Detection of Protective Apparatus for Municipal Engineering Construction Personnel Based on Improved YOLOv5s

Shuangyuan Li[1], Yanchang Lv[2], Mengfan Li[3], Zhengwei Wang[4]
Information Construction Office, Jilin Institute of Chemical Technology, Jilin, China[1]
School of Information and Control Engineering, Jilin Institute of Chemical Technology, Jilin, China[2, 3, 4]

***Abstract*—With the rapid economic development, the government has increased investment in municipal construction, which usually takes a long time, involves many open-air operations, and is affected by cross-construction, traffic, climate and environment, and so on. The safety protection of urban construction workers has been a concern. In this paper, an improved algorithm based on YOLOv5s for the simultaneous detection of helmets and reflective vests is proposed for municipal construction management. First, a new data enhancement method, Mosaic-6, is used to improve the model's ability to learn local features. Second, the SE attention mechanism is introduced in the focus module to expand the perceptual field, strengthen the degree of association between channel information and the detection target, and improve the detection accuracy. Finally, the features of small-scale targets are interacted and fused in multiple dimensions according to the Swin transformer network structure. The experimental results show that the improved algorithm achieves accuracy, recall, and mean accuracy rates of 98.5%, 97.0%, and 92.7%, respectively. These results show an average improvement of 3.4 percentage points in mean accuracy compared to the basic YOLOv5s. This study provides valuable insights for further research in the area of urban engineering security and protection.**

Keywords*—YOLOv5s; hard hat; reflective vest; simultaneous detection

I. Introduction

With the relentless advancement of science and technology, Artificial Intelligence (AI) has emerged as a key player in numerous fields. One such field is municipal engineering, where ensuring the safety of construction workers is of paramount importance. Wearing helmets and reflective vests is an essential measure to ensure their well-being. Combined with artificial intelligence target detection method, it can realize automatic detection and monitoring of whether construction workers are wearing helmets and reflective vests. It is necessary to use artificial intelligence target detection method to detect municipal engineering construction workers wearing helmets and reflective vests [1-3]. First of all, AI target detection method can improve the efficiency and accuracy of site safety monitoring. Traditional safety monitoring methods usually rely on manual inspection, there is a waste of human resources, and blind area coverage is not complete [4]. The use of AI target detection methods, combined with cameras and image processing technology, can monitor the construction site in real time and automatically detect whether there are construction workers wearing helmets and reflective vests. This automated monitoring method can not only reduce the burden of site managers, but also detect and deal with violations in a timely manner, improving the overall safety level of the site [5]. In addition, the AI target detection method can reduce the interference of human factors on safety monitoring results. In traditional safety monitoring, manual judgment can be affected by factors such as subjective awareness, perspective limitations and fatigue, resulting in inaccurate or missing detection results for construction workers wearing helmets and reflective vests. In contrast, AI target detection methods use advanced algorithms and models to quickly and accurately detect target objects without the interference of subjective factors [6, 7]. This ensures more objective and accurate monitoring results for construction workers wearing helmets and reflective vests, effectively improving the reliability of safety monitoring.

The remaining sections of this paper are organized as follows. In Section II, we introduce two deep learning-based detection algorithms and discuss the current state of research in this area, Section III elaborates the basic architecture of the YOLOv5 model and details the optimization part for YOLOv5, Section IV of this paper provides an overview of the dataset sources, the experimental environment, and the model parameter settings. In Section V, we present the experimental results and discuss the performance of the optimized model introduced in the previous section. Experimental tests are conducted for the three optimized parts and compared with the current mainstream algorithms.

II. Current Status of Research

A. Introduction of Deep Learning Algorithm

Deep learning is an advanced and complex machine learning algorithm in the field of AI, which has the advantages of powerful learning ability, high generalization ability, and wide applicability. Deep learning based algorithms in the field of target detection often use a technique where anchor boxes of different sizes are placed on the image to achieve target detection, regressing and classifying the anchor boxes. Based on the way of regression box generation, target detection algorithms can be divided into two main categories: two-stage algorithms and one-stage algorithms [8].

Well-known two-stage target detection algorithms include R-CNN, Fast R-CNN, Faster R-CNN, etc. The two-stage target

detection algorithm is a commonly used target detection method, which consists of two main stages: candidate image generation (region proposal) and target classification and localization [9].

Region Proposal Phase: First, a series of candidate frames (often called candidate regions or candidate frames) that may contain targets are generated from the input image by using some advanced techniques such as selective search, Edge Boxes, or deep learning-based methods [10]. The goal of this stage is to generate as many candidate frames as possible to cover the potential targets in the image.

Target Classification and Pinpointing Stage: In this stage, a target classifier is applied to each candidate frame to determine whether it contains the target of interest and to pinpoint the target, i.e., to determine the exact bounding box location of the target. For each candidate frame, the classifier discriminates its extracted features and determines whether it belongs to a specific target class. In addition, the target bounding box can be further fine-tuned to more accurately match the location and shape of the target.

In computer vision, single-stage target detection algorithms are widely used and can be classified as a common type of target detection algorithm. They contrast with another common class of target detection algorithms known as two-stage algorithms. One-stage target detection algorithms have the key feature of performing target detection and localization directly on the image itself. This approach is typically faster and more suitable for real-time application scenarios [11]. The following are some common representatives of single-stage target detection algorithms:

The YOLO (You Only Look Once) family: YOLO is a popular collection of single-stage target detection algorithms. By dividing the image into grids and predicting the location and class of targets within each grid, the YOLO algorithm effectively transforms the target detection problem into a regression problem. The YOLO algorithm performs target localization and classification simultaneously by a single neural network, which is faster, and YOLOv5 is the latest version of the YOLO series.

SSD (Single Shot MultiBox Detector): SSD is another popular single-stage target detection algorithm. The SSD algorithm uses multi-scale feature maps to detect targets of different sizes, and uses anchor boxes of different sizes for target position prediction. The SSD extracts image features through a convolutional neural network and performs target classification through subsequent convolutional layers and bounding box regression.

The following is a detailed description of the YOLO family of target detection algorithms (YOLOv1, YOLOv2, YOLOv3, YOLOv4, YOLOv5):

YOLOv1 is the first release in the YOLO series. YOLOv1 transforms the target detection problem into a regression problem by directly predicting bounding box locations and class probabilities through a single forward pass. A convolutional neural network (CNN) is used to extract features from the image and output predictions through a fully connected layer [12]. YOLOv1 has a fast detection speed, but performs poorly on small and dense targets and has limited effectiveness for scenes requiring high localization accuracy.

YOLOv2 incorporates several enhancements to improve the detection capabilities for small objects, such as the introduction of multi-scale prediction and the implementation of the anchor box mechanism. The network structure used in YOLOv2 is Darknet-19, and a novel loss function is employed to effectively deal with classification errors and bounding box regression errors [13]. YOLOv2 also proposes a data enhancement technique called Random Gamma Adjustment.

YOLOv3 introduced a number of improvements over YOLOv2, including the use of a deeper Darknet-53 network structure, the use of residual connectivity, the application of multi-scale prediction, and the use of more anchor frames. YOLOv3 introduced a Feature Pyramid Network (FPN) structure, which extracts feature maps at different scales and performs target detection. By using feature maps of different sizes for target detection, YOLOv3 can effectively detect targets of different sizes [14]. YOLOv3 achieves a better balance between detection speed and detection performance.

YOLOv4 further improves the network structure and training strategy to increase the detection accuracy and speed. CSPDarknet53 is adopted as the backbone network, and modules such as SAM (Spatial Attention Module) and PAN (Path Aggregation Network) are introduced to improve the feature representation and perception capabilities. YOLOv4 also uses multi-scale inference and multi-scale training strategies by combining different YOLOv4 also uses multi-scale inference and multi-scale training strategies to achieve more accurate target detection by combining feature maps of different scales. The GIoU (Generalized Intersection over Union) loss function is introduced to more accurately compute the overlap between bounding boxes [15].

YOLOv5 introduces some improvements in the network architecture with lighter model architecture while maintaining high detection performance. YOLOv5 introduces a new target detection architecture called CSPNet (Cross Stage Partial Network), which balances model size and performance by reducing computational complexity and improving accuracy. YOLOv5 also introduces more data augmentation techniques to improve model generalization. YOLOv5 achieves faster training and inference through model lightweighting and optimization. Overall, the goal of the YOLO family is to achieve real-time target recognition while balancing speed and accuracy. Each version of YOLO has its own unique contributions and enhancements to improve detection performance and efficiency.

B. Current Status of Research

There have been many researchers who have studied helmet detection to propose various optimized detection methods for the above algorithms, for example, Chen et al. presented a novel integration of Retime image enhancement technology into the Faster R-CNN framework. They addressed the challenges posed by various factors such as lighting conditions and distance, which often hinder accurate detection. To overcome these obstacles and improve detection performance, they employed the K-means++ algorithm. By

using this approach, they achieved automatic helmet detection with improved accuracy [16]. Guo et al. added a VGG16 feature extraction module to Faster R-CNN to determine whether the helmet is correctly worn using Euclidean distance, which improved the detection accuracy and speed [17]. Song added the R-SSE module to YOLOv3, reduced the network depth, improved the network detection speed and accuracy, and the dual module Res 2 was used to improve the feature reusability and detection efficiency of small target [18]. Chen et al. used a lightweight network PP-LCNet to improve YOLOv4, reduced the model parameters with depth-separable convolution, used a new SIoU loss function, reduced the model size, and improved the speed of helmet detection [19]. To improve the YOLOv5 detector, Jia et al. used triple attention fusion and soft NMS, which can achieve high accuracy and real-time results under complex weather conditions [20]. Jin et al. used K-means++ algorithm and Deep Coordinated Attention (DWCA) mechanism in YOLOv5 to enhance the information propagation between features and improve the accuracy of helmet detection [21].

In summary, although certain results have been achieved in the field of helmet research, there is little research in the field of reflective vest detection. All of the above studies have optimized and improved the algorithms to some extent, but most of the models are more complex and not easy to implement, or the detection accuracy is difficult to meet the demand. In addition, in the urban engineering construction environment, weather conditions and obstacles can significantly affect the reliability of detection, and more in-depth research is required to obtain a good detection model. This paper proposes to combine the two, studies the simultaneous detection and identification methods, and improves and optimizes the speed and accuracy of real-time detection, which can further protect the personal safety of municipal engineering construction personnel.

III. Improved YOLOv5 Model Architecture

A. Review of the Fundamentals of the YOLOv5 Model

Backbone Network: YOLOv5 uses CSPDarknet as its backbone network. CSPDarknet is a lightweight convolutional neural network that uses a CSP (Cross-Stage Partial) connection structure to divide the feature maps in the channel dimension and bypass some of the convolutional layers, thus reducing the computational complexity.

Neck Network: Instead of an explicit neck network structure, YOLOv5 introduces multiple cross-level connectivity (PANet) modules in the backbone network. These PANet modules perform spatial pyramidal pooling operations on feature maps at different scales to fuse semantic information at different levels.

Head Network: YOLOv5's head network is responsible for generating predictions for target detection. It has three main components: Spatial Pyramid Pooling (SPP) module, feature pyramid pooling layer, and detection head. SPP Module: The SPP module performs multi-scale pyramid pooling operations on the feature graph to capture contextual information at different scales. Feature pyramid pooling layer: The feature pyramid pooling layer combines feature maps at different scales by upsampling and fusion to produce a feature map that incorporates information from multiple scales. Detection head:

The detection head consists of a series of convolutional layers responsible for predicting the target's bounding box location, category, and confidence score [22]. YOLOv5 uses a multi-scale prediction strategy, i.e., target detection is performed on feature maps at different levels to handle targets of different sizes.

Loss function: YOLOv5 uses a loss function called YOLOv5 Loss, which combines the loss calculation of target position, category and confidence.

In general, YOLOv5 is structured with a lightweight backbone network, cross-stage connectivity modules, and a multi-scale prediction strategy to achieve efficient and accurate target detection. It achieves a good balance between speed and accuracy for a variety of real-time or offline target detection tasks. The schematic structure of YOLOv5 is shown in Fig. 1.

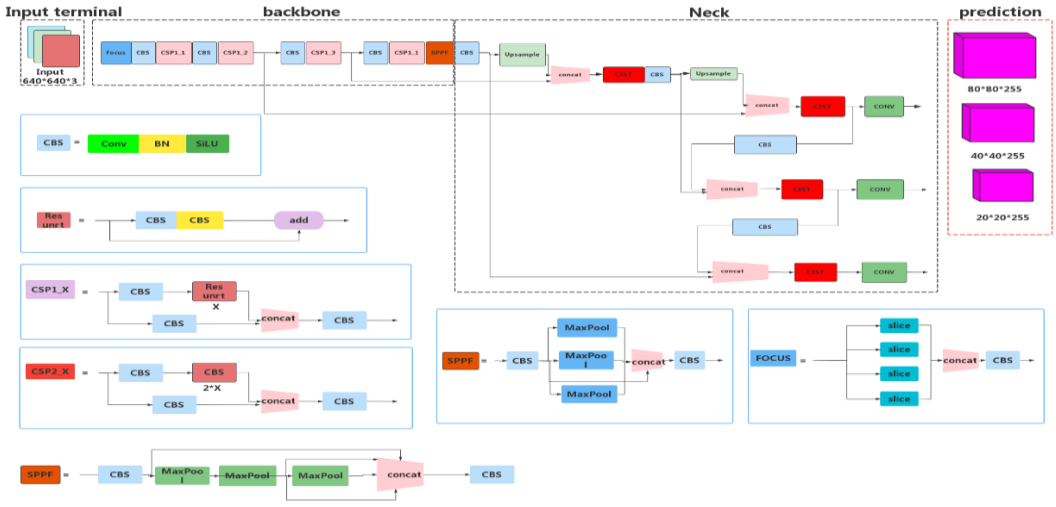

Fig. 1. YOLOv5 principle structure diagram.

B. Improving the YOLOv5 Algorithm

1) Mosaic-6: In YOLOv5, a new data enhancement method called Mosaic has been introduced as a complement to the basic data enhancement techniques. The primary concept behind this method is to randomly crop and scale four images, which are then combined in a random order to create a single image. This method not only enriches the data set, but also increases the number of small sample targets, thus improving the training speed of the network. An important advantage of using the mosaic data enhancement method is that the data from four images can be computed at once when performing the normalization operation, thus reducing the memory requirement of the model [23]. This batch processing reduces memory requirements and allows the model to process large data more efficiently.

By randomly cropping, scaling, and arranging the four images, the mosaic data enhancement method can generate training samples with more diversity and complexity. In this way, the model can better learn how to handle different scenes, scales, and target combinations during training, improving the model's ability to generalize to different situations.

Inspired by the Mosaic data enhancement method, this paper proposes a new data enhancement method called Mosaic-6. In this method, six images are randomly selected for cropping, arranging, and scaling operations to combine them into a new image. The Mosaic-6 detail enhancement is shown in Fig. 2. This can effectively increase the amount of sample data and control the random noise within a reasonable range. In this way, the network model can be improved to distinguish small target samples in images. The final renderings are shown in Fig. 3.

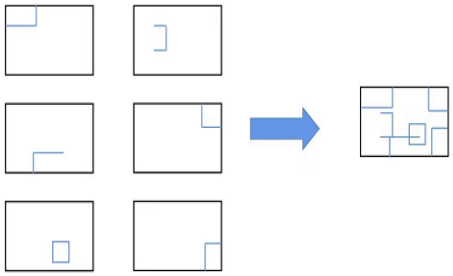

Fig. 2. Mosaic-6 detail enhancement map.

Fig. 3. Mosaic-6 effect.

2) Optimization of attentional mechanisms: Given the nature of the helmet image, where there are numerous small targets that make up a small portion of the overall image, they can be easily influenced by the background and other factors. The original YOLOv5s network tends to lose crucial feature information of these small target helmets during deep convolution, which hinders their effective detection. To enhance the helmet features in the given image, this study introduces a work that assigns different weights to different positions of the image in the channel domain. This approach aims to extract more critical feature information [24]. The paper incorporates the Squeeze-and-Excitation (SE) attention mechanism, which is fused with the input feature map. This fusion produces a feature map with channel attention, which mitigates the loss of information related to small targets in the helmet image.

To ensure the effective functioning of the attention mechanism within the network, the paper integrates the SE attention mechanism into the focus module of the backbone network. This module is located at the front of the network and is placed before the first convolution operation. As a result, the network can prioritize the channel feature information of the target for detection at an early stage, thereby improving its representational capability.

As shown in Fig. 4, the structure of the SEFocus enhanced slicing module consists of two main parts. One part is responsible for slicing the original image to reduce the model computation, as shown in the figure for the slicing module; the other part embeds the channel feature information into the sliced and reconstructed image through the attention mechanism to prevent the channel feature information loss.

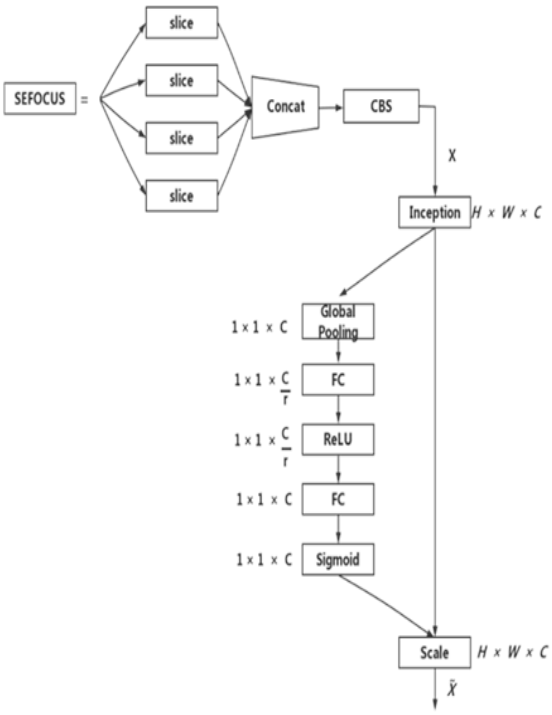

Fig. 4. SEFocus structure diagram.

Transforming the sliced image into an attention image with channel feature information requires three steps. The first step first performs spatial feature compression on the feature map, the second step learns by FC fully connected layer to obtain the channel attention feature map, and the third step outputs the channel attention feature map [25].

In order to reduce the computation, the number of parameters, and the introduction of inter-channel relationships, the squeeze operation is first performed on the input feature map as shown in equation (1), assuming that the input feature map is X with dimensions C*H*W, where C is the number of channels, and H and W are the height and width, respectively.

$$z_c = \mathrm{F}_{sq}(\mathrm{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \qquad (1)$$

This step is completed by two fully connected layers to generate the weight information required in this paper by weight W, which are obtained by learning. Then, the vector z obtained in the previous step is processed through two fully connected layers W_1 and W_2, and the channel weight s is obtained as shown in Equation (2), and after two fully connected layers, the different values in s represent the weight information of different channels, and the channels are given different weights.

$$\mathrm{s}=\mathrm{F}_{ex}(\mathrm{z},\mathrm{W}) = \sigma\big(g(\mathrm{z},\mathrm{W})\big) = \sigma\big(\mathrm{W}_2\delta(\mathrm{W}_1\mathrm{z})\big) \qquad (2)$$

Finally, the weight vector s generated in the second step is assigned to the feature map U, and the feature map is obtained as shown in Equation (3), whose size is exactly the same as that of the feature map U.

$$\tilde{\mathrm{x}}_c = \mathrm{F}_{scale}(\mathrm{u}_c, s_c) = s_c \mathrm{u}_c \qquad (3)$$

3) Swin transformer: When YOLOv5s is used for feature extraction, the typical convolutional neural network CSPDarknet53 is selected to expand the perceptual field by continuously stacking convolutional layers to complete the capture of local information into global information. However, the increasingly deep convolutional layers lead to an increasingly complex model, and with the deepening of the network structure, the position information of the small-scale targets in the helmet dataset may be coarse, and the feature information may be easily lost after multiple convolutional operations [26].

This study presents an improved convolutional structure, as shown in Fig. 5, by incorporating the Swin Transformer into the C3 convolution module.

The aim is to enhance the semantic information and feature representation of small targets by utilizing the self-attentive window module in the feature fusion process.

In the Swin Transformer model, after the linear transformation of the input vectors, the resulting matrix is equally divided into three parts, which become the three features of query vector Q, key vector K and position vector V in the Transformer, and the attention mechanism is calculated as shown in equation (4).

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (4)$$

In the above equation, Attention denotes the attention weight matrix, Softmax denotes the normalized exponential function, QK^T is the matrix obtained by matrix multiplication, d is the feature dimension, and denotes the length of the keyword vector.

The method not only takes into account the displacement invariance, dimensional invariance, perceptual field, and hierarchical relationships characteristic of convolutional neural networks, but also has the ability to extract global information and learn long dependencies. Swin Transformer improves efficiency by limiting self-attentive computation to non-overlapping local windows with moving windows, while allowing cross-window connections [27]. In addition, this hierarchical structure provides the flexibility to model at different scales and improves the ability of the model to capture multiscale information.

In summary, the above three improved methods are applied to the YOLOv5s algorithm, and the improved network structure is shown in Fig. 6.

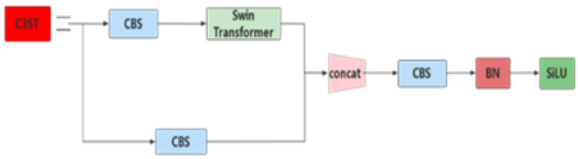

Fig. 5. C3ST structure.

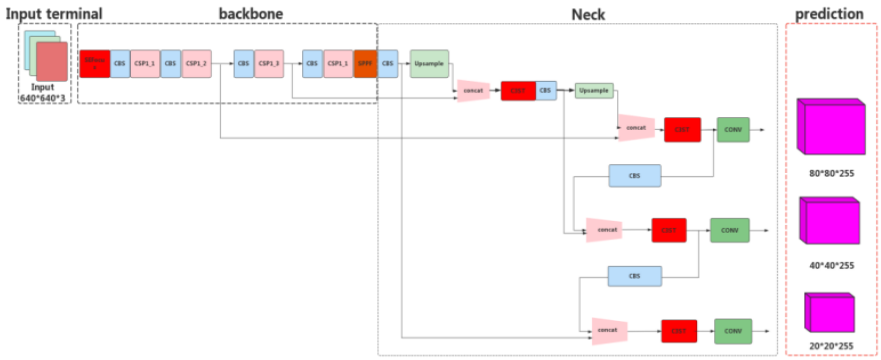

Fig. 6. Improved structure diagram of YOLOv5.

IV. Introduction to the Data Set and Experimental Environment

A. Introduction to the Data Set

1) Dataset sources: In the study of helmet and vest recognition, due to the lack of publicly available large-scale datasets, this paper collects relevant images through on-site photography and from the Internet. On-site photography is a common way to collect data, and this paper chooses to shoot at municipal construction sites and other scenes where helmets and vests are required. On-site photography captures image samples with real environmental backgrounds, different lighting conditions, and different poses, which is very beneficial for the robustness and generality of the algorithm.

In addition, a large number of relevant image resources are available on the Internet. In this paper, we obtain the necessary image information by searching and downloading user-uploaded construction site photos and construction site surveillance video screenshots through engineering websites and social media platforms.

2) Data pre-processing: In the data pre-processing stage, the images that meet the requirements are first converted to .jpg format for further processing and model training. To improve the robustness of the model, some of the positive sample data in this paper are inverted, which can increase the diversity of the data and make the model more generalizable when dealing with mirror image or symmetry problems. The saturation and exposure of the images are also adjusted to further increase the diversity of the data and the adaptability of the model.

Next, manual labeling is performed using the labelImg labeling tool. The construction workers in the image are labeled according to four categories: wearing a helmet (hat), wearing a safety vest (reflective clothes), not wearing a helmet (head), and not wearing a safety vest (other clothes). The labeling tool draws rectangular boxes in the image to frame the target location and assigns the appropriate category label to each box. This creates an XML tag file containing the four coordinates and category information of the target within the frame. The final PASCAL VOC format is obtained, which is a commonly used annotation format for target detection datasets and can be easily compatible and interactive with various deep learning frameworks.

Through the above processing steps, the pre-processed and annotated dataset with a total of 12,000 images is obtained in this paper. Each image has a corresponding image file in .jpg format as well as a tag file in XML format containing the location and class of the target.

B. Experimental Environment and Parameter Settings

In this study, a Linux system was chosen as the experimental development platform and the Ubuntu 18.04 operating system was used.

The experimental environment is equipped with a V100 GPU, and CUDA 10.1 is used as the GPU acceleration library. For the deep learning framework, PyTorch 1.8 was used as the main framework and Python 3.7 as the programming language. To meet the task requirements, the image resolution was set to 640x640, the training batch was set to 16, the training process was set to 100 rounds, and the initial learning rate was set to 0.001. The impulse was set to 0.937.

V. Experimental Results and Analysis

A. Evaluation Metrics

The performance metrics commonly used in target detection algorithms include precision, recall, average precision (AP), mean average precision (mAP), and number of parameters and floating point operations (FLOPs).

Accuracy is the proportion of all samples with a positive prediction that are actually positive. It measures the degree of accuracy of the model in predicting positives. Recall rate, also known as sensitivity or true positive rate, is the ratio of correctly predicted positive samples to the total number of actual positive samples. The formula is shown in equation (5), which measures the ability of the model to detect positive specimens.

$$P = \frac{TP}{TP+FP} \tag{5}$$

Recall is measured by calculating the ratio of correctly detected positive samples (True Positive, TP) to all true positive samples. TP refers to the positive targets that are successfully classified in the model, i.e., the true positive samples that are correctly detected. Equation (6) represents the formula for determining the recall rate.

$$R = \frac{TP}{TP+FN} \tag{6}$$

Here, FN denotes positive samples that were incorrectly classified as negative during the classification process, i.e., true positive samples that were not correctly detected. The recall value, which ranges from 0 to 1, increases as the model becomes better at identifying true positive samples.

Average Precision (AP) is a comprehensive metric that calculates the accuracy of the model at different confidence levels and averages these accuracies. The formula is shown in equation (7) and is used to assess the balance between accuracy and recall of the model.

$$AP = \int_0^1 P(R)dR \tag{7}$$

The mAP is the average of the average accuracies of all categories. It measures the overall performance of the model over multiple categories and is calculated as shown in equation (8).

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{8}$$

The parameter count represents the total number of learnable parameters in the model. Reducing the number of parameters in a model can reduce memory requirements and computational complexity, making model deployment and inference easier. The number of FLOPs quantifies the number of floating point operations required by the model during inference. FLOPs serve as a measure of the computational complexity and inference speed of the model.

B. Presentation of Experimental Results

In this paper, a modified YOLOv5 model is used for security detection in the field of municipal engineering. The effectiveness of the improved model is tested using datasets consisting of homemade helmets and reflective vests to evaluate its detection capabilities. The accuracy, recall, and MAP graphs of this paper are shown in Fig. 7, 8, and 9, respectively.

According to Fig. 7, the accuracy of the proposed algorithm can reach 0.985. The x-axis represents the confidence threshold and the y-axis represents the precision. The overall shape and trend of the curve reflect the detection performance of the algorithm at different confidence thresholds. The improved algorithm in this paper can maintain high precision even at high confidence, indicating good accuracy.

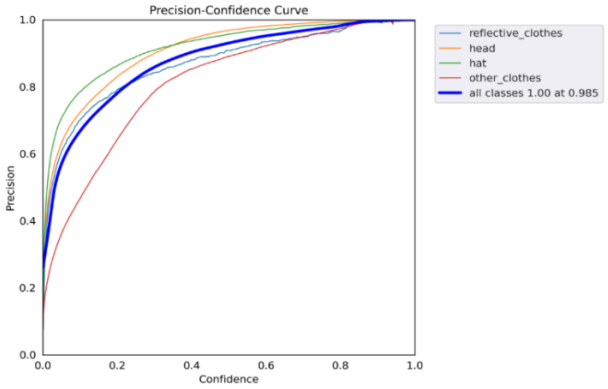

Fig. 7. Accuracy graph.

From Fig. 8, the recall of the proposed algorithm reaches 0.97. The x-axis represents the confidence threshold, and the y-axis represents the recall. The curve describes the recall at different confidence thresholds. Higher recall indicates that the algorithm can detect more targets at a given confidence threshold, but it may be associated with some false positives. Lower recall may indicate that the algorithm misses some targets at a given confidence level, but with a lower false positive rate. The proposed algorithm maintains high recall even at high confidence.

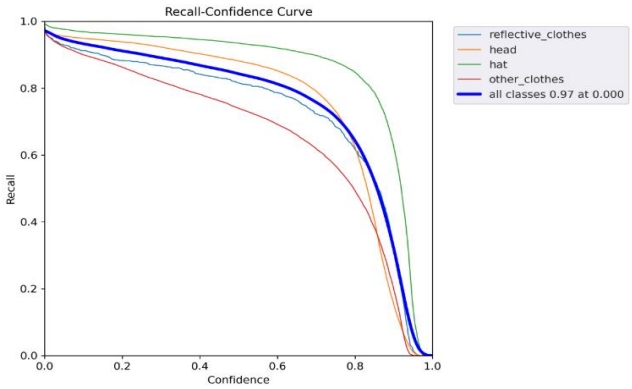

Fig. 8. Recall rate graph.

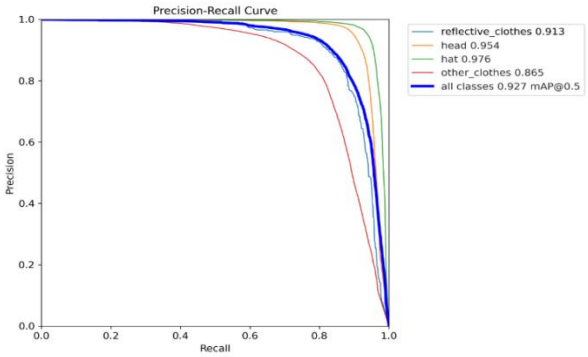

Fig. 9. mAP graph.

According to Fig. 9, the Mean Average Precision (mAP) value in this paper reaches 0.927. The x-axis represents recall and the y-axis represents precision. The curve is plotted by connecting points representing precision and recall at different thresholds. The position of the curve closer to the upper right corner indicates higher precision and recall at different thresholds, implying better performance of the model. This indicates that the model proposed in this paper has good performance.

C. Analysis of Experimental Results

As shown in the table, in this paper, the operation of adding each of the three improvement modules one by one is performed and tested according to the three evaluation metrics of P, R and mAP. Among them, optimization model 1 modified the Mosaic-4 data enhancement method of the original model, optimization model 2 modified the focus structure of the original model, and optimization model 3 modified the C3 module of the feature fusion part of the original model. From the data in the table, the mAP value of Optimized Model 1 is improved by 1.3%, which indicates that the improved Mosaic method enriches the context in which the detection objects appear, which makes the model better learn the local features of the detection objects and optimizes the training of the model. The optimized model 2 mAP value is improved by 2%, indicating that the model can focus on the channel feature information of the target to be detected earlier by adding the SE attention mechanism in the focus part, which effectively reduces the problem of small target feature information loss. The mAP value of the optimized model 3 is improved by 1.8%, which proves that Swin Transformer can effectively improve the global perception of small features of the detected targets. In this paper, the combined use of three enhancement methods, whose P, R, and mAP are improved by 5.3%, 5.7%, and 3.4%, respectively, optimizes the detection effect of small and dense targets in complex scenes.

To better demonstrate the advantages of the algorithms in this paper, the improved algorithms in this paper are compared with several mainstream algorithms for experiments, including Faster R-CNN, SSD, YOLOv3, YOLOv4, YOLOv5s, and the improved algorithms in this paper. To ensure the fairness of the experimental results, the experimental environment, including experimental settings such as input image size, learning rate, and momentum, should be kept identical for the comparative experiments. As shown in Table I, with P, R, mAP, Parameters, and FLOPs as evaluation metrics.

TABLE I. OPTIMIZATION RESULTS OF EACH IMPROVEMENT MODULE FOR YOLOV5S

Algorithm model	Mosaic-6	SEFocus	Swin Transformer	P/%	R/%	mAP/%
YOLOv5s				93.2	91.3	89.3
Optimized model 1	√			94.1	91.5	90.6
Optimized model 2		√		95.4	90.6	91.3
Optimized model 3			√	96.4	95.3	92.1
Algorithm in this paper	√	√	√	98.5	97.0	92.7

TABLE II. COMPARISON OF DIFFERENT ALGORITHMS ON THE DATA SET OF THIS PAPER

Algorithm model	P/%	R/%	mAP/%	Parameters /M	FLOPs/G
Faster-RCNN	87.3	79.5	83.4	136.5	396.2
SSD	82.5	73.6	79.7	24.2	281.3
YOLOv3	84.3	77.4	81.2	116.3	55.1
YOLOv4	85.1	76.3	83.0	101.3	54.7
YOLOv5s	93.2	91.3	89.3	7.05	16.6
Algorithm in this paper	98.5	97.0	92.7	7.6	21.8

As shown in Table II, the algorithm in this paper has a better detection performance with an average accuracy mAP of 9.3%, 13%, 11.5%, 9.7%, and 3.4% over the constructed helmet and reflective vest datasets than the Faster-RCNN, SSD, YOLOv3, YOLOv4, and YOLOv5s algorithms, respectively. In terms of model computation size, this paper is slightly larger than YOLOv5s, but much smaller than other mainstream algorithms, especially the two-stage detection algorithm. In addition, the algorithm in this paper has a great improvement in accuracy (P) and recall (R). Overall, compared with the YOLOv5s model, the algorithm in this paper is slightly larger in model size, but has a significant improvement in accuracy, and still meets the demand for real-time detection, and also significantly improves the detection performance, with significant overall performance advantages.

In order to visually compare the detection effect, this paper performs the detection of the same image with the YOLOv5 model and the improved algorithm, and shows the comparison results, as shown in Fig. 10. The left side of the figure shows the detection effect of the YOLOv5 model, and the right side shows the detection effect of the improved algorithm. By comparing the results of the two, the performance improvement of the improved algorithm on the target detection task can be clearly observed. Such a comparison helps to illustrate the effectiveness of the improved algorithm proposed in this paper in improving the detection results.

(a) Missed detection.

(b) Dark scenario.

(c) Dense targets.

Fig. 10. Comparative detection.

As shown above, in image (a), the original YOLOv5 model has failed to detect and identify whether reflective vests are worn or not, while the improved model can successfully detect them. The second set of images (b) is a dim scene, and the improved model can detect the protective gear worn by the construction workers even in very low light. The last set of images (c) is dense target detection, and the algorithm is able to accurately identify the presence of personnel occlusion. In short, the improved algorithm has been greatly improved in the case of missed detection, dim scenes and dense targets, which are not conducive to the correct identification of target detection.

VI. Conclusion

This paper proposes a YOLOv5s-based algorithm for the simultaneous detection of helmet and reflective vest wearing, which provides a new technological means for urban engineering supervision. First, the Mosaic-4 data enhancement method used in YOLOv5s is improved, and the Mosaic-6 data enhancement method is used to give synthesized images more complex backgrounds, which is used to enhance the learning ability of the model for local features, thus improving the generalization ability of the whole model. Second, the attention mechanism is introduced in the focus module to enhance the perceptual field of the model and improve the correlation between the channel information and the detection target features. Finally, the feature fusion part of the original model is improved by adopting a new hierarchical construction method inspired by the Swin Transformer. This enhancement aims to improve the semantic information and feature representation of the helmet mini-target by incorporating the window self-attention module, which effectively replaces the C3 module.

The experimental results show that the enhanced approach presented in this research paper is capable of accurately identifying helmets and reflective vests, and the average accuracy can reach 92.7%, which is 9.3, 13, 11.5, and 9.7 percentage points higher than the average of Faster-RCNN, SSD, YOLOv3, and YOLOv4 models, respectively. Li et al. improved the YOLOv5s network structure and loss function. The improved algorithm has a map value of 92.5%, a precision rate of 87.7%, and a recall rate of 86.8% [28]. Compared with this algorithm, the map value is only 0.2% higher. However, our algorithm shows a significant advantage in terms of accuracy, exceeding by 10.8%, as well as recall, exceeding by 10.2%. Although there is a slight increase in model size, the accuracy and recall rates are significantly improved and still meet the requirements of real-time detection with high application value.

VII. Future Outlook

While the improved algorithm in this study has achieved certain results in the detection of safety gear wearing in municipal construction projects, the authors acknowledge that there is still room for improvement. This study focused only on the method for detecting the wearing of safety protective equipment. In practical applications, it is necessary to design and develop a standardized system that includes the interaction interface between the backend algorithm and the frontend system. This will enable real-time detection of construction sites and ultimately establish a complete system. In future work, the authors will continue to strive for higher detection accuracy and explore more effective algorithms and techniques. Additionally, the authors will design and develop the system to ensure the algorithm's compatibility with real-world application environments. By considering both accuracy and real-time requirements, the goal is to provide a high-performance and reliable system that offers a feasible and effective solution for detecting the wearing of safety protective equipment in the field of municipal construction projects.

Acknowledgment

The authors would like to thank the reviewers for taking the time to guide the completion of this paper, and also thank the Education Department of Jilin Province, China for their financial support.

References

[1] Mneymneh B E, Abbas M, Khoury H. Vision-based framework for intelligent monitoring of hardhat wearing on construction sites[J]. Journal of Computing in Civil Engineering, 2019, 33(2): 04018066.

[2] Park M W, Elsafty N, Zhu Z. Hardhat-wearing detection for enhancing on-site safety of construction workers[J]. Journal of Construction Engineering and Management, 2015, 141(9): 04015024.

[3] R. Kamal, A. J. Chemmanam, B. A. Jose, S. Mathews and E. Varghese, "Construction Safety Surveillance Using Machine Learning," 2020

International Symposium on Networks, Computers and Communications (ISNCC), Montreal, QC, Canada, 2020, pp. 1-6.

[4] Nath, Nipun D., Amir H. Behzadan, and Stephanie G. Paal. "Deep learning for site safety: Real-time detection of personal protective equipment." Automation in Construction 112 (2020): 103085.

[5] Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., Rose, T. M. and An, W.. "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos." Automation in construction 85 (2018): 1-9.

[6] Kamal, R., Chemmanam, A. J., Jose, B. A., Mathews, S. and Varghese, E. "Construction safety surveillance using machine learning." 2020 International Symposium on Networks, Computers and Communications (ISNCC). IEEE, 2020.

[7] Li, J., Zhao, X., Zhou, G. and Zhang, M. "Standardized use inspection of workers' personal protective equipment based on deep learning." Safety science 150 (2022): 105689.

[8] Rajeshwari P, Abhishek P, Srikanth P and Vinod, T. Object detection: an overview[J]. Int. J. Trend Sci. Res. Dev.(IJTSRD), 2019, 3(1): 1663-1665.

[9] Ansari M F, Lodi K A. A survey of recent trends in two-stage object detection methods[C]//Renewable Power for Sustainable Growth: Proceedings of the International Conference on Renewable Power (ICRP 2020). Singapore: Springer Singapore, 2021: 669-677.

[10] Du L, Zhang R, Wang X. Overview of two-stage object detection algorithms[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1544(1): 012033.

[11] Aksoy T, Halici U. Analysis of visual reasoning on one-stage object detection[J]. arXiv preprint arXiv:2202.13115, 2022.

[12] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," Paper presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, 2016.

[13] J. Redmon and A. Farhadi, "Yolo9000: Better, Faster, Stronger," paper presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7263-7271, 2017.

[14] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[15] A. Bochkovskiy, C.-Y. Wang and H.-Y. M. Liao, "Yolov4: Optimal Speed and Accuracy of Object Detection," ArXiv, vol. abs/2004.10934, 2020.

[16] Chen, S., Tang, W., Ji, T., Zhu, H., Ouyang, Y. and Wang, W. "Detection of safety helmet wearing based on improved faster R-CNN." 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020.

[17] Guo, S., Li, D., Wang, Z. and Zhou, X. "Detection of safety helmet wearing based on improved faster R-CNN." Artificial Intelligence and Security: 6th International Conference, ICAIS 2020, Hohhot, China, July 17-20, 2020, Proceedings, Part II 6. Springer Singapore, 2020.

[18] Song, Hongru. "Multi-scale safety helmet detection based on RSSE-YOLOv3." Sensors 22.16 (2022): 6061.

[19] Chen, J., Deng, S., Wang, P., Huang, X. and Liu, Y. "Lightweight helmet detection algorithm using an improved YOLOv4." Sensors 23.3 (2023): 1256.

[20] Jia, W., Xu, S., Liang, Z., Zhao, Y., Min, H., Li, S. and Yu, Y. "Real-time automatic helmet detection of motorcyclists in urban traffic using improved YOLOv5 detector." IET Image Processing 15.14 (2021): 3623-3637.

[21] Jin, Z., Qu, P., Sun, C., Luo, M., Gui, Y., Zhang, J. and Liu, H. "DWCA-YOLOv5: An improve single shot detector for safety helmet detection." Journal of Sensors 2021 (2021): 1-12.

[22] Liao, Zhihao and Ming Tian. "A bird species detection method based on YOLO-v5." 2021 International Conference on Neural Networks, Information and Communication Engineering. Vol. 11933. SPIE, 2021.

[23] Zhou, Fangbo, Huailin Zhao and Zhen Nie. "Safety helmet detection based on YOLOv5." 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA). IEEE, 2021.

[24] Zhao, Yu, Yuanbo Shi and Zelong Wang. "The improved YOLOV5 algorithm and its application in small target detection." Intelligent Robotics and Applications: 15th International Conference, ICIRA 2022, Harbin, China, August 1-3, 2022, Proceedings, Part IV. Cham: Springer International Publishing, 2022.

[25] Yao, H., Dong, P., Cheng, S. and Yu, J. "Regional attention reinforcement learning for rapid object detection." Computer and Electrical Engineering 98 (2022): 1077-47.

[26] Wang, Kun, Maozhen Liu and Zhaojun Ye. "An advanced YOLOv3 method for small-scale road object detection." Applied Soft Computing 112 (2021): 107846.

[27] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[28] Li, Weihao and Yan Wei. "A lightweight YOLOv5 model used for safety helmet and reflective clothing detection." 2022 2nd International Conference on Algorithms, High Performance Computing and Artificial Intelligence (AHPCAI). IEEE, 2022.