ITC 2/53

Information Technology and Control
Vol. 53 / No. 2 / 2024
pp. 453-469
DOI 10.5755/j01.itc.53.2.35273

Helmet Detection Based on Context Enhancement Pyramid Under Surveillance Images

Received 2023/10/05 | Accepted after revision 2024/01/06

HOW TO CITE: Xu, Z., Li, Y. (2024). Helmet Detection Based on Context Enhancement Pyramid Under Surveillance Images. *Information Technology and Control*, 53(2), 453-469. https://doi.org/10.5755/j01.itc.53.2.35273

Helmet Detection Based on Context Enhancement Pyramid Under Surveillance Images

Zhigang Xu

School of Computer and Communication, Lanzhou University of Technology,
Lanzhou, 730050, China; e-mail: xzg_cn@163.com

Yugen Li

School of Computer and Communication, Lanzhou University of Technology,
Lanzhou, 730050, China; e-mail: rootlyg@163.com
School of Information Engineering, Yangling Vocational & Technical College, Yangling, 712100, China

Corresponding author: xzg_cn@163.com

Helmet detection is of great significance for realizing the automated management of industrial safety. To address the problem that existing object detection methods have insufficient ability to detect helmet small objects under surveillance images, this paper proposes a helmet detection based on context enhancement pyramid under surveillance images to realize the automatic detection task of helmet objects. The method helps the network improve position localization for small-scale helmet objects by adding a high-resolution detection layer to YOLOv5. Also, the proposed context enhancement pyramid reduces the semantic differences between different scale features and generates rich contextual features to enhance the network's discriminative learning ability for helmet small object features. In addition, the proposed multi-scale attention module improves the feature fusion effect in the pyramid network to further capture multi-scale features and expand the receptive field to enhance the network's detection precision of helmet objects under surveillance images. The experimental analysis shows that the proposed method has good detection effect compared to existing object detection methods on the Safety Helmet Wearing Dataset (SHWD) as well as the customized dataset.

KEYWORDS: Surveillance images, Helmet detection, YOLOv5, Context enhancement pyramid, Multi-scale attention.

1. Introduction

Due to the rapid development of convolutional neural networks, deep learning based methods have not only achieved remarkable results in the field of computer vision [15, 18], but also been rapidly developed and widely applied in the fields of atmospheric monitoring [6, 44], wireless transmission [45-47] and health assistance [16, 19, 26, 32]. And in the field of industrial safety monitoring, the use of deep learning based object detection methods to realize the automated detection of helmets has become one of the urgent problems to be solved in the current industrial safety management.

As we know, in the process of industrial production, the helmet can effectively avoid workers' head injury [39], thus protecting the safety of workers' lives [30]. However, the current situation of workers wearing helmets still relies on manual methods for on-site supervision [5]. This is very inefficient for complex construction sites and a large number of workers, and can easily lead to safety accidents. With the development of object detection technology based on deep learning, some scholars have begun to utilize advanced object detection methods and make various improvements based on them to realize the automatic detection task of helmet objects [8-9, 12, 20]. However, for specific scenes such as surveillance images, the helmet object has a low imaging resolution and is susceptible to interference from the complex background environment as well as the influence of easy changes in the color of the target, which leads to the lack of discriminative learning ability of the existing deep learning based object detection methods on the features of small-scale helmet objects, which affects the detection model to differentiate between the small-scale helmet objects in the complex background environment. This affects the ability of the detection model to distinguish small-scale helmet objects in complex background environments, which in turn leads to the insufficient detection ability of the existing detection methods for small helmet objects.

In order to be able to improve the detection precision of small helmet objects under surveillance images, this paper proposes a context enhancement pyramid based helmet detection method under surveillance images. The method utilizes YOLOv5 [34] as a baseline and adds a high-resolution detection layer on top of it to help the detection network improve the positional spatial localization of helmet objects. Meanwhile, the proposed context enhancement pyramid structure is utilized to enhance the network's discriminative learning ability for helmet object features. In addition, the multi-scale attention module is used to further refine and capture the multi-scale context to help the detection network to predict the helmet object. Results of the experiments show that the method in this paper has good detection performance compared to mainstream object detection methods on both publicly available helmet datasets and custom datasets. The main contributions of this paper are as follows:

1 Add a high-resolution detection layer to the YOLOv5 network to reduce the feature loss of helmet small object targets during downsampling.

2 Propose a context enhancement pyramid to interactively fuse image features from shallow and deep layers to generate context-rich features, reduce semantic differences existing between features at different scales, and improve the discriminative learning ability of the network.

3 Propose a multi-scale attention module to further capture multi-scale features as well as expand the receptive field to improve the network's detection precision for small helmet objects.

The remainder of this paper is organized as follows. Section 2 presents the work related to helmet detection. Section 3 describes the baseline YOLOv5 method and the proposed method in this paper. Section 4 gives the experimental results and analysis. Section 5 concludes the paper.

2. Related Work

In recent years, thanks to the rapid development of deep learning in various fields [48], computer vision technology has been widely used. At present, deep learning based object detection technology is widely used in the fields of intelligent manufacturing [25] and industrial automation [17, 35, 49] due to its excellent detection performance. In the field of industrial production safety monitoring, with the increasing popularity of the current video surveillance system, the use of object detection technology to realize the automatic detection of workers' helmet wearing [24] has become more and more important. In the early days, most of the helmet object detection utilized manual feature extraction to detect the helmet object [41]. However, these traditional methods are more complicated and inefficient for helmet object detection. In recent years, many scholars have begun to use advanced object detection methods to realize the automatic detection task for helmets.

Wang [38] et al. propose a real-time helmet wear detection method by introducing a cross-stage localized network CSP [37] and a spatial pyramid pooling structure in the YOLOv3 [29] backbone to improve the learning ability of the network, and combining top-down and bottom-up feature fusion strategies to improve the detection of helmet objects at construction sites. Yang [42] et al. enhanced the multi-scale feature extraction capability of the YOLOv4 [1] backbone network and introduced a channel attention mechanism to dynamically focus on the channel features of helmet objects, thus improving the network's detection performance for small helmet objects. Fang [7] et al. added an attention mechanism to the backbone of YOLOv5 to make the network pay more attention to the region of interest, and at the same time combined the BiFPN [31] network structure in YOLOv5 to better fuse the fine-grained features of the helmet object and to improve the detection accuracy of the helmet object under the condition of combining with migration learning. Chen [3] et al. reduced the computational complexity of the model by introducing a lightweight Ghost module into the backbone and neck feature extraction portions of the YOLOv5-S network, and combined it with a BiFPN structure to reconfigure the network for the helmet detection task. Gao [10] et al. propose a real-time helmet detection method based on the YOLOX [11] method, which strengthens the feature extraction capability of the network by adopting the recursive gated volume and BiFPN structure, and at the same time, adopts the training strategy of the SIOU cross-entropy loss function to further improve the detection precision of helmets. Lin [22] et al. added CBAM and super-resolution modules to YOLOX to extract foreground features and optimize image features as much as possible, and added a detection head for small objects of helmets to further improve the detection accuracy of helmets. Yu [43] et al. propose an improved helmet detection model based on YOLOv4, which significantly reduces the computational effort of the model by adding a depth-separable convolution to the YOLOv4 network to replace the traditional 3×3 convolution, and at the same time combines with multiscale prediction to realize real-time detection of helmets. Although the above methods can improve the detection precision of helmets to a certain extent and effectively increase the detection speed, they still do not propose effective solutions for the automated detection of small objects of helmets in surveillance image scenes. Therefore, the current mainstream target detection methods are still challenging for detecting small objects of helmets in surveillance image scenes.

3. Method

3.1. Overview of the YOLOv5

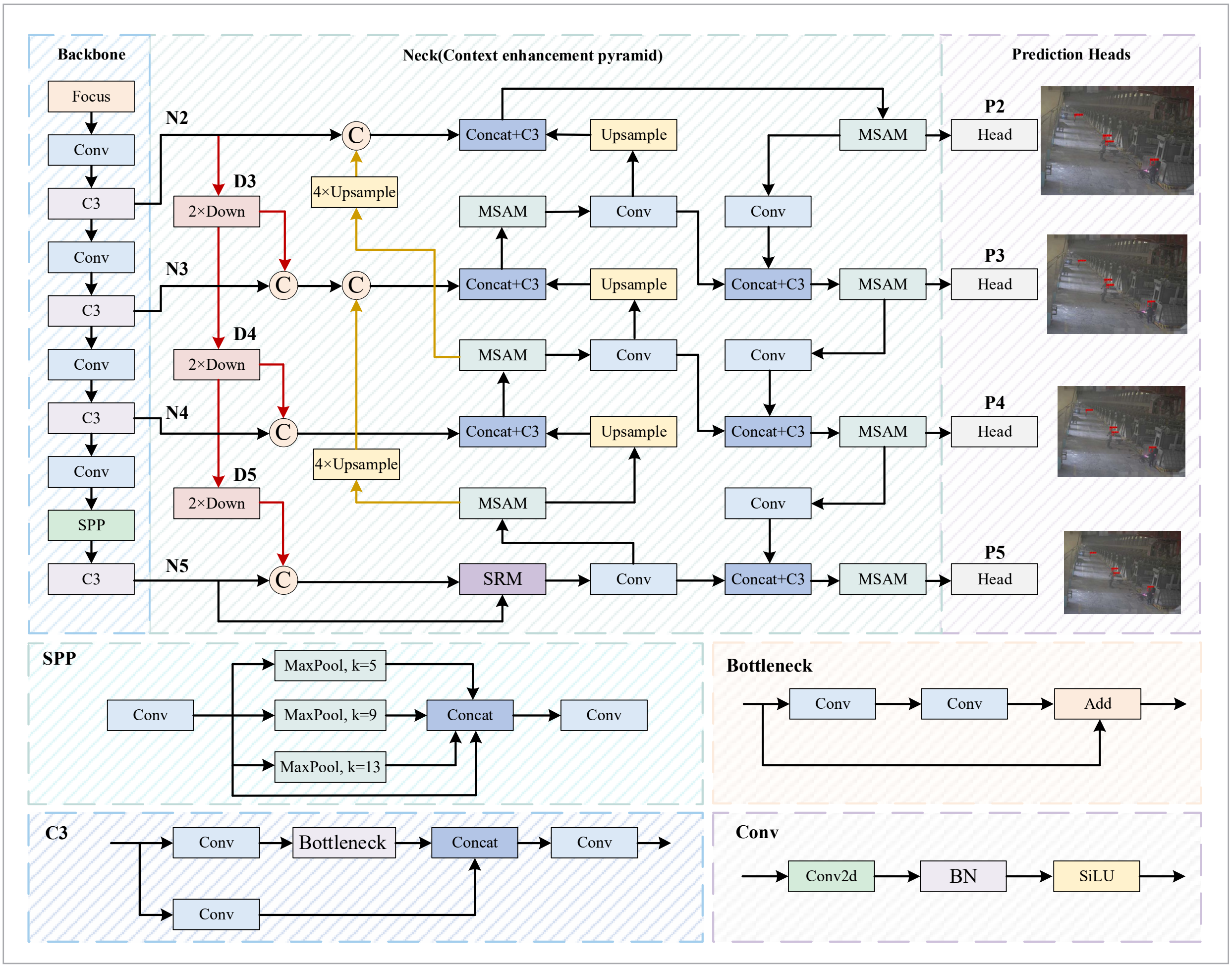
According to the depth and width of the network, the YOLOv5 method can be divided into four different versions: YOLOv5-S, YOLOv5-M, YOLOv5-L, and YOLOv5-X. The YOLOv5 method continues the network structure of the YOLOv4 method, which consists of a backbone network, a neck structure, and a prediction network, respectively. Foucs, CSP structure, and SPP [13] modules are used in the backbone network for image feature extraction. The CSP structure enhances the learning expression of the backbone using cross-stage connectivity. Meanwhile, the SPP module is able to perform multi-scale feature mapping so as to fuse feature information from different scales. The neck structure adopts the PANet [23] structural idea to build a pyramid network by top-down and bottom-up paths, which further enhances the network's ability to extract image features. The prediction network is able to detect objects of different sizes with three different scales of features extracted from the neck, and calculate the object class, confidence score, and predicted object frame information.

3.2. Network Framework of the Proposed Method

The network structure of the proposed context enhancement pyramid based helme detection method under surveillance images is shown in Figure 1. The method extends the network structure of the YOLOv5 method by the proposed context enhancement pyramid and multi-scale attention module. Also, in order to

Figure 1

Network framework of the proposed method



be able to accurately localize the position of small-scale helmet objects under surveillance images, this paper adds a high-resolution detection layer to the YOLOv5 network. Among them, the context enhancement pyramid can reduce the semantic differences between different scales and generate rich contextual features to improve the network's discriminative learning ability for helmet objects. The multi-scale attention module can further refine the features and establish multi-scale mapping and expand the receptive field to improve the network's detection precision for helmets.

3.3. High-resolution Detection Layer

Since the backbone network of YOLOv5 needs to perform multiple down-sampling processes when feature extraction of helmet objects is carried out, which is likely to lead to the gradual loss of spatial location information of the helmet objects in the surveillance image, thus reducing the accurate localization of the detection network for helmet small objects. Therefore, this paper adds a high-resolution detection layer P2 to the original 3-level detection layer of YOLOv5, as shown in Figure 1, in order to improve the accurate localization ability of the detection network for helmet objects. Specifically, the shallow image features N2 extracted from the backbone are extended and the proposed context enhancement pyramid network is utilized to obtain the high-resolution detection layer P2, which enables the detection layer P2 to retain richer spatial detail information of the helmet objects as much as possible, making the network more sensitive in dealing with small-scale helmet objects.

3.4. Context Enhancement Pyramid

The YOLOv5 method adopts the idea of PANet (Figure 2(b)) network structure to construct the pyramid network, although it can improve the feature information transfer of small objects in the FPN [21] (Figure 2(a)) structure in the network to a certain extent, the semantic differences that exist between features of different scales are still ignored. The deeper features in the PANet structure need to undergo many times of up-sampling in order to be fused with the shallowest features, and a large amount of abundant abstract semantics is gradually diluted, which in turn causes the lack of semantic information of shallow features. Meanwhile, the deep features lack sufficient contextual information around the object, resulting in the inability to precisely localize small-scale objects, which leads to the insufficient discriminative learning ability of the detection network for the helmet objects.

Therefore, inspired by the literature [5, 14], we propose the context enhancement pyramid in this paper, as shown in Figure 2©. This structure is able to transfer shallow image features to deeper feature layers, and transfer the rich semantics contained in deeper image features to shallower feature layers, so as to generate rich contextual semantics, reduce the semantic differences between features of different scales in the pyramid network, and then guide the feature construction process of the pyramid network, in order to improve the network's discriminative learning ability for the helmet objects. Among them, the semantic refinement module is used to eliminate the redundant contexts in the deep image features and refine the helmet object feature information.

1 **Context enhancement pyramid:** The context enhancement pyramid is able to inject shallow image features containing a large number of helmet object spatial details directly into the deeper feature layer, so that smaller scale helmet object features will not be easily lost. Moreover, deep image features containing rich semantics are up-sampled across stages and transferred to shallow image features to make up for the lack of semantic information of shallow image features and generate rich contextual features to reduce the semantic differences of features of different scales and improve the network's discriminative learning ability of the helmet objects. Specifically, in order to interactively fuse the shallow features with the deeper ones, this paper adds a bottom-up extended path and two top-down extended cross-stage paths to the PANet network structure, so that the deeper feature layer and the shallower one can effectively obtain the required spatial information of the target location as well as sufficient abstract semantics. The fusion approach uses Concat to preserve as many contextual features as possible. It can be described as:

$$Ni' = Concat\left(Di,\ \ Ni\right), \left(i = 3,4,5\right) \tag{1}$$

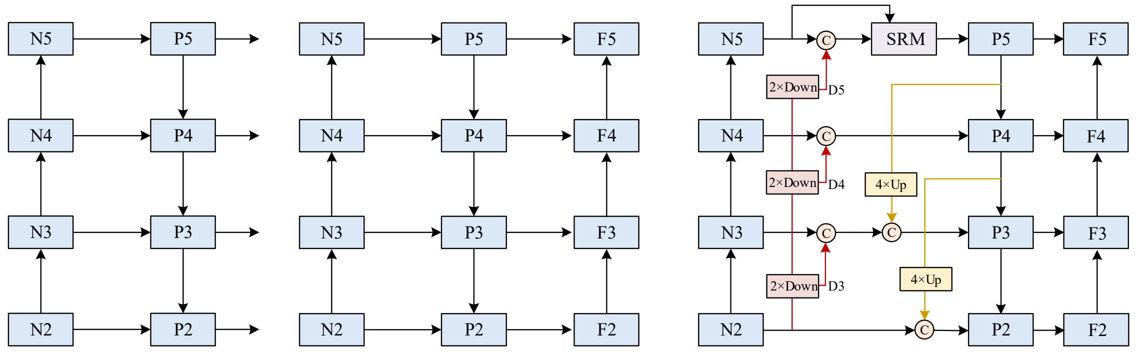
$$N3'' = Concat\left(Upsample_{\times 4}\left(P5\right), N3'\right) \tag{2}$$

$$N2' = Concat\left(Upsample_{\times 4}\left(P4\right), N2\right) \tag{3}$$

Where $Upsample_{\times 4}$ indicates a 4x upsampling operation using nearest neighbor interpolation.

Figure 2

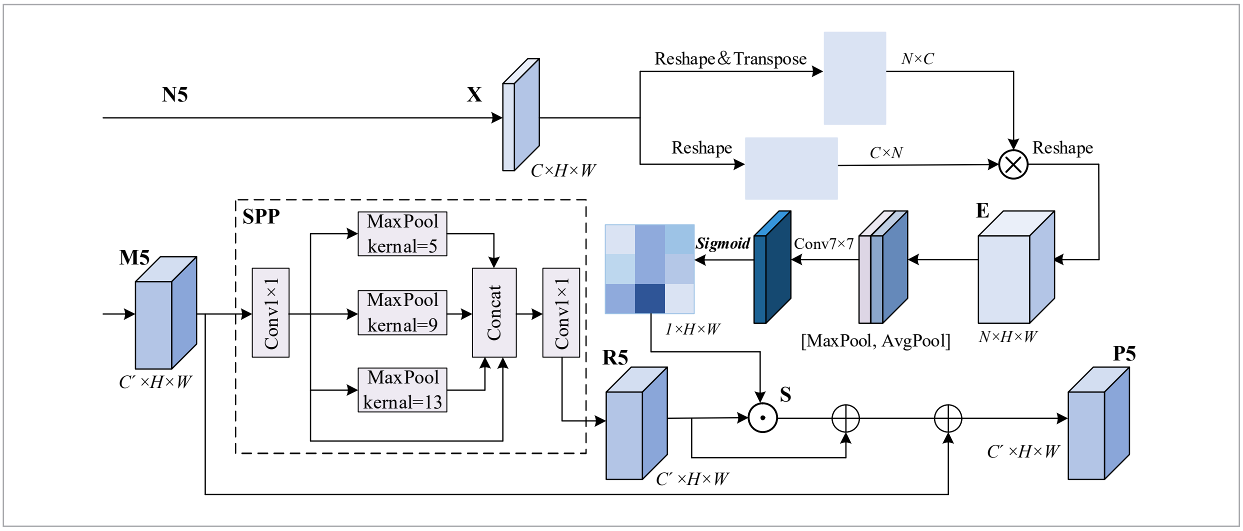
Different feature pyramid structures



(a) FPN (b) PANet (c) Context enhancement pyramid

Figure 3

Semantic refinement module(SRM)



2 **Semantic refinement module (SRM):** In order to avoid the original rich semantics in the deep image features being interfered by the redundant background information from the shallow features, thus causing feature confusion and resulting in a reduction in the network's ability to learn the semantic features of the helmet target. Therefore, this paper proposes a semantic refinement module based on the literature [2], as shown in Figure 3. This module is able to establish long-range dependencies on the original deep image features and generate spatial attention weights to refine the helmet object information in the image features, thus enabling the network to improve the discriminative learning ability for the helmet object. Among them, the SPP module is used to fuse multi-scale spatial features to mine richer small-scale helmet object feature information.

Specifically, for the fused feature M5 first, the SPP module is utilized for multi-scale mapping to obtain the feature R5. Then, the deep features N5 extracted by the backbone are regarded as $X \in \mathbb{R}^{C\times H\times W}$, and they are convolutionally transformed using 1×1 convolutional layers W_q and W_k to obtain $Q = W_q X$ and $K = W_k X$, respectively. Then Q and K are multiplied by matrix to obtain the similarity matrix $E \in \mathbb{R}^{N\times N}$ and $N = H \times W$, which is expressed by the formula:

$$E = Q(X)^{T} \times K(X) \qquad (4)$$

Then, $E \in \square^{N\times N}$ is adjusted to $E \in \square^{N\times H\times W}$, and then $E^{max} \in \square^{1\times H\times C}$ and $E^{avg} \in \square^{1\times H\times C}$ are obtained using the max pooling and average pooling process, followed by fusion using 7×7 convolution and using Sigmoid to obtain attention weights and weighting to feature R5. Then, they are output after element summing with feature R5 and feature M5, respectively, expressed by the formula:

$$S = R5 \odot \sigma\left(f^{7\times 7}\left(Concat\left(E^{max}; E^{avg}\right)\right)\right) \qquad (5)$$

$$P5 = (S \oplus R5) \oplus M5 \qquad (6)$$

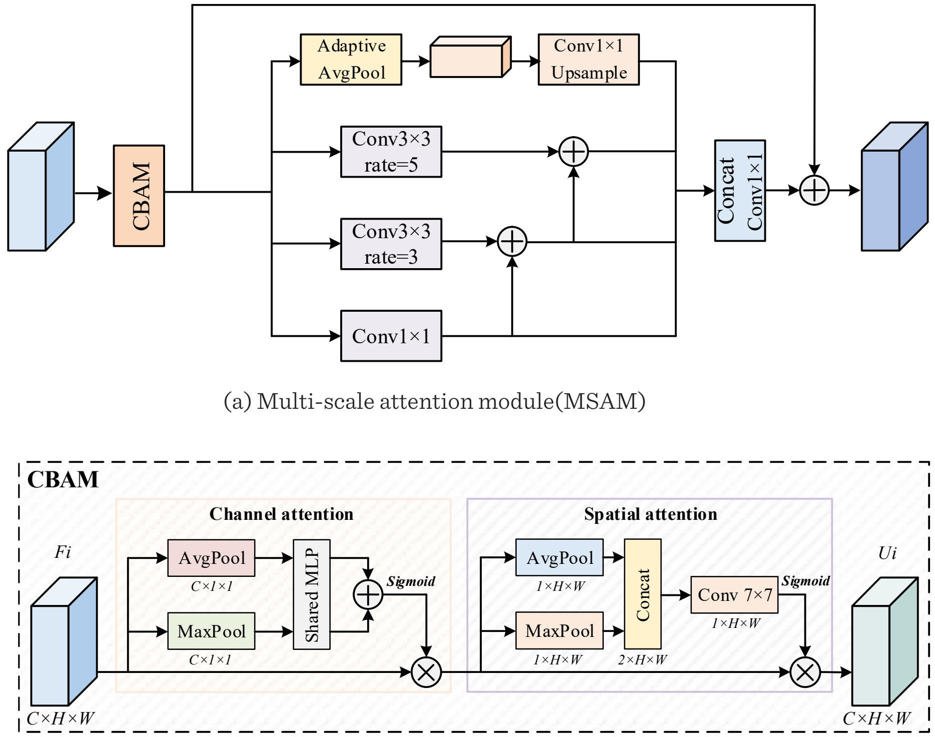
Where σ is a Sigmoid function, while $f^{7\times 7}$ denotes a convolutional operation with a convolutional kernel size of 7 × 7.

3.5. Multi-scale Attention Module (MSAM)

In order to improve the effect of feature fusion in the pyramid network and avoid the region where the helmet object is located in the image features from being affected by redundant information, we proposed the multi-scale attention module, as shown in Figure 4(a). This module is able to refine the fused features during the construction of the pyramid network, while capturing the multi-scale information inside

Figure 4

(a)Multi-scale attention module(MSAM).(b)Convolutional block attention module(CBAM)



(a) Multi-scale attention module(MSAM)

(b) Convolutional block attention module(CBAM)

the image features of different scales and further expanding the receptive field as a way to improve the feature fusion effect in the pyramid network and increase the network's detection precision for helmet objects.

Specifically, the multi-scale attention module first refines the features from both the channel and spatial dimensions using CBAM (Figure 4(b)) [40] from channel attention and spatial attention, respectively, so as to filter out a large amount of noise and redundant information contained in the image features and to highlight the key helmet object features, which can be described as:

$$L_i = \sigma\left(MLP\left(F_i^{max}\right) + MLP\left(F_i^{avg}\right)\right) \otimes F_i \qquad (7)$$

$$U_i = \sigma\left(f^{7\times7}\left(Concat\left(L_i^{max}, L_i^{avg}\right)\right)\right) \otimes L_i \qquad (8)$$

Where, σ is the Sigmoid function, MLP is the multi-layer perceptron, and $f^{7\times7}$ is the 7×7 convolution operation. $F_i^{max} \in \mathbb{R}^{C\times1\times1}$ and $F_i^{avg} \in \mathbb{R}^{C\times1\times1}$ denote the maximum pooling and average pooling process from the channel dimension. And $L_i^{max} \in \mathbb{R}^{1\times H\times W}$ and $L_i^{avg} \in \mathbb{R}^{1\times H\times W}$ denote the maximum pooling and average pooling process from the spatial dimension.

Then, for the image features that have been refined by CBAM, adaptive global average pooling is then used to aggregate the global semantic information, while the input features are mapped using 1×1 convolution as well as 3×3 convolution with different dilation rates, and then the output features are subjected to the elemental summation operation in order to avoid the loss of helmet object feature information. Finally, the output features from the four branches are fused using Concat and 1×1 convolution. One long residual edge is used to maintain the integrity of the information inside the image features.

4. Experiments

4.1. Experimental Datasets

4.1.1. Safety Helmet Wearing Dataset (SHWD)

The publicly available Safety Helmet Wearing Dataset (SHWD) [27] has 7581 images containing a total of two categories of labels, hat and person. person labels are derived from the SCUT-HEAD [28] dataset to simulate the unworn helmet objects. Among them, there are 9047 hat tags and 111514 person tags. In this paper, we divide the dataset into training and test sets according to the ratio of 8:2, and 10% of the training set is classified as the validation set for validation. There are 5457 images in the training set, 607 images in the validation set, and 1517 images in the test set. Where, the training set, validation set as well as the test set contains the number of hat labels and person labels as shown in the Table 2.

Table 1

Comparison of the number of images in the training set, validation set, and test set in SHWD and Custom dataset

Datasets	Number of images			
	train	val	test	Total
SHWD	5457	607	1517	7581
Custom dataset	7617	847	2117	10581

4.1.2. Custom Dataset

To further evaluate the effectiveness of the proposed method in detecting helmet objects in surveillance image scenarios, this paper expands 3000 surveillance images on the basis of SHWD. The expanded surveillance images are all from real industrial production environments and manually labeled according to the Pascal VOC data format, as shown in Figure 5. The expanded dataset has 10581 images containing 131586 labeled frames. Among them, there are 20072 hat labels and 111,514 person labels. We divide the dataset into training set and test set according to the ratio of 8:2, while 10% of the training validation set is randomly divided into validation set. Table 1 demonstrates the comparison of the number of images in the custom dataset and the SHWD dataset, the training set in the final divided custom dataset has a total of 7,617 have images, the validation set has 847 images, and the test set has a total of 2,117 images. Among them, the number of hat labels and person labels included in the training set, validation set, and test set are shown in Table 2.

In addition, since the expanded surveillance images are all from real construction environments, this paper utilizes mosaics in the detection effect images shown to obscure textual information such as the construction location and time displayed in the upper left corner of the surveillance images.

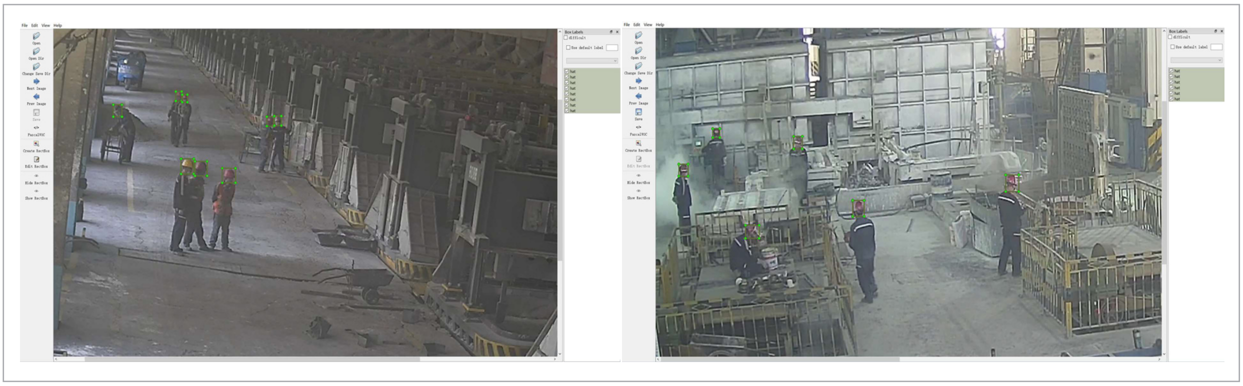
Table 2

Comparison of the number of labels included in the training set, validation set, and test set in SHWD and Custom dataset

Labels	SHWD			Custom dataset		
	train	val	test	train	val	test
hat	6379	596	2072	14449	1800	3823
person	78973	9257	23284	81475	8670	21369

Figure 5

Example of labeling



4.2. Training Setup

In this paper, the experiments are conducted in the Pytorch deep learning framework, and the GPU used is an Nvidia RTX 3090.The proposed method is implemented on top of the original YOLOv5 network, and thus follows the original YOLOv5 tuning strategy for the depth and width of the network. Stochastic Gradient Descent (SGD) is used to optimize the weights of the network during the training process, and a strategy of cosine annealing learning rate decay and Mosaic data augmentation is used, while the backbone CSPDarkNet, which has been pre-trained on ImageNet, is loaded during the training. in addition, the initial learning rate is set to 1e-2, and the minimum learning rate is set to 1e-4. furthermore, the batch size is set to 4, epochs is set to 100, momentum is set to 0.937, and weight_decay is set to 5e-4. For a fair comparison, the same training settings are used for each method.

4.3. Evaluation Metrics

This paper uses average precision (AP), mean average precision (mAP) and Frames Per Second (FPS) as evaluation metrics. Where AP is denoted as the area under the curve after the multiplication of precision(P) and recall(R), and mAP is denoted as the average of the AP of the two object categories. The formulas for precision(P) and recall(R) are denoted as:

$$\begin{cases} P = \dfrac{TP}{TP+FP}; \\ R = \dfrac{TP}{TP+FN}; \end{cases} \tag{9}$$

where, TP denotes the number of samples that were detected as correct and were actually positive samples. FP denotes the number of samples that were detected as correct but were actually negative samples. FN denotes the number of samples that were categorized as negative samples but were actually positive samples. mAP is calculated using an IoU threshold of 0.5.

4.4. Experiments on SHWD

To verify the detection effect of the proposed method for the helmet objects, the proposed method is compared with different methods on the SHWD in this paper, and the experimental results are shown in Table 3.

Table 3

Different methods for comparison of results on SHWD

Method	Input size	AP(hat)/%	AP(person)/%	mAP/%
YOLOv3	416×416	83.42	80.42	81.92
YOLOv3	608×608	87.76	90.74	89.25
YOLOv4	416×416	86.82	79.45	83.14
YOLOv4	608×608	91.08	90.30	90.69
FCOS	640×640	86.29	84.54	85.41
CenterNet	512×512	83.14	75.21	79.17
YOLOv5-S	640×640	84.56	85.56	85.06
YOLOv5-M	640×640	87.91	88.61	88.26
YOLOv5-L	640×640	90.41	90.63	90.52
YOLOv5-X	640×640	91.65	92.19	91.92
YOLOX-S	640×640	83.99	87.18	85.58
YOLOX-M	640×640	85.16	88.82	86.99
YOLOX-L	640×640	87.04	90.09	88.56
YOLOX-X	640×640	88.23	92.76	90.50
YOLOv7	640×640	91.78	92.64	92.21
Ours-S	640×640	85.07	87.50	86.28
Ours-M	640×640	89.55	90.72	90.14
Ours-L	640×640	91.48	92.33	91.90
Ours-X	640×640	93.50	93.25	93.37

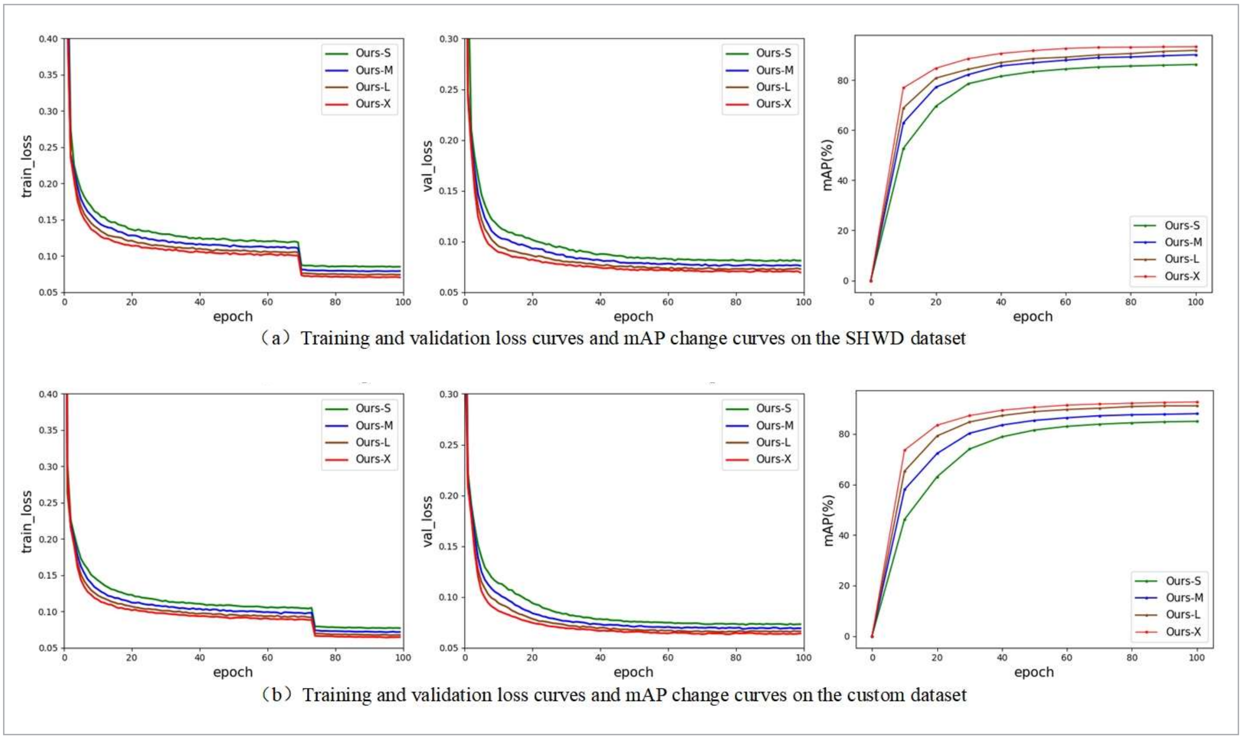
From Table 3, the proposed method shows good detection results on SHWD. The proposed Ours-S method improves the mAP of YOLOv3 and YOLOv4 by 4.36% and 3.14%, respectively, compared to the YOLOv3 and YOLOv4 with an input size of 416 × 416, and the mAP value of Ours-S can be improved by 7.11% compared to the CenterNet [50] method with an input size of 512 × 512. Also, the mAP of Ours-S can be improved by 1.22% and 0.7% compared to the baseline method YOLOv5-S and the advanced YOLOX-S method, respectively. For the Ours-M, Ours-L, and Ours-X methods, the mAP can be improved by 1.88%, 1.38%, and 1.45% compared to the baseline methods YOLOv5-M, YOLOv5-L, and YOLOv5-X, respectively. In addition, Ours-X was able to increase the mAP by 4.12% and 2.68% compared to the YOLOv3 and

YOLOv4 methods with input size of 608 × 608, respectively, and by 7.96% compared to the FCOS [33] method, and Ours-X was able to increase the mAP by 2.87% and 1.87% compared to the advanced YOLOX-X and YOLOv7 [36] methods, respectively.

Figure 6(a) shows the training and validation loss curves and the mAP variation curves of the proposed method on SHWD, respectively, and it can be seen that with the adoption of a more powerful baseline network, it can make the training and validation loss converge faster and make the mAP value increase continuously. Figure 7 shows the detection results of Ours-X method on the test set classified in this paper.

Figure 6

Training and validation loss curves and mAP change curves of the proposed method on SHWD and custom dataset

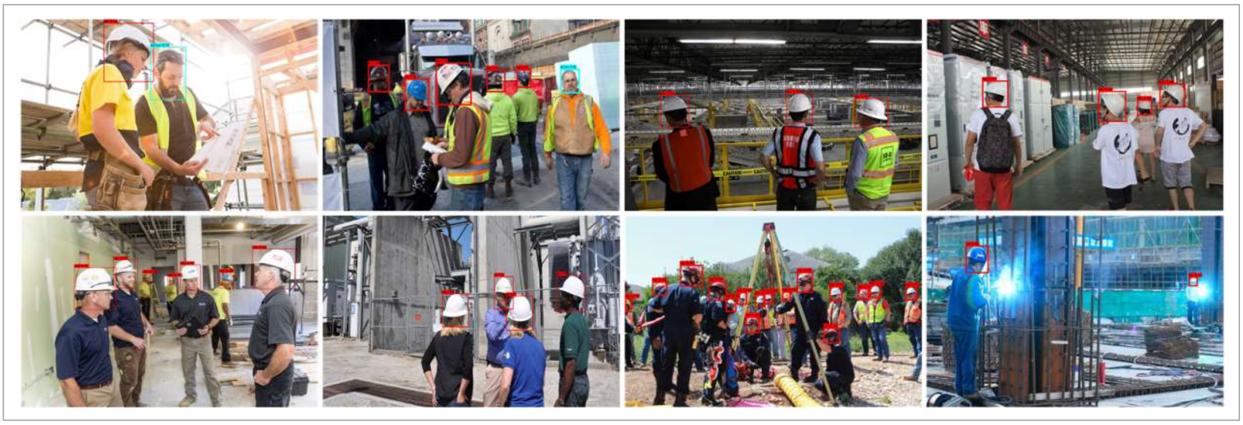


（a）Training and validation loss curves and mAP change curves on the SHWD dataset

（b）Training and validation loss curves and mAP change curves on the custom dataset

Figure 7

The detection effect of Ours-X in the SHWD test set



4.5. Experiments on Custom Dataset

To further verify the proposed method's detection effect for small-scale helmet objects under surveillance images, the proposed method is compared with different detection methods on the custom dataset in this paper, and the experimental results are shown in Table 4.

From Table 4, it can be seen that the proposed method achieves the best detection performance in terms of mAP values on Custom dataset as well as for helmet objects AP values. It can be seen that the proposed methods Ours-S, Ours-M, Ours-L, and Ours-X are able to improve the mAP values by 3.6%, 2.87%, 2.64%, and 2.48% compared to the baseline methods YOLOv5-S, YOLOv5-M, YOLOv5-L, and YOLOv5-X, respectively, and for the helmet objects the AP values can be improved by 5.08%, 4.12%, 3.82%and 3.55%, but the detection speed is slightly reduced with FPS values of 37.3, 31.2, 25.2, and 20.3 respectively. The proposed method Ours-X was able to improve the mAP by 6.7% and 5.42% compared to the YOLOv3 and YOLOv4 methods with input size of 608 × 608, and it was able to improve the AP for the helmet objects by 9.08% and 7.93%, respectively. Also, the proposed method Ours-X has a detection speed FPS value of 20.3, which is slightly reduced compared to the YOLOv3 and YOLOv4 methods that exhibit FPS values of 65.9 and 43.7. Ours-X was able to improve the mAP values by 3% and 2.43% compared to the state-of-the-art YOLOX-X and YOLOv7 methods, and was able to improve the AP values by 3.84% and 3.75% for the helmet objects, respectively. In addition, although the proposed method Ours-X exhibits slightly lower FPS values of 34.8 and 61.4 compared to the advanced YOLOX-X and YOLOv7 methods, the proposed method is able to significantly improve the detection accuracy for the helmet objects.

Figure 6(b) further demonstrates the training and validation loss curves as well as the mAP variation curves of the proposed method on Custom dataset.

Table 4

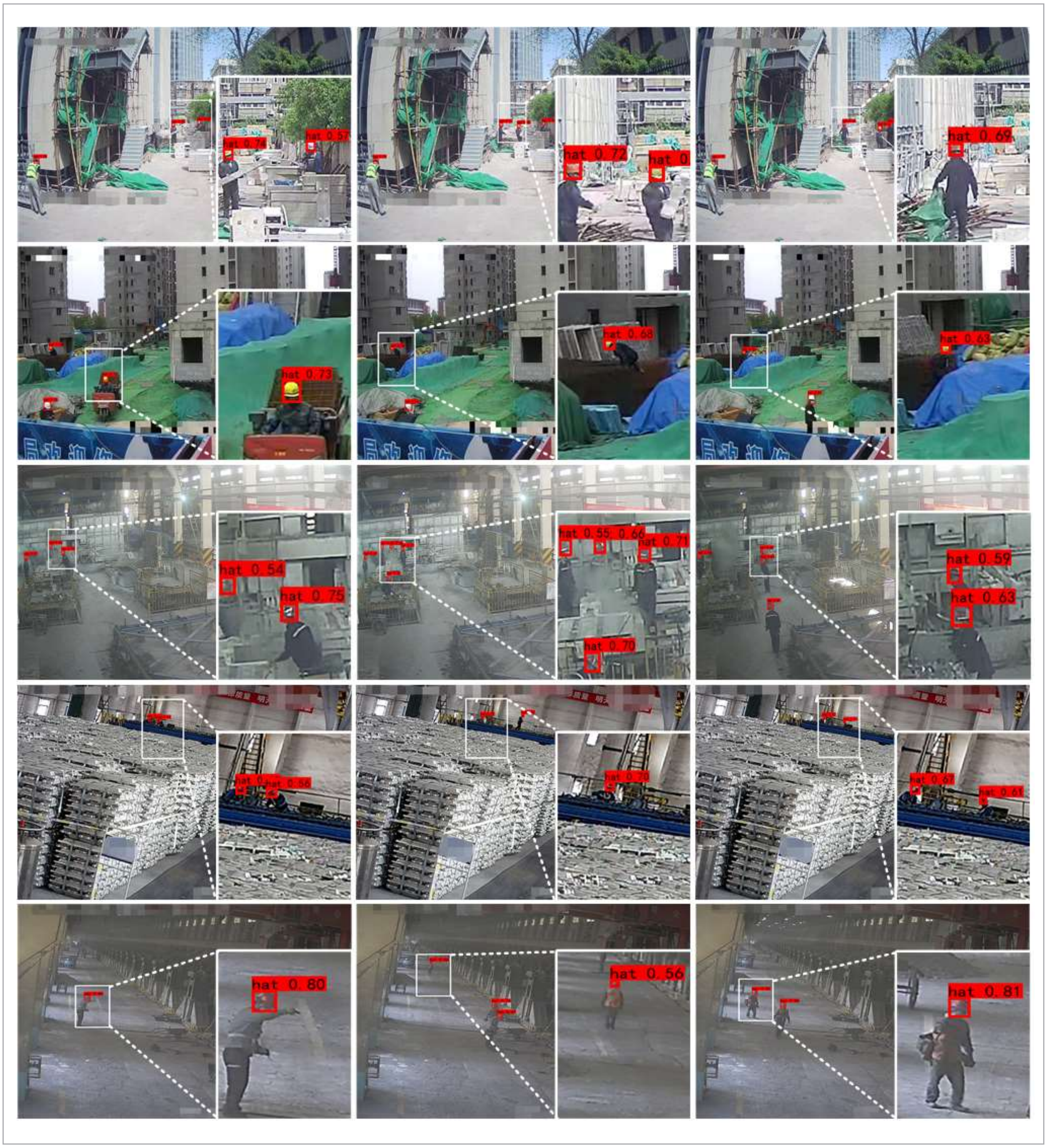
Different methods for comparing results on custom dataset

Method	Input size	AP(hat)/%	AP(person)/%	mAP/%	FPS
YOLOv3	608×608	81.98	89.92	85.95	65.9
YOLOv4	608×608	83.13	91.33	87.23	43.7
FCOS	640×640	82.36	88.13	85.25	36.1
CenterNet	512×512	76.55	79.32	77.94	54.5
YOLOv5-S	640×640	75.15	87.63	81.39	75.0
YOLOv5-M	640×640	80.33	89.98	85.16	55.1
YOLOv5-L	640×640	85.06	91.92	88.49	45.3
YOLOv5-X	640×640	87.51	92.84	90.17	37.9
YOLOX-S	640×640	77.82	86.95	82.39	52.9
YOLOX-M	640×640	81.86	88.05	84.96	44.4
YOLOX-L	640×640	86.64	91.67	89.15	37.5
YOLOX-X	640×640	87.22	92.08	89.65	34.8
YOLOv7	640×640	87.31	93.13	90.22	61.4
Ours-S	640×640	80.23	89.75	84.99	37.3
Ours-M	640×640	84.45	91.62	88.03	31.2
Ours-L	640×640	88.88	93.39	91.13	25.2
Ours-X	640×640	91.06	94.24	92.65	20.3

Figure 8 demonstrates some of the detection results of Ours-X on the test set, and it can be seen that the proposed method is able to better detect the distant small-scale helmet objects in the surveillance images in different construction operation environments.

Figure 8

The detection effect of Ours-X in the test set with custom dataset



4.6. Ablation Analysis

1 **Impact of different modules:** In order to evaluate the effects of context enhancement pyramid (CEP), multi-scale attention module (MSAM), and the addition of high-resolution detection layer on the baseline YOLOv5 detection performance, we conduct ablation experiments on a customized dataset using YOLOv5-S as the baseline, and the results of the experiments are shown in Table 5. Where, P2 denotes for the added high resolution detection layer and SRM denotes semantic refinement module.

From Table 5, it can be seen that adding the high-resolution detection layer P2 to the baseline can make the APs of mAP and helmet improve by 1.94% and 2.29%, respectively, indicating that P2 can enhance the network's location localization of helmet objects. Adding CEP containing SRM to the baseline can improve the AP of mAP and helmet by 2.44% and 3.09% respectively, indicating that CEP can utilize the generated rich context to enhance the network's discriminative learning ability for helmet objects. However, when CEP without SRM is added, it only enhances the mAP and the AP of helmet by 1.86% and 2.76%, respectively, indicating that without the help of SRM, a large amount of redundant contextual information will semantically interfere with the deep image features and reduce the network's ability to learn about the helmet objects. When MSAM is added to the baseline, it can make the mAP and the AP value for helmet improve by 1.71% and 2.59%, respectively, indicating that MSAM can improve the feature fusion effect during the construction of pyramid network and increase the detection precision of the network for helmet objects. In addition, when the CEP containing SRM is added on top of P2, the mAP can be improved by 3.08%, indicating that the context enhancement pyramid combined with the high-resolution detection layer P2 can further improve the detection effect of helmet objects. And when MSAM is added to P2, it can make the mAP improve by 2.73%, which illustrates the contribution of MSAM to improve the pyramid network construction. Adding CEP and MSAM containing SRM to the baseline only improves the mAP and helmet AP by 1.94% and 2.49%, respectively, indicating that the network is not capable of localizing the position of small-scale helmet objects in the absence of a high-resolution detection layer. In comparison, the proposed method in this paper is able to improve the mAP and the AP for helmet by 3.6% and 5.08%, respectively, which further illustrates the effectiveness of the proposed method.

Table 5

The results of the ablation experiments

Method	AP(hat)/%	AP(person)/%	mAP/%	FPS
Baseline	75.15	87.63	81.39	75.0
Baseline+P2	77.44	89.22	83.33	58.4
Baseline+CEP(W/o SRM)	77.91	88.78	83.35	61.2
Baseline+CEP(W/ SRM)	78.24	89.41	83.83	58.3
Baseline+MSAM	77.74	88.45	83.10	51.3
Baseline+P2+CEP(W/o SRM)	78.98	89.35	84.16	57.5
Baseline+P2+CEP(W/ SRM)	79.55	89.40	84.47	55.8
Baseline+P2+MSAM	79.30	88.95	84.12	38.9
Baseline+CEP(W/o SRM)+MSAM	77.65	88.66	83.15	40.4
Baseline+CEP(W/ SRM)+MSAM	77.64	89.01	83.33	39.1
Baseline+P2+CEP(W/o SRM)+MSAM	79.50	89.57	84.53	38.5
Baseline+P2+CEP(W/ SRM)+MSAM	80.23	89.75	84.99	37.3

2 **Impact of pre-trained backbone networks:** In addition, in order to further evaluate the impact of loading the pre-trained backbone network during the training process of the proposed method on the experimental results, this paper analyzes the proposed method experimentally on a custom dataset, and the experiment results are shown in Table 6.

From Table 6, it can be seen that the proposed Ours-S, Ours-M, Ours-L, and Ours-X methods were able to improve the mAP values by 8.29%, 9.4%, 11.19%, and 10.2% under the condition of loading the pre-trained backbone network as compared to the training without loading the pre-trained backbone, respectively. Meanwhile, the AP values for the helmet target were able to increase by 12.84%, 14.65%, 17.03%, 15.84% respectively. Therefore, it can be inferred that loading the pre-trained backbone network for migration learning during the training process can significantly improve the detection performance of the proposed method for helmets compared to not loading the pre-trained backbone network.

Table 6

Comparison of experimental results on whether to load a pre-trained backbone on custom dataset

Method	pre-trained CSPDarket	AP(hat)/%	AP(person)/%	mAP/%
Ours-S	×	67.39	86.02	76.70
	√	80.23	89.75	84.99
Ours-M	×	69.80	87.46	78.63
	√	84.45	91.62	88.03
Ours-L	×	71.85	88.03	79.94
	√	88.88	93.39	91.13
Ours-X	×	75.22	89.68	82.45
	√	91.06	94.24	92.65

3 **Impact of different hyper-parameter settings:** In order to further evaluate the impact of hyper-parameters settings of the proposed method on the experimental results during the training process, this paper evaluates the adopted strategies of Monsaic data augmentation as well as cosine annealing learning rate, and the experimental analysis is carried out on a custom dataset, and the results are shown in Table 7.

From the Table 7, it can be seen that when the proposed method does not use the Monsaic data augmentation and Cosine annealing learning rate strategy, the mAP can only reach 79.35%, and the AP of the helmet can only reach 74.93%, which indicates that the network does not fully learn the complex feature information of the helmet objects in the dataset during the training process and does not converge to the optimal solution. When only Monsaic data augmentation is used for training, the mAP can reach 81.66%, which indicates that the Monsaic data augmentation strategy can help the network to improve the learning ability of the helmet objects. When only the Cosine annealing learning rate is used, the mAP reaches 83.28%, which indicates that the Cosine annealing learning rate can help the network to find the optimal solution during the training process and improve the training effect and detection performance of the network. In contrast, the proposed method in this paper uses both Monsaic data augmentation and Cosine annealing learning rate training strategy, which can make the mAP reach 84.99%, and further improve the network's detection performance for helmet objects.

Table 7

Comparison of experimental results with hyper-parameters settings on custom dataset

Method	Mosaic data augmentation	Cosine annealing scheduler	AP(hat)/%	AP(person)/%	mAP/%
Ours-S	×	×	74.93	83.76	79.35
	√	×	75.72	87.59	81.66
	×	√	78.72	87.85	83.28.
	√	√	80.23	89.75	84.99

5. Conclusion

This paper proposes a helmet detection based on context enhancement pyramid under surveillance images to realize the automatic detection task for helmets in industrial production processes. The method helps the network to accurately localize the helmet objects by adding a high-resolution detection layer to the YOLOv5 network. Meanwhile, the proposed context enhancement pyramid interactively fuses image features from both shallow and deep layers to enhance the network's discriminative learning ability for helmet object features. In addition, the proposed multiscale attention module is used to improve the feature fusion effect during the construction of the pyramid network, which further improves the detection precision of the network for helmet. Experimental results show that the method proposed in this paper has good detection performance for helmet objects in surveillance image scenarios compared with mainstream object detection methods. Future work will further consider automated helmet object detection tasks in more complex environments.

Acknowledgement

This work was supported by a grant from the National Nature Science Foundation of China (No. 62161020).

Data availability

The Safety Helmet Wearing Dataset (SHWD) from https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset. Custom dataset requires contacting the author.

Code availability

Code is available upon request by contacting the author.

References

1. Bochkovskiy, A., Wang, C. Y., Liao, H. Y. M. Yolov4: Optimal Speed and Accuracy of Object Detection. arXiv Preprint, 2020, arXiv:2004.10934.
2. Cao, J., Chen, Q., Guo, J., Shi, R. Attention-Guided Context Feature Pyramid Network for Object Detection. arXiv Preprint, 2020, arXiv:2005.11475.
3. Chen, W., Li, C., Guo, H. A Lightweight Face-Assisted Object Detection Model for Welding Helmet Use. Expert Systems with Applications, 2023, 221, 119764. https://doi.org/10.1016/j.eswa.2023.119764
4. Chen, W., Liu, M., Zhou, X., Pan, J., Tan, H. Safety Helmet Wearing Detection in Aerial Images Using Improved YOLOv4. Computers, Materials and Continua, 2022, 72, 3159. https://doi.org/10.32604/cmc.2022.026664
5. Chen, P. Y., Hsieh, J. W., Wang, C. Y., Liao, H. Y. M. Recursive Hybrid Fusion Pyramid Network for Real-Time Small Object Detection on Embedded Devices. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, 402-403.
6. Cui, B., Liu, M., Li, S., Jin, Z., Zeng, Y., Lin, X. Deep Learning Methods for Atmospheric PM2. 5 Prediction: A Comparative Study of Transformer and CNN-LSTM-Attention. Atmospheric Pollution Research, 2023, 14(9), 101833. https://doi.org/10.1016/j.apr.2023.101833
7. Fang, Y., Ma, Y., Zhang, X., Wang, Y. Enhanced YOLOv5 Algorithm for Helmet Wearing Detection via Combining Bi-Directional Feature Pyramid, Attention Mechanism and Transfer Learning. Multimedia Tools and Applications, 2023, 1-25. https://doi.org/10.1007/s11042-023-14395-0
8. Fan, Z., Peng, C., Dai, L., Cao, F., Qi, J., Hua, W. A Deep Learning-Based Ensemble Method for Helmet-Wearing Detection. Peer Journal of Computer Science, 2020, 6, e311. https://doi.org/10.7717/peerj-cs.311
9. Farooq, M. U., Bhutto, M. A., Kazi, A. K. Real-Time Safety Helmet Detection Using YOLOv5 at Construction Sites. Intelligent Automation & Soft Computing, 2023, 36(1). https://doi.org/10.32604/iasc.2023.031359
10. Gao, T., Zhang, X. Investigation into Recognition Technology of Helmet Wearing Based on HBSYOLOX-s. Applied Sciences, 2022, 12(24), 12997. https://doi.org/10.3390/app122412997
11. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J. YOLOX: Exceeding YOLO Series in 2021. arXiv Preprint, 2021, arXiv:2107.08430.
12. Han, G., Zhu, M., Zhao, X., Gao, H. Method Based on the Cross-Layer Attention Mechanism and Multiscale Perception for Safety Helmet-Wearing Detection. Com-

puters and Electrical Engineering, 2021, 95, 107458. https://doi.org/10.1016/j.compeleceng.2021.107458

13. He, K., Zhang, X., Ren, S., Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9), 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824
14. Huang, Z., Li, W., Xia, X. G., Wu, X., Cai, Z., Tao, R. A Novel Nonlocal-Aware Pyramid and Multiscale Multitask Refinement Detector for Object Detection in Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing, 2018, 60, 1-20. https://doi.org/10.1109/TGRS.2021.3059450
15. Jayakanth, K. Comparative Analysis of Texture Features and Deep Learning Method for Real-Time Indoor Object Recognition. In 2019 International Conference on Communication and Electronics Systems (ICCES), 2019, 1676-1682. https://doi.org/10.1109/ICCES45898.2019.9002551
16. Jabnoun, H., Benzarti, F., Morain-Nicolier, F., Amiri, H. Video-Based Assistive Aid for Blind People Using Object Recognition in Dissimilar Frames. International Journal of Advanced Intelligence Paradigms, 2019, 14(1-2), 122-139. https://doi.org/10.1504/IJAIP.2019.102967
17. Jeyaraj, P. R., Samuel Nadar, E. R. Computer Vision for Automatic Detection and Classification of Fabric Defect Employing Deep Learning Algorithm. International Journal of Clothing Science and Technology, 2019, 31(4), 510-521. https://doi.org/10.1108/IJCST-11-2018-0135
18. Kulikajevas, A., Maskeliūnas, R., Damaševičius, R., Ho, E. S. 3D Object Reconstruction from Imperfect Depth Data Using Extended YOLOv3 Network. Sensors, 2020, 20(7), 2025. https://doi.org/10.3390/s20072025
19. Li, B., Zhang, X., Muñoz, J. P., Xiao, J., Rong, X., Tian, Y. Assisting Blind People to Avoid Obstacles: A Wearable Obstacle Stereo Feedback System Based on 3D Detection. In 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2015, 2307-2311. https://doi.org/10.1109/ROBIO.2015.7419118
20. Li, N., Lyu, X., Xu, S., Wang, Y., Wang, Y., Gu, Y. Incorporate Online Hard Example Mining and Multi-Part Combination into Automatic Safety Helmet Wearing Detection. IEEE Access, 2020, 9, 139536-139543. https://doi.org/10.1109/ACCESS.2020.3045155
21. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 2117-2125. https://doi.org/10.1109/CVPR.2017.106
22. Lin, Y., Liu, M., Yang, C., Li, S., Zhang, W. AC-YOLO: A Safety Helmet Detection Based on YOLOX. In Proceedings of the 2022 4th International Conference on Robotics, Intelligent Control and Artificial Intelligence, 2022, 827-832. https://doi.org/10.1145/3584376.3584524
23. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, 8759-8768. https://doi.org/10.1109/CVPR.2018.00913
24. Li, Z., Xie, W., Zhang, L., Lu, S., Xie, L., Su, H., Hou, W. Toward Efficient Safety Helmet Detection Based on YOLOv5 with Hierarchical Positive Sample Selection and Box Density Filtering. IEEE Transactions on Instrumentation and Measurement, 2022, 71, 1-14. https://doi.org/10.1109/TIM.2022.3169564
25. Marchewka, A., Ziółkowski, P., Aguilar-Vidal, V. Framework for Structural Health Monitoring of Steel Bridges by Computer Vision. Sensors, 2020, 20(3), 700. https://doi.org/10.3390/s20030700
26. Meshram, V. V., Patil, K., Meshram, V. A., Shu, F. C. An Astute Assistive Device for Mobility and Object Recognition for Visually Impaired People. IEEE Transactions on Human-Machine Systems, 2019, 49(5), 449-460. https://doi.org/10.1109/THMS.2019.2931745.
27. njvisionpower. Safety-Helmet-Wearing-Dataset, 2019. https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset. Accessed on June 01, 2023. https://doi.org/10.1109/THMS.2019.2931745
28. Peng, D., Sun, Z., Chen, Z., Cai, Z., Xie, L., Jin, L. Detecting Heads Using Feature Refine Net and Cascaded Multi-Scale Architecture. In 2018 24th International Conference on Pattern Recognition (ICPR), 2018, 2528-2533. https://doi.org/10.1109/ICPR.2018.8545068
29. Redmon, J., Farhadi, A. YOLOv3: An Incremental Improvement. arXiv Preprint, 2018, arXiv:1804.02767.
30. Song, H., Zhang, X., Song, J., Zhao, J. Detection and Tracking of Safety Helmet Based on DeepSort and YOLOv5. Multimedia Tools and Applications, 2023, 82(7), 10781-10794. https://doi.org/10.1007/s11042-022-13305-0
31. Tan, M., Pang, R., Le, Q. V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 10781-10790. https://doi.org/10.1109/CVPR42600.2020.01079

32. Trabelsi, R., Jabri, I., Melgani, F., Smach, F., Conci, N., Bouallegue, A. Indoor Object Recognition in RGBD Images with Complex-Valued Neural Networks for Visually-Impaired People. Neurocomputing, 2019, 330, 94-103. https://doi.org/10.1016/j.neucom.2018.11.032

33. Tian, Z., Shen, C., Chen, H., He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, 9627-9636. https://doi.org/10.1109/ICCV.2019.00972

34. Ultralytics. YOLOv5, 2020. https://github.com/ultralytics/yolov5. Accessed on June 10, 2023.

35. Urbonas, A., Raudonis, V., Maskeliūnas, R., Damaševičius, R. Automated Identification of Wood Veneer Surface Defects Using Faster Region-Based Convolutional Neural Network with Data Augmentation and Transfer Learning. Applied Sciences, 2019, 9(22), 4898. https://doi.org/10.3390/app9224898

36. Wang, C. Y., Bochkovskiy, A., Liao, H. Y. M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, 7464-7475.

37. Wang, C. Y., Liao, H. Y. M., Wu, Y. H., Chen, P. Y., Hsieh, J. W., Yeh, I. H. CSPNet: A New Backbone That Can Enhance Learning Capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, 390-391. https://doi.org/10.1109/CVPRW50498.2020.00203

38. Wang, H., Hu, Z., Guo, Y., Yang, Z., Zhou, F., Xu, P.A real-time safety helmet wearing detection approach based on CSYOLOv3. Applied Sciences, 2020, 10(19), 6732. https://doi.org/10.3390/app10196732

39. Wang, Z., Wu, Y., Yang, L., Thirunavukarasu, A., Evison, C., Zhao, Y. Fast Personal Protective Equipment Detection for Real Construction Sites Using Deep Learning Approaches. Sensors, 2021, 21(10), 3478. https://doi.org/10.3390/s21103478

40. Woo, S., Park, J., Lee, J. Y., Kweon, I. S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), 2018, 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

41. Wu, H., Zhao, J. An Intelligent Vision-Based Approach for Helmet Identification for Work Safety. Computers in Industry, 2018, 100, 267-277. https://doi.org/10.1016/j.compind.2018.03.037

42. Yang, B., Wang, J. An Improved Helmet Detection Algorithm Based on YOLOv4. International Journal of Foundations of Computer Science, 2022, 33(06n07), 887-902. https://doi.org/10.1142/S0129054122420205

43. Yu, H., Tao, Y., Cui, W., Liu, B., Shi, T. Research on Application of Helmet Wearing Detection Improved by YOLOv4 Algorithm. Mathematical Biosciences and Engineering, 2023, 20(5), 8685-8707. https://doi.org/10.3934/mbe.2023381

44. Zheng, Q., Tian, X., Yu, Z., Jiang, N., Elhanashi, A., Saponara, S., Yu, R. Application of Wavelet-Packet Transform Driven Deep Learning Method in PM2.5 Concentration Prediction: A Case Study of Qingdao, China. Sustainable Cities and Society, 2023, 92, 104486. https://doi.org/10.1016/j.scs.2023.104486

45. Zheng, Q., Tian, X., Yu, Z., Wang, H., Elhanashi, A., Saponara, S. DL-PR: Generalized Automatic Modulation Classification Method Based on Deep Learning with Priori Regularization. Engineering Applications of Artificial Intelligence, 2023, 122, 106082. https://doi.org/10.1016/j.engappai.2023.106082

46. Zheng, Q., Zhao, P., Wang, H., Elhanashi, A., Saponara, S. Fine-Grained Modulation Classification Using Multi-Scale Radio Transformer with Dual-Channel Representation. IEEE Communications Letters, 2022, 26(6), 1298-1302. https://doi.org/10.1109/LCOMM.2022.3145647

47. Zheng, Q., Zhao, P., Zhang, D., Wang, H. MR-DCAE: Manifold Regularization-Based Deep Convolutional Autoencoder for Unauthorized Broadcasting Identification. International Journal of Intelligent Systems, 2021, 36(12), 7204-7238. https://doi.org/10.1002/int.22586

48. Zheng, Q., Zhao, P., Li, Y., Wang, H., Yang, Y. Spectrum Interference-Based Two-Level Data Augmentation Method in Deep Learning for Automatic Modulation Classification. Neural Computing and Applications, 2021, 33(13), 7723-7745. https://doi.org/10.1007/s00521-020-05514-1

49. Zhou, Q., Chen, R., Huang, B., Liu, C., Yu, J., Yu, X. An Automatic Surface Defect Inspection System for Automobiles Using Machine Vision Methods. Sensors, 2019, 19(3), 644. https://doi.org/10.3390/s19030644

50. Zhou, X., Wang, D., Krähenbühl, P. Objects as Points. arXiv Preprint, 2019, arXiv:1904.07850.