand, the dataset is randomly divided into training set, validation set and test set by script, and the ratio is set to 8:1:1. The size of the experimental input image is set to 640 × 640, and the number of training epochs is set to 150. Figure 3 shows the basic situation analysis of the dataset. The upper-left figure is the number of samples of the dataset that are "hat" and "person." The upper-right figure shows the size and number of marking frames, the bottom-left figure shows the coordinate distribution of the center point, and the bottom-right figure shows the width and height of the marking frames. From the figure, it is also possible to analyze the basic situation of the target size, with darker areas concentrating on a higher number of smaller targets.

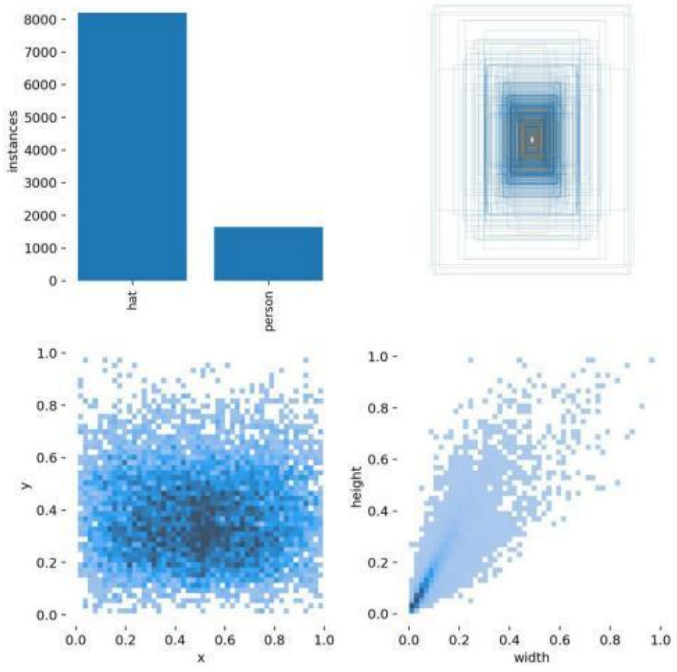

Figure 3: Data set basic situation analysis

3.2 Evaluation Indicators

Validating the performance of a target detection model is important for visually assessing whether the means taken to improve the optimized model have a positive impact on the model [21].

1. Precision(P) reflects the proportion of all samples for which the model is judged to be a positive example that are actually positive examples. In general, an increase in precision usually implies an enhancement in the model performance.

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

TPrepresents the total number of correctly identified positive examples and FP represents the total number of incorrectly determined positive examples.

2. Recall(R) measures the ability of the model to accurately identify all actual positive category samples. A higher recall rate indicates that the model captures more positive category samples, thus reflecting a superior model performance.

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

3. Average Precision (AP) is used to measure the balanced performance of the detection model across categories. It provides a comprehensive evaluation of the overall performance of the model by calculating the average precision across all categories. It can be obtained by calculating the area under the precision-recall curve.

$$AP = \int_{P}^{1}(R)dR \tag{11}$$

4. Mean Average Precision (mAP) calculates the average precision under different confidence thresholds, usually between 0.5 and 1.0. The mAP integrates the precision and recall of the model, and is an important indicator for evaluating the overall performance of the model.

$$mAP = \frac{1}{N}\sum_{i-1}^{N} A_{Pi} \tag{12}$$

AP_i is the average precision of category i, and is the total number of categories. Additionally, mAP@0.5 indicates the mAP when the IoU threshold is 0.5, whereas mAP@0.5:0.95 represents the average mAP when the IoU threshold ranges from 0.5 to 0.95.

5.The number of frames processed per second (FPS) is used to measure the inference speed of the model.

3.3 Model Training

Figure 4 shows a performance comparison between YOLOv5s and its improved model over 150 training epochs. The performance metrics include the mean accuracy (mAP) and total loss, which vary as the training epochs increase.

Figure 4(a) describes the mAP@0.5:0.95 of the two models at different training epochs. The curves show that the mAP values of YOLOv5s+BiFormer+WIoU (orange line) are higher than those of the original YOLOv5s model (blue line) for most of the training epochs. This indicates that the model integrating the BiFormer and WIoU modules has higher accuracy in the object detection task.

Figure 4(b) shows the total loss curves for both models. In deep learning, the loss function estimates the inconsistency between the predicted and actual values of a mode. The loss of both

models decreases as the training epochs increase, but the loss of YOLOv5s+BiFormer+WIoU decreases more significantly, converges faster, and remains lower than that of YOLOv5s.

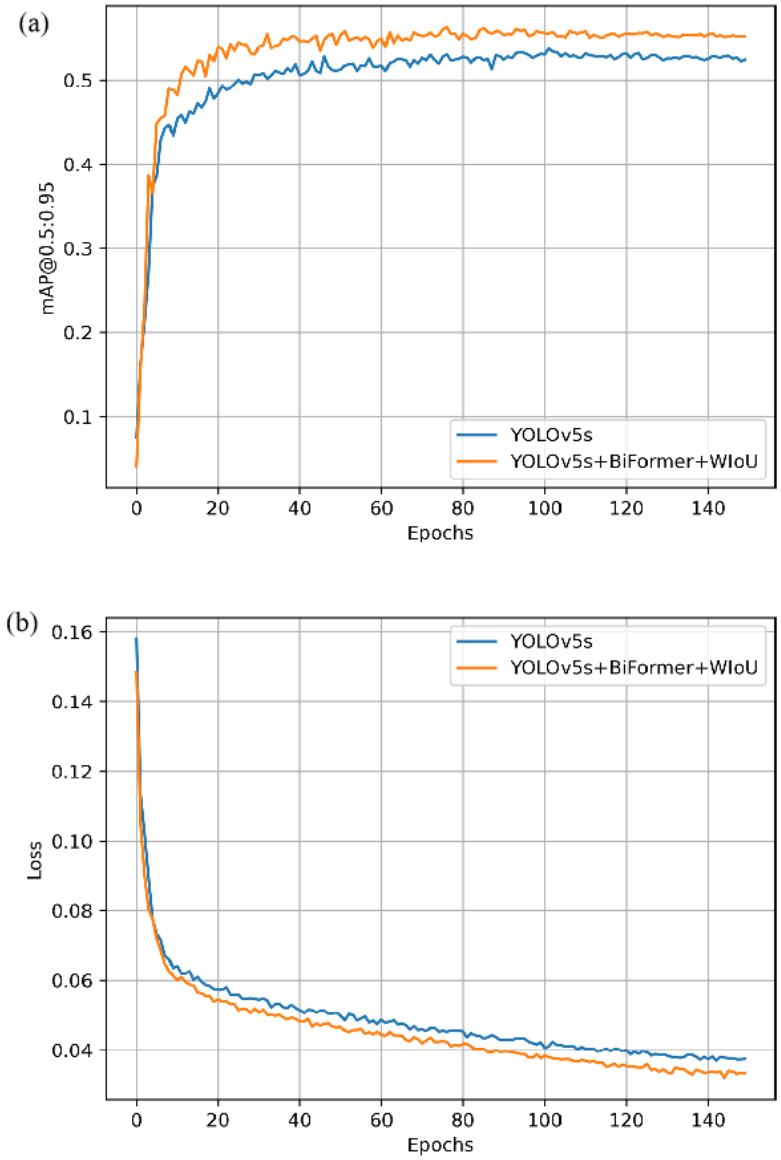

Figure 4: Comparison of mAP and loss between YOLOv5s and improved YOLOv5s.(a) mAP(%).(b) loss.

Figure 4 shows that the performance of object detection is improved by the improved model. Specifically, the improved model improves the accuracy and converges faster with less loss during the learning process.

3.4 Comparative Experiments of Different Attention Mechanisms

To validate the effectiveness of incorporating the BiFormer module into YOLOv5s proposed in this study, comparative experiments of attention mechanisms were conducted on the same dataset, hardware and software conditions. The more popular current attention mechanisms, CBMA, CA, ECA, and SimAM are incorporated in YOLOv5s. Each row in the table represents a combination of experiments with different modules. The results of the experiments are listed in Table 1:

1. The last row, using BiFormer, shows the highest performance metrics, with 2.4%, 2.5%, 0.4%, and 1.1% improvements in precision, recall, mAP@0.5 and mAP @0.5:0.95, respectively, compared with the YOLOv5s model. This indicates that the BiFormer module works best when used in conjunction, and provides the best detection performance. The mAP@0.5 and mAP@0.5:0.95 reach 85.4% and 54.8%, respectively, which indicates that the model is highly robust under different IoU thresholds.

2. Precision, recall, mAP@0.5 and mAP@0.5:0.95 fluctuate in all the experimental parameters. However, the introduction of the BiFormer module led to significant increases in all the indicators. In particular, the FPS increases to 31, indicating that the detection rate of the model is improved. Thus, it proves that the incorporation of the BiFormer attention mechanism has a better effect on model improvement.

3.5 Comparative Experiments of Actual Scene Test

In order to more intuitively feel the application effect of the improved target detection algorithm in the actual scene, four groups of photos taken in the actual scene were selected to visualize and compare the test results, and the following Figure 5 shows the comparison of the four groups of test results.

(a) Low-light：It can be seen that the model has a good detection effect in low light scenes both before and after improvement, but the accuracy is obviously improved after improvement.

(b) Long-distance: This shows that the base model has missed detection in the shooting of a long-distance scene, where there are three people wearing safety helmets, however, the original algorithm only recognizes two objects, compared to the improved algorithm, which successfully recognizes all the targets. This is due to the embedding of the BiFormer attention mechanism and the Wise-IoU loss function, both of which contribute to the performance of small and dense target detection.

(c) Numerous-equipment: It shows the results of a wide range of scenarios tested with field devices, and the improved algorithm also shows better detection performance for small distant targets.

(d) Partial-obstruction: The target on the right is partially obscured by the billet, and the test results show that the improved algorithm also has good accuracy for target detection in the obscured scenes.

Table 1: Comparison of experimental results of different attention mechanisms

model	Precision	Recall	mAP@0.5	mAP@0.5:0.95	FPS
YOLOv5s	0.891	0.768	0.85	0.537	29
YOLOv5s+CBMA	0.897	0.787	0.853	0.536	20
YOLOv5s+CA	0.905	0.772	0.844	0.529	27
YOLOv5s+ECA	0.892	0.796	0.858	0.539	28
YOLOv5s+SimAM	0.887	0.785	0.836	0.523	27
YOLOv5s+BiFormer	0.915	0.793	0.854	0.541	31

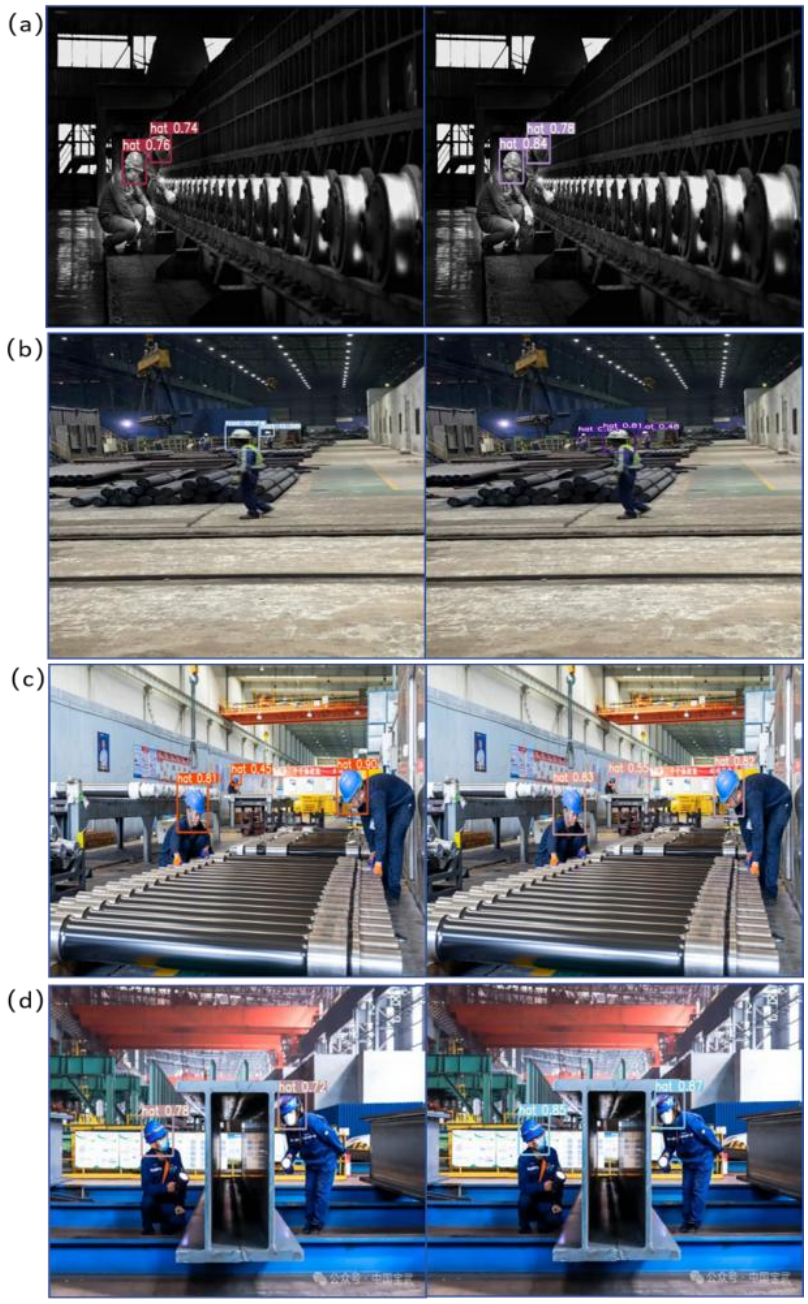

Figure 5: Comparison of detection performance in various complex scenarios between YOLOv5s (first column) and improved YOLOv5s (second column).(a)Low-light.(b)Long-distance.(c)Numerous-equipment.(d)Partial-obstruction.

3.6 Ablation Experiment

To validate the effectiveness of incorporating the BiFomer attention mechanism and Wise-IoU loss function, ablation experiments are performed on the same dataset as well as under the same experimental conditions.YOLOv5s is used as the base model. Precision, recall, mAP, and FPS are used as the evaluation indicators. The experimental results are listed in Table 2.

1. The precision and recall are improved by 0.8% and 0.4%, respectively, after modifying the loss function to WIoU in YOLOv5s. However, there is some loss in FPS.

2. After adding the BiFormer attention mechanism to the neck structure of YOLOv5s, there is a significant improvement in the model precision, which increases by 2.4%, whereas there is a

small loss in the recall metric. However, the FPS reaches 31, thus effectively verifying that the introduction of the BiFormer attention mechanism does not result in a decrease in the efficiency of model checking owing to the high computation and memory usage.

3. Finally, BiFormer and WIoU models are added to YOLOv5s at the same time, and it is obvious from the table data that the model precision, recall, and FPS have been improved to different degrees.Which the precision and recall have been improved the most obviously, by 4.4% and 2.3% respectively compared with original model.Detection speed has also increased by 14 FPS.

Table 2: Results of Ablation Experiment

model	Precision	Recall	FPS
YOLOv5s	0.891	0.793	19
YOLOv5s+WIoU	0.899	0.797	17
YOLOv5s+BiFormer	0.915	0.768	31
YOLOv5s+WIoU+BiFormer	0.935	0.816	33

Figure 6(a) describes mAP@0.5:0.95. Figure 6(b) describes mAP@0.5. Each model is represented separately using a different color. The blue curve indicates the transformation of the original model training. The green and orange curves indicate the training transformation process after adding the BiFomer attention mechanism and WIoU loss function, respectively. The red curve indicates the model training transformation process with the integration of the BiFomer module and WIoU loss function. The mAP values of YOLOv5s+BiFomer+WIoU (red line) are higher than those of the original YOLOv5s model (blue color) in most of the training epochs. This indicates that the model integrating the BiFomer and WIoU moduls has higher accuracy in the object detection.

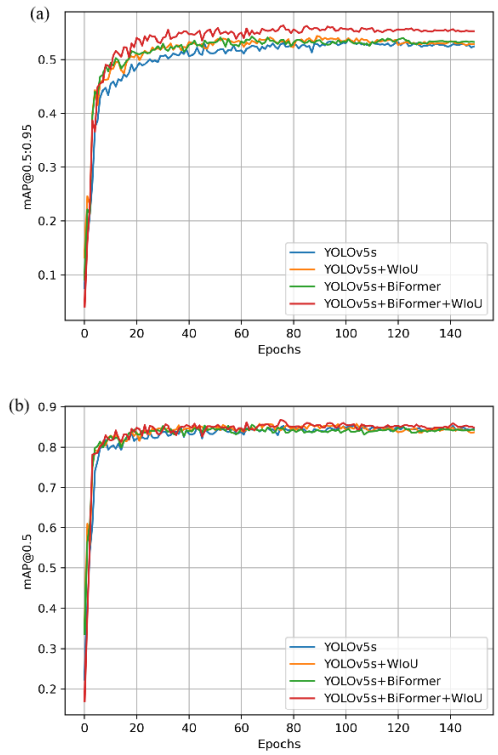

Figure 6: Comparison of ablation experiments.(a)mAP@0.5:0.95(%).(b)mAP@0.5.

4 Conclusions and Future Work

The YOLOv5s model is effectively improved by introducing the BiFormer attention mechanism and Wise-IoU loss function to adapt to the special needs of complex environments in this issue. Through a series of experimental validations, the improved model demonstrates higher accuracy and robustness in the safety helmet wearing detection, especially in dealing with complex backgrounds, low-light conditions, and target occlusion, etc. The integration of BiFormer attention mechanism significantly enhances the ability of the model to capture key features, enabling it to more accurately recognize the location and contour of safety helmets. Meanwhile, the introduction of Wise-IoU provides more reasonable criteria for evaluating the model's performance, especially in the detection of small and overlapping targets, which makes the detection results more in line with the actual application scenarios. Tests in real steel production environments show that the improved YOLOv5s model not only meets the demand for real-time monitoring in terms of detection speed, but also reaches industrial safety standards in terms of detection accuracy. This provides a new technical support for safety management, and helps to reduce safety accidents caused by failure to wear safety helmets.

Despite the results achieved in this study, there is still room for further optimization and improvement. Future work can focus on the following areas: further optimizing the BiFormer attention mechanism to improve the ability of the model to recognize different types of safety helmets, exploring more efficient network structures to reduce the computational complexity of the model, and testing and validating the model's generalization ability in a wider range of industrial scenarios.

Funding

This work was supported by the Natural Science Foundation of Anhui Province under Grant 2024AH051773, and in part by the Natural Science Foundation of Anhui Province under Grant 2022AH052835.

Author's Profile

Taoying Hu was born in Wuhu, Anhui, China in 1987. She obtained her bachelor's degree from Anqing Normal University in 2010 and the M.S. from Hefei University of Technology in 2013.Her main research is Object Detection Algorithm and applied computer technology.

Jiusheng Zhou was born in Chuzhou, Anhui, China in 1987. He obtained his bachelor's degree from Anqing Normal University in 2010.His main research is Artificial Intelligence and Process Management System.

References

[1] X. Zhang, Y. Zhang, F. Wang, and Y. Liu Yiming, “Improved YOLOv7 helmet wearing detection algorithm for steel rolling workshop”. Computerized Measurement and Control, Beijing, China,Oct.2023,pp.15-22,dio:10.16526/j.cnki.11-4762/tp.2024.07.003.

[2] H. Xu, Z. Deng, C. Yao, and C. Ye, “Improved helmet wearing detection algorithm for YOLOv5”, Software Guide, Beijing, China, Aug. 2023, pp.33-41, dio:10. 11907/rjdk. 221885.

[3] Y. Liu, Y. Ilhamu, L. Xi, and A. Intezar, “Research on helmet wearing detection algorithm with improved YOLOv5s”, Computer Engineering and Applications,Beijing, China,Aug.2023,pp.184-191,doi: 10.3778/j.issn.1002-8331.2305-0237.

[4] Ojala T, Pietikainen M, Harwood D, “Performance evalua tion of texture measures with classification based on Kull back discrimination of distributions”,in Proceedings of 12th international conference on pattern recognition. IEEE, Oct.1994, pp. 582-585, doi: 10.1109/ICPR.1994.576366.

[5] Papageorgiou C P, Oren M, Poggio T, “A general framework for object detection”, in Sixth international conference on computer vision (IEEE Cat. No. 98CH36271). IEEE, Jan. 1998,pp.555-562,dio:10.1109/ICCV.1998.710772.

[6] Dalal N, Triggs B. “Histograms of oriented gradients for hu man detection”, in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, Jun.2005, pp. 886-893.dio:10.1109/CVPR.2005.177.

[7] Girshick R, Donahue J, Darrell T, et al, “Rich feature hierar chies for accurate object detection and semantic segmentation”, Proceedings of the IEEE conference on computer vision and pattern recognition, Jun. 2014,pp.580-587.

[8] Girshick R.Fast r-cnn[C]//Proceedings of the IEEE interna tional conference on computer vision. 2015: 1440-1448.

[9] Ren S, He K, Girshick R, et al, “Faster r-cnn: Towards real time object detection with region proposal networks”, Advances in neural information processing systems, Jun. 2015, pp.1137 - 1149,dio:10.1109/TPAMI.2016.2577031

[10] Redmon J, Divvala S, Girshick R, et al, “You only look once: Unified, real-time object detection”, Proceedings of the IEEE conference on computer vision and pattern recogni tion. 2016, pp. 779-788.

[11] W. Liu W, Anguelov D, Erhan D, “Ssd: Single shot multi box detector”, Computer Vision-ECCV 2016: 14th Euro pean Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I 14. Springer International Publishing, Beijing, China, 2016, pp.21-37.

[12] Z. Deng, Y. XIONG, R. Yang, and Y Chen, “Improved YOLOv5 Helmet Wear Detection Algorithm for Small Targets”, Beijing, China, Computer Engineering and Applications, Aug.2023, pp.184-191.

[13] H. ZHOU, y. LI, and A. DANG, “Edge-enhanced small target detection of floating garbage on water surface based on YOLOv7”, Journal of Langfang Normal College, Beijing, China, Jun. 2024, pp.45-51.

[14] L. Zhu,X. Wang,Z. Ke,W. Zhang,and Rynson Lau, “BiFormer: Vision Transformer with Bi-Level Routing Attention”, arXiv preprint arXiv:2303.08810,Beijing, China, 2023.

[15] Z. Liu, H. Xu, X. Zhu, C. Li, Z. Wang, Y. Cao, and K. Dai, “Bi-YOLO: a lightweight target detection algorithm based on the improvement of YOLOv8”, Computer Engineering and Science, Beijing, China,Vol.46, No.8, Aug.2024,pp.1444-1454.

[16] D. LI, Y. SUN, P. WANG, and M. YE, “Improved surface defect detection algorithm for bushing parts with YOLOv7-tiny”, Inspection and Quality, Beijing, China, Jun.2024, pp.133-140.

[17] L. Zheng, and Y. Zhang, “Traffic signal detection based on improved YOLOv7”, Journal of System Simulation, Beijing, China, Mar. 2024,dio:10.16182/j.issn1004731x.joss.23-1562.

[18] C. MA, H. ZHANG, and X. MA, “A lightweight wheat disease detection method based on improved YOLOv8”, Journal of Agricultural Engineering, Beijing, China, Vol40, No.5, Mar.2024, pp.187-195.

[19] D. Xu, S. Wang, K. Yin, and Z. Wang, “Improved urban vehicle target detection algorithm for YOLOv8”, Computer Engineering, Beijing, China, Jul. 2004, dio:10.19678/j.issn. 1000-3428.0069125.

[20] Q. Yu, Q.Wan, and W. HU, “Safety helmet Detection based on Improved YOLOv5s”, Computer Processing, Beijing, China, Dec.2 023, No.6, pp.50-54.

[21] Y. Gong, M. XIA, K. WANG, and J. ZHAI, “Based on improved YOLOv5s helmet wear detection Algorithm”, Journal of Harbin University of Commerce, Beijing, China,Vol.39, No.5,Oct.2023,pp.550-557.