Enhancing YOLOv5 with Swin Transformer and Multi-Scale Attention for Improved Helmet Detection in Power Grid Construction Sites

Jindong He[1,2], Tianming Zhuang[1,*], Jianwen Min[1], Botao Jiang[1], Chaosheng Feng[3], Zhen Qin[1]

[1] Network and Data Security Key Laboratory of Sichuan Province, University of Electronic Science and Technology of China, 610054, Chengdu, Sichuan, China
[2] Digital Technology Research Center, Fujian Power Co., Ltd. Electric Power Research Institute, 350007, Fuzhou, Fujian, China
[3] School of Computer Science, Sichuan Normal University, 610101, Chengdu, Sichuan, China
202021090101@std.uestc.edu.cn

Abstract. With the progress of computer vision technology, helmet wearing detection has become increasingly important for site safety, particularly in complex environments like power grid construction sites. However, this remains challenging due to issues with incorrect and missed detections in crowded environments. To it, in this paper, we introduce a novel approach that refines the YOLOv5 network with swin Transformer. This method is able to address the limitations of YOLO's convolutional architecture, which struggles with long-range dependencies and dense target detection. Our hybrid strategy combines the Transformer's ability to capture global dependencies with YOLO's processing speed, resulting in a robust and real-time detection system for power grid environments. Additionally, a novel Multi-Scale Convolutional Attention (MSCA) module is proposed, which overcomes the single-scale focus of existing attention mechanisms. By integrating attention across various scales, the MSCA module captures the semantic richness of different feature sizes, enhancing the model's performance in long-range contextual understanding and fine-grained semantic awareness. Extensive experiments are conducted on the safety helmet wearing detect and VOC2028-SafeHelmet datasets, the superior performance validates the effectiveness and generalization of our proposed method.

Keywords: safety helmet wearing · target detection · vision Transformer · multi-scale representations

1 Introduction

The helmet is a head protection product used in the workplace, which protects the user's head from injuries caused by falling objects or small splashing objects and other factors. The recurrence of safety accidents linked to non-compliance with helmet use has led to both personal harm and significant corporate losses. Therefore, the development of effective helmet-wearing monitoring algorithms is of paramount importance, offering both academic value and practical applications.

As computer vision technology advances, the detection of helmet wearing has become a critical research area for ensuring site safety. Target detection algorithms are designed to distill semantic information from images, enabling efficient applications in real-world scenarios, such as autonomous vehicles[1], security systems[2] and intelligent robotics[3], etc. However, achieving high accuracy in helmet detection on power grid construction sites remains challenging due to issues with incorrect and missed detections in crowded environments. Addressing this, our work refines the YOLOv5 network to enhance the precision of detecting helmet-wearing workers in these complex scenarios. As the model size of YOLOv5 is smaller than YOLOv9, and the computational complexity is also lower, making deployment of YOLOv5 on mobile devices easier.

The Transformer model, with its self-attention mechanism, adeptly handles long-range dependencies and contextual cues, effectively managing target occlusion and dense detection in complex scenes. Conversely, the YOLO network's convolutional architecture limits its ability to capture such dependencies, and its global receptive field struggles with the intricacies of dense target arrangements, often resorting to local information for detection. This can lead to merged detections of multiple targets in crowded scenarios. To address helmet-wearing detection challenges in power grid environments, where small and occluded targets or densely packed individuals can lead to mis-identifications and missed detections, we propose a hybrid approach merging the Transformer's dependency capture capabilities with YOLO's swift processing. This fusion enhances the accuracy and robustness of helmet detection in dense power grid settings without compromising real-time efficiency.

Attention mechanisms enhance computer vision models by emphasizing crucial image features. However, existing methods often focus on a single scale, limiting their effectiveness for varied target sizes and details. Our multi-scale convolutional attention module addresses this by integrating attention across different scales, harnessing the semantic richness of varied feature sizes. This approach leverages the robust feature extraction of convolutional neural networks, significantly boosting model performance in target detection and recognition within complex scenes.

The main contributions of our approach can be summarized as follows:

- A novel YOLOv5-enhanced strategy with Transformer is proposed, which creates a C3W-T structure that integrates Swin Transformer blocks into CSPDarknet53. This design enables to expand the network's receptive field, enhancing its capability in capturing global information and contextual richness, thereby addressing YOLOv5's detail loss in complex, large-scale scenes and bolstering detection accuracy.
- A novel Multi-Scale Convolutional Attention (MSCA) module is proposed, which can replace the traditional self-attention, efficiently aggregating local information and capturing multi-scale contexts. Enlightened by the SegNeXt design[4], this module improves long-distance contextual understanding and

fine-grained positional awareness, mitigating detection errors in safety helmet recognition within power grid operations.

- Our overall framework, integrating the YOLO fusion transformer strategy and the MSCA module, significantly elevates safety helmet detection accuracy in power grid scenarios. Through extensive experiments on relevant datasets, our method surpasses state-of-the-art (SOTA) models, validating its effectiveness and broad applicability.

2 Relative works

In this section, we will have a brief review about the approaches related to one-phase-based and two-phase-based helmet detection in the following.

2.1 One-stage target detection

One-stage target detection methods, such as YOLO, SSD, and Retina-Net, eliminate the need for a region proposal stage, directly outputting object class probabilities and positional coordinates.

YOLOv3[5], introduced in 2018, increased model complexity for a balance between speed and accuracy, outperforming R-CNN series in detection speed, making it suitable for time-sensitive tasks like helmet detection. Researchers have since refined YOLOv3-Tiny[6] by adjusting input size and anchor frame proportions, reducing layer counts to optimize feature extraction and resource usage for real-time detection on edge devices. Cheng et al.[7] further enhanced YOLOv3-Tiny with separable convolution and channel attention mechanisms, replacing max pooling and improving the loss function for better performance on custom datasets. Building upon the prior versions, YOLOv5 sums four independent feature layers and expands input channels for improved dimensional integration. Zhou et al.[8] applied YOLOv5 to construction site inspections, achieving high speed and accuracy, though the model's generalization for special environments needed improvement. Ma et al. [9] added a small target detection layer to YOLOv5, significantly increasing Mean Average Precision (mAP) on both open-source and non-open-source datasets, with easier model deployment on mobile devices, enhancing the monitoring efficiency of unsafe behaviors at construction sites.

2.2 Two-stage based detection method

Candidate region-based helmet wearing detection method is also called two-stage target detection method, which is divided into two steps of extracting the object region and CNN classification and recognition of the region. Chen et al.[10] cited Retinex image enhancement technology to improve the image quality of outdoor scene of substation for electric power construction scenario, and K-means++ algorithm is used to apply the helmet target of small size, and the method has an 8.1% detection accuracy. detection accuracy of 8.1%, effectively overcoming the interference of light, distance and other factors, while being able to identify the situation of multiple people wearing, due to the relatively single application scene, the method's generalization ability is poor, and it can't be widely used in other scenes; Rabbi et al. [11]to solve the problem of detecting the detection performance of small-sized targets in remote sensing images, GAN-based image enhancement technology is introduced, which significantly improves the image quality in the low-illumination and high dynamic range scenes. In addition, by applying the improved non-maximal suppression (NMS) strategy and multi-scale detection strategy, the model is able to effectively distinguish multiple pedestrian targets in a dense scene and enhance the detection rate of small targets. Although the method performs well in the field of urban traffic monitoring and achieves a double improvement in detection speed and accuracy, the model training and preprocessing phases are more complex and have a higher demand for computational resources, which restricts its scope of application on resource-constrained devices.

Current helmet detection research combines image recognition, target detection, and segmentation techniques, with advancements such as separable convolution and channel attention mechanisms showing promise on construction sites but still requiring improvement for general industrial scenarios and small target detection. In this paper, the proposed C3W-T+MSCA framework aims to overcome the limitations of detail capture and generalization in complex environments, significantly enhancing the accuracy and efficiency of safety helmet detection and offering a novel direction for the field.

3 Methodology

In this work, we propose a novel approach that addresses the detail loss issue in traditional detection algorithms for complex scenes by integrating the Swin Transformer and a multi-scale convolutional attention (MSCA) module with YOLOv5. The Swin Transformer broadens the network's sensory capabilities and global information capture. Then, the MSCA module is proposed to refine target recognition by synthesizing local details across scales. This integration streamlines feature extraction and bolsters the model's efficacy in detecting small to medium targets within complex environments, particularly improving the precision and speed of identifying standardized helmet use by workers in power grid operations.

3.1 Integration of Swin Transformer modules

In order to solve the problem of YOLOv5 losing detail information when processing large-scale complex scene images, we integrate the Swin Transformer Block into the CSPDarknet53 structure of the original network, and the improved C3W-T structure is able to extend the network's sensory field to enhance the receptive range, better capture global information and enrich contextual information.

The structure of the Swin Transformer module includes Layer Normalization (LN), Window-based Multihead Self-Attention (W-MSA), Shift-Window Multihead Self-Attention (SW-MSA), and a two-layer Multi-Layer Perceptron (MLP) with a GELU nonlinear in the middle, as shown in Fig. 1.

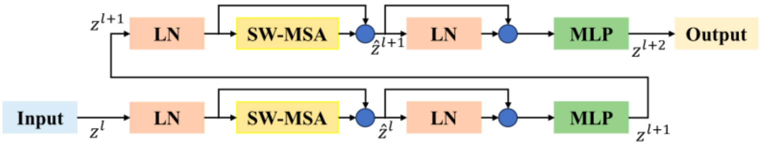


Fig. 1. Swin Transformer structure

However, the LN layer used in conventional CNNs may destroy the learned sample features, and the Swin Transformer module introduced in this paper applies residual connectivity directly after the W-MSA and SW-MSA and does not pass through the LN layer. Therefore, we introduce the Swin Transformer module in the CSP structure, as shown in Fig. 2, to replace the C3 module in the YOLOv5 backbone and necking.

This architecture refines features through a CBS block followed by multiple Bottleneck layers. The Transformer's output is merged with parallel CBS paths, synthesizing a diverse set of feature representations. The design leverages the Swin Transformer's strengths to capture features pertinent to helmet detection in power grid operations, enhancing feature mobility without compromising on efficiency.

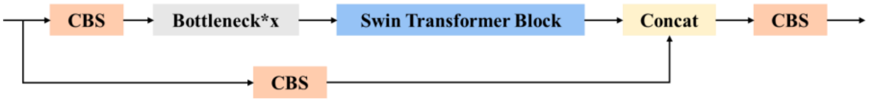
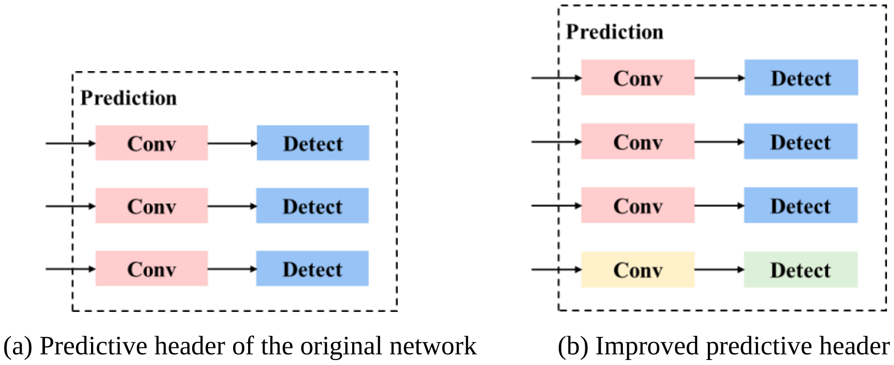


Fig. 2. Improved CSP structure

YOLOv5 employs feature maps of three varying sizes to detect objects of diverse scales, as depicted in Fig. 3. However, detection can fail when object dimensions fall below 8 pixels, leading to inadequate feature extraction. Processing a vast number of images with simple downsampling risks losing critical data due to an oversized downsampling factor. Conversely, an excessively small downsampling factor burdens the GPU with excessive feature map storage, potentially causing memory overflow and hindering training and inference. To optimize performance in outdoor power grid operations, we've enhanced YOLOv5 by appending a multi-scale feature extraction layer and a feature fusion layer after the initial feature extraction, evolving it into a four-tiered detection architecture as shown in Fig. 3. This configuration leverages deeper network layers for feature extraction, bolstering the model's capacity for multi-scale learning and enabling it to capture and assimilate multi-level target features more effectively, thereby refining helmet detection accuracy within grid operation scenarios.



(a) Predictive header of the original network (b) Improved predictive header

Fig. 3. Comparison of the original YOLOv5 with the improved network prediction header in this work

3.2 Multiscale Convolutional Attention (MSCA)

Existing attention mechanisms, which use pooling and nonlinear transformations to direct the detector's attention to the target region, tend to rely heavily on previous downsampling layers to define the region of interest, which may lead to the loss of details and semantic information of some objects, and the detection accuracy is often poor due to the uncertainty of the personnel density and the size of helmet targets in the images of outdoor working scenes of the power grid. To solve this problem, we introduce a novel multi-scale convolutional attention (MSCA) module to replace the self-attention mechanism based on the original network, as shown in Fig. 4.

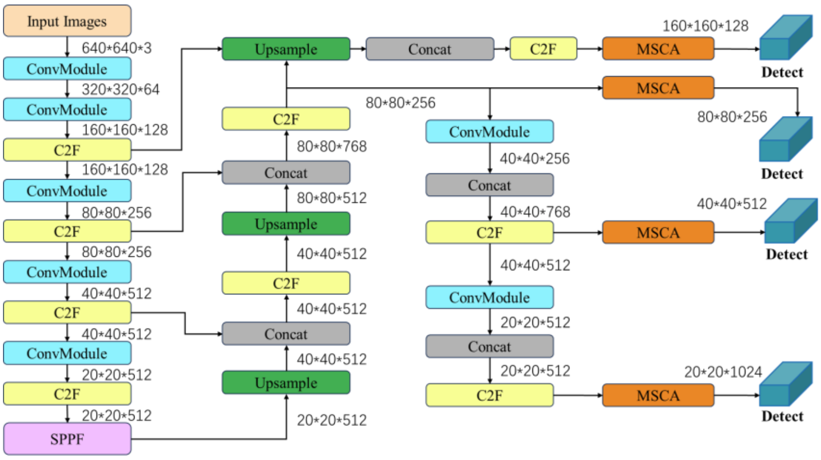


Fig. 4. YOLOv5 network after introduction of MSCA

The multi-scale convolutional attention (MSCA) module comprises three key elements: a channel-separated convolution that consolidates local details, a multi-branch channel-separated strip convolution designed to seize multi-scale contexts, and a 1x1 convolution that facilitates inter-channel relationship mapping. Drawing mathematical inspiration from the Channel-Attention (CA) model, our MSCA incorporates specialized pooling layers to selectively filter information along the horizontal and vertical axes of each channel. Specifically, we apply a (H, 1) pooling kernel to extract horizontal insights and a (1, W) kernel to gather vertical cues. This dual-direction pooling strategy effectively captures and integrates spatial information across both dimensions, described as:

$$Z_h^c = \frac{1}{W}\sum_{i=0}^{W-1} x_{h,i}^c \quad , \tag{1}$$

$$Z_v^c = \frac{1}{H}\sum_{i=0}^{H-1} x_{i,w}^c \quad , \tag{2}$$

$Z_h^c Z_h^c$ and Z_v^c denote the processed horizontal and vertical feature information, $x_{h,t}^c x_{h,i}^c$ and $x_{i,w}^c$ are the horizontal and vertical pixel values of the original feature maps on a specific channel c, and W and H represent the width and height of the feature maps, respectively. This design allows the model to capture long-range contextual relationships in one dimension while maintaining accurate positional information in the other, thus making the model more flexible and responsive in the detection process.

However, it is not enough to consider spatial attention alone. To ensure that the model captures features at different scales and fully considers them at different spatial scales, we further introduce 3×3 and 5×5 convolutional layers. The smaller 3×3 convolutional kernel captures more detailed local information, while the larger 5×5 convolutional kernel captures a wider range of contextual information.

By fusing and transforming all the above features, our attention module achieves a balance: it can capture long-range contextual relationships in one dimension while maintaining precise positional information in the other. Its mathematical expression is as follows:

$$F_l = \sum_{i=1}^{N} Conv_i(X) \quad , \tag{3}$$

where the size of each convolutional kernel is defined. This design aims to capture various features in the image from fine-grained to coarse-grained. After completing the above feature extraction, we first merged the feature maps obtained by the Z_h^c and Z_v^c processes, and then obtained the following results by a shared 1×1 convolutional transform:

$$f = \delta(F_1(Z_h^c, Z_v^c)), \tag{4}$$

δ is a nonlinear activation function, and f can be viewed as an intermediate feature map that encodes horizontal and vertical spatial information. This design uses intermediate feature maps that encode horizontal and vertical spatial information to help the model better understand the target region in the picture.

To effectively detect target features of various shapes and scales. To better understand and capture the long range interactions between these features and further enhance the model's ability to capture spatial relationships, the raw features are fed into two convolutional layers of different scales for feature weight generation. The results of the operation are shown below:

$$g_h = \sigma(F_h(f_h))$$
$$g_w = \sigma(F_w(f_w)), \tag{5}$$

The intermediate feature map is divided into two tensors$f_h$andf_w, each addressing spatial information along different dimensions. These tensors are processed by separate 1×1 convolutions to match the input channel number and then passed through a nonlinear activation to produce attention weights $g_h$and g_w. These weights are applied to the original feature map to create an attention-enhanced version. The horizontal and vertical attention weights are multiplied element-wise with the feature map, adjusting the emphasis on different spatial locations. This process refines the model's focus on critical semantics, particularly small target regions, to enhance helmet detection accuracy in grid operation scenarios by capturing detailed spatial information and multi-scale features.

4 Experiment

4.1 Dataset

The Safety helmet (hardhat) wearing detect dataset (SHWD) includes 7581 images with 9044 human safety helmet wearing objects(positive) and 111514 normal head objects(not wearing or negative). It has been the widely-used benchmark datasets in safety helmet wearing detection tasks.

4.2 Experimental Settings

The extensive experiments are conducted on PyTorch deep learning framework with i7-9700k CPU, 2080ti GPU. During the training phase, the batch size and epoch are configured to 16 and 200, with the learning rate initialized to 0.01. We select SGD optimizer with momentum to optimize the overall network.

4.3 Evaluation Metrics

In order to better reflect the good properties of the proposed method, we use precision (P), recall (R), inference time, and mean-on-average precision (mAP) as quantitative evaluation metrics. Mathematically,

$$P = \frac{TP}{(TP+FP)} \ , \tag{4}$$

$$R = \frac{\mathrm{TP}}{(\mathrm{TP+FN})} \ , \tag{5}$$

Where TP represents the number of targets correctly detected by the model, FP represents the number of targets incorrectly detected by the model, and FN represents the number of correct targets missed by the model.

Mean Average Precision (mAP) is a comprehensive evaluation metric for multi-class object detection tasks. It yields an overall evaluation value by calculating the integral of precision and recall for each class with the following formula:

$$AP = \int_0^1 p(r)dr \ , \tag{6}$$

$$\mathrm{mAP} = \frac{1}{n_j \sum_{j=1}^{n_j} AP_j} \ , \tag{7}$$

4.4 Ablation Study

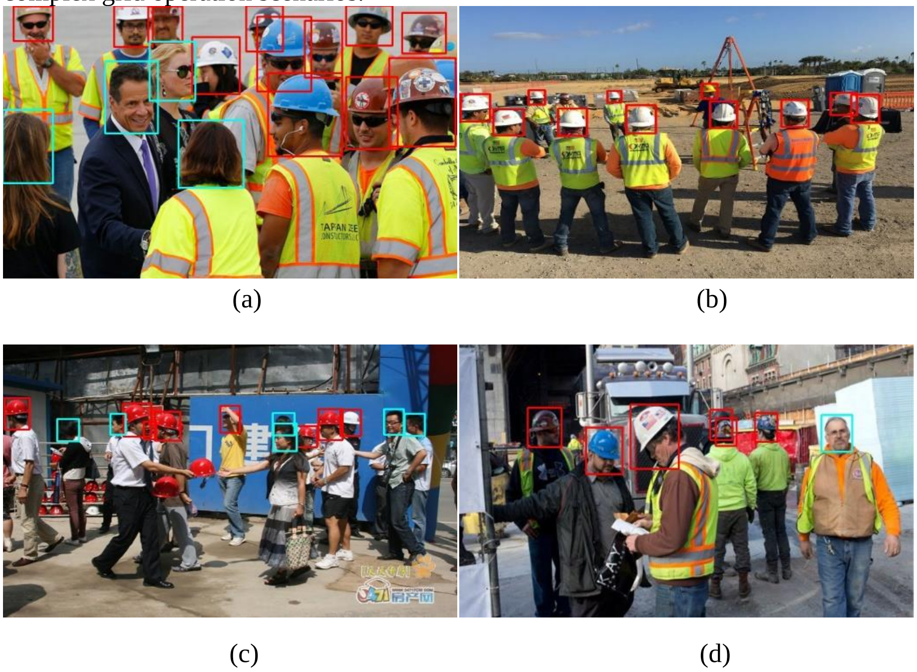
We performed ablation experiments on the Safety helmet wearing detect and VOC2028-SafeHelmet datasets, comparing the original YOLOv5 with versions enhanced by the Swin Transformer (ST-YOLOv5) and MSCA (MSCA-YOLOv5) modules, as well as a combined approach (STMA-YOLOv5). Using identical training data and parameters, the evaluation results are presented in Table 1.

Table 1. Test results after adding different modules

Method	Precision(%)	Recall(%)	mAP50-95(%)
YOLOv5	0.939	0.89	0.939
ST-YOLOv5	0.938	0.896	0.94
MSCA-YOLOv5	0.941	0.904	0.947
STMA-YOLOv5	0.944	0.899	0.948

The ablated results are shown in Table 1. It can be seen that, each part of the improvement in this paper has led to an improvement in the detection accuracy of the network. The simultaneous introduction of both mechanisms improves the mAP50-95 value to 94.8%, which indicates that the proposed work in this paper can significantly improve the performance of helmet detection in grid operation scenarios.

In order to visualize the detection performance of the proposed method, we have selected the images from the VOC2028-SafeHelmet dataset for qualitative verification. As shown in Fig.5, the proposed method can recognize most of the targets in the images, indicating the superior recognition capability in dealing with helmet recognition in complex grid operation scenarios.



(a) (b)

(c) (d)

Fig. 5. Recognition results of the improved method in real scenarios

4.5 Comparison with other state-of-the-art works

In order to further quantitatively evaluate the performance of the proposed work, the method proposed in this paper is compared with other leading approaches on the dataset, and the results are reported in Table 2. Compared with RoI-Transformer approach, our proposed method can exceed by 25.24% in mAP score. The experimental results show that our proposed method achieves the highest mean average accuracy mean among all the algorithms, demonstrating its effectiveness and generalization.

Table 2. Comparison with other state-of-the-art works

Method	mAP (%)
TPH-YOLOv5[12]	71.4
ViT-YOLO[13]	73.1

MDCF2Det[14]	72.86
RoI-Transformer[15]	69.56
Our work	94.8

5 Future Work

Due to objective limitations, our proposed method has achieved good performance through experimental verification on large-scale public datasets. In the future work, we will go to the power grid site to collect large-scale real scene images and conduct further algorithm adjustments and experiments. In addition, the large-scale dataset we are currently using already includes different weather conditions, lighting conditions, and different types of helmets, but our consideration of extreme external conditions is limited, which is a direction worth further research in the future.

6 Conclusion

In this paper, we introduce a novel target detection paradigm by integrating the YOLOv5 network with swin Transformer. We enhance the CSPDarknet53 structure with a customized Swin Transformer block and replace YOLOv5's CSP backbone and neck structures for further capturing global and local information semantics. Additionally, a multi-scale convolutional attention module is proposed, which enables to efficiently aggregating local information and capturing multi-scale contexts. The proposed method is validated on the Safety helmet wearing detect and VOC2028-SafeHelmet datasets, achieving notable improvements in mAP, recall, and precision metrics. The extensive experimental results prove the effectiveness and feasibility of the proposed method in detecting helmets within grid operation scenarios.

Acknowledgement This work is supported in part by the National Natural Science Foundation of China (No.62372083, No.62072074, No.62076054, No.62027827, No.62002047), the Sichuan Science and Technology Program (No.2022NSFSC0552, No.2022YFQ0045, No.2022YFS0220, No.2023YFS0020, No. 2023YFS0197, No. 2023YFG0148), the CCF-Baidu Open Fund (No.202312), the Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China (No.ZYGX2021YGLH212, No.ZYGX2022YGRH012).

References

1. Zhang L, Xiong N, Pan X, et al. Improved object detection method utilizing yolov7-tiny for unmanned aerial vehicle photographic imagery[J]. Algorithms, 2023, 16(11): 520.
2. Pouyan S, Charmi M, Azarpeyvand A, et al. Propounding first artificial intelligence approach for predicting robbery behavior potential in an indoor security camera[J]. IEEE Access, 2023.

3. Liu G, Hu Y, Chen Z, et al. Lightweight object detection algorithm for robots with improved YOLOv5[J]. Engineering Applications of Artificial Intelligence, 2023, 123: 106217.
4. Guo, Meng-Hao, et al. "Segnext: Rethinking convolutional attention design for semantic segmentation." Advances in Neural Information Processing Systems 35 (2022): 1140-1156.
5. Redmon J, Farhada A. YOLOv3: an incremental improve-ment[J]. arXiv:1804.02767, 2018.
6. Adarsh P, Rathi P, Kumar M. YOLO v3-Tiny: Object Detection and Recognition using one stage improved model[C]//2020 6th international conference on advanced computing and communication systems (ICACCS). IEEE, 2020: 687-694.
7. Cheng R, He X, Zheng Z, et al. Multi-scale safety hel-met detection based on SAS-YOLOv3-tiny[J]. Applied Sci-ences, 2021, 11(8): 3652.
8. Ma Y, Fang Y. Safety helmet wearing recognition based on YOLOv5[M]//Mobile wireless middleware, operating systems and applications.Cham: Springer, 2022: 137-150.
9. Zhu X C, Chen Z T. Safety Helmet wearing detection based on improved YOLO v5[J].Journal of Nanjing Insti-tute of Technology(Natural Science Edition), 2021，19(4): 7-11.
10. Chen S B, Tang W H, Yang Y O, et al. Detection of safety helmet wearing based on improved faster R-CNN[C]// 2020 International Joint Conference on Neural Networks (IJCNN), 2020: 7-15.
11. Rabbi J, Ray N, Schubert M, et al. Small-object detection in remote sensing images with end-to-end edge-enhanced GAN and object detector network[J]. Remote Sensing, 2020, 12(9): 1432.
12. Zhu X, Lyu S, Wang X, Zhao Q. In TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
13. Zhang Z, Lu X, Cao G, Yang Y, Jiao L, Liu F. ViT-YOLO: Transformer-Basd YOLO for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2799–2808.
14. Tian T, Yang J. Remote sensing image target detection based on multi-scale feature fusion network. Laser Optoelectron. Prog. 2022, 59, 427–435.
15. Ding J, Xue N, Long Y, Xia G.-S, Lu Q. Learning roi transformer for oriented object detection in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2849–2858.