Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.055115

ARTICLE

HWD-YOLO: A New Vision-Based Helmet Wearing Detection Method

Licheng Sun[1], Heping Li[2,3] and Liang Wang[1,4,*]

[1]College of Information Science and Technology, Beijing University of Technology, Beijing, 100124, China

[2]Chinese Institute of Coal Science, Beijing, 100013, China

[3]State Key Laboratory for Intelligent Coal Mining and Strata Control, Beijing, 100013, China

[4]Engineering Research Center of Digital Community of Ministry of Education, Beijing, 100124, China

*Corresponding Author: Liang Wang. Email: wangliang@bjut.edu.cn

Received: 17 June 2024 Accepted: 12 August 2024 Published: 12 September 2024

ABSTRACT

It is crucial to ensure workers wear safety helmets when working at a workplace with a high risk of safety accidents, such as construction sites and mine tunnels. Although existing methods can achieve helmet detection in images, their accuracy and speed still need improvements since complex, cluttered, and large-scale scenes of real workplaces cause server occlusion, illumination change, scale variation, and perspective distortion. So, a new safety helmet-wearing detection method based on deep learning is proposed. Firstly, a new multi-scale contextual aggregation module is proposed to aggregate multi-scale feature information globally and highlight the details of concerned objects in the backbone part of the deep neural network. Secondly, a new detection block combining the dilate convolution and attention mechanism is proposed and introduced into the prediction part. This block can effectively extract deep features while retaining information on fine-grained details, such as edges and small objects. Moreover, some newly emerged modules are incorporated into the proposed network to improve safety helmet-wearing detection performance further. Extensive experiments on open dataset validate the proposed method. It reaches better performance on helmet-wearing detection and even outperforms the state-of-the-art method. To be more specific, the mAP increases by 3.4%, and the speed increases from 17 to 33 fps in comparison with the baseline, You Only Look Once (YOLO) version 5X, and the mean average precision increases by 1.0% and the speed increases by 7 fps in comparison with the YOLO version 7. The generalization ability and portability experiment results show that the proposed improvements could serve as a springboard for deep neural network design to improve object detection performance in complex scenarios.

KEYWORDS

Object detection; deep learning; safety helmet wearing detection; feature extraction; attention mechanism

1 Introduction

The safety helmet is the most common personal protection equipment in workplaces with a high risk of safety accidents, such as construction sites and mine tunnels, which can effectively prevent and reduce head injuries from external threats [1]. Although workers appreciate the necessity of wearing the safety helmet and have been trained or qualified to wear it, some still would not like to wear it due

to the discomfort caused by the pinching and friction of the chin strap and high temperature. So, it is vital to inspect whether workers always wear safety helmets in workplaces.

In the early stage, some sensors, such as the accelerometer [2], altitude sensor, radio frequency identification sensor [3], and pressure sensor, are employed to perform safety helmet-wearing inspection. Although these nonvisual sensors can detect safety helmet wearing to a certain degree, there are still two obstacles to overcome. One is that these sensors cannot identify whether the safety helmet is worn, held, or placed in a particular place. The other is that these intrusive sensors prevent some workers from wearing safety helmets due to discomfort and privacy.

Recently, with the progress of artificial intelligence, methods based on visual sensors have been proposed to detect safety helmet-wearing automatically. These methods can continuously and even in real-time identify safety helmet-wearing from surveillance videos by exploiting features designed by handcraft [4] or learned by deep learning [5–8]. However, the accuracy and efficiency of these methods do not meet the needs of real applications, especially in workplaces where complex, cluttered, and large-scale scenes result in severe occlusion, scale variation, and perspective distortion in surveillance video images.

So, a new method of safety helmet-wearing detection based on the improved You Only Look Once (YOLO) and visual sensor, i.e., HWD-YOLO, is proposed, which can significantly improve detection accuracy and speed. Firstly, a multi-scale context aggregation module (MSCAM) is proposed in the backbone of HWD-YOLO to enhance the global and deep feature representation ability. Secondly, a detection block combining the perceptual field expansion and the attention mechanism is proposed in the prediction part to suppress uninterested objects and highlight small concerned objects effectively. Finally, the overall performance is balanced by exploiting some newly emerged modules. Extensive experiments on open datasets validate the proposed HWD-YOLO. Compared with the baseline and the state-of-the-art methods, it can better balance the accuracy and efficiency of safety helmet-wearing detection. The main contributions can be summarized as follows:

1. A new helmet-wearing detection method is proposed, which can significantly enhance helmet-wearing detection performance in accuracy and speed.

2. A multi-scale context aggregation network module is proposed to aggregate global context information. It can retain and fuse more multi-scale features and enhance the attention weight of concerned objects in aggregated features.

3. A detection network module combining perceptual field expansion and attention scheme is proposed to suppress uninterested objects and highlight concerned objects.

4. To balance performance, some newly emerged modules, such as implicit information, Transformer, and Ghostnet, are incorporated into the proposed network.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 elaborates on the proposed method. Section 4 reports on experiments. Finally, Section 5 concludes this paper.

2 Related Work

This section reviews developments in helmet-wearing detection, improvements on network detection performance, and YOLO.

2.1 Helmet-Wearing Detection

According to the sensor type, methods of helmet-wearing detection can be roughly classified into two categories: methods based on non-visual sensors and based on visual sensors.

The first category of methods usually detects helmet-wearing state with sensors assembled on the helmet. Kim et al. [2] used a three-axis accelerometer assembled on the helmet to detect whether the helmet was worn. Barro-Torres et al. [3] used the chinstrap sensor, altitude sensor, and radio frequency identification assembled on the helmet to improve detection accuracy. Besides pressure sensors, Bluetooth devices [4] are also assembled on the helmet to transmit sensing data to a computer for processing and recording. These methods can detect safety helmet-wearing to some extent. However, they are restricted to few applications due to the limitations of these sensors. Moreover, considering the comfort and privacy, most workers do not wear helmets due to these intrusive sensors.

The methods based on visual sensors collect data in a non-contact way, which can alleviate the limitations of non-visual sensor-based methods. They can be further divided into methods based on hand-crafted features and methods based on features learned by deep learning. The former leverages image information, such as intensity, color, and geometry shapes, to detect helmet-wearing. Hand-crafted features, such as histograms of oriented gradient and color histograms, Gaussian mixture models [9] and color histogram [10], and local binary patterns, Hu moment invariant and color histogram [11] are used to perform helmet-wearing detection. These methods work well in general cases. However, their performance is greatly affected by the cluttered environment and illumination change in images.

With the development of artificial intelligence, more and more helmet-wearing detection methods based on deep learning emerge. For example, the faster region-based convolutional neural network (Faster-RCNN) [6] is applied to safety helmet-wearing detection [7]. However, there is a big gap in detection accuracy and speed between these early deep learning-based methods and the need for applications. Recently, Redmon et al. [12] proposed an excellent end-to-end object detection framework, i.e., YOLO, and then different improved versions [13,14] were proposed to enhance the performance. Various algorithms based on YOLO [15–17] are proposed to fulfill object detection. For example, Wang et al. [15] introduced the relative-distance-aware transformer and Squeeze-and-Excitation attention into YOLOv5 to improve the performance of object detection. Variants of YOLOs were also applied to perform helmet detection [1,5,8,16]. For example, the AT-YOLO [5] introduces a channel attention module and a spatial attention module into the backbone and neck of YOLOv3, respectively, to improve the helmet-wearing detection performance. However, these methods still cannot be applied well to complex workspaces due to the limited accuracy and efficiency performance and poor generalization ability.

2.2 Improvements on Networks Detection Performance

2.2.1 Multi-Scale Feature Fusion

For deep neural networks, if there are too many layers, semantic information and feature details will diminish with the increase of network depth, thus causing a decline in detection performance. Multi-scale feature fusion via building a feature pyramid is an effective way to alleviate this problem. Following this idea, many methods have been proposed.

Most multi-scale feature fusion modules are based on the Cross Stage Partial (CSP) [17] structure. CSP uses the cross-stage feature fusion strategy to enhance the variability of features in separate layers. Therefore, it is widely used in many multi-scale feature fusion modules. Following the structure of

CSP, the first multi-scale feature fusion module is the spatial pyramid pooling (SPP) [18] network. It first applies three different size pooling operations on the input, then transmits the results to a fully connected layer. However, this module generally ignores the important feature information of small objects such as distant helmets. So, the atrous spatial pyramid pooling (ASPP) network [19] is proposed. It combines dilate convolution with different expansion rates to increase the receptive field to obtain multi-scale information. Both SPP and ASPP take the CSP structure.

In summary, existing multi-scale feature fusion structures can alleviate the diminishment of semantic information and feature details to some extent. However, like CSP, they still often ignore the feature information of small objects located at the edge of feature images and cannot fully exploit multi-scale features to obtain complete feature information. It is because that they either only use some local operations, such as dilate convolution, rather than the global information to enlarge the receptive field to increase the scale as much as possible, or only locally aggregate the scales together without considering the relationship between different scales of feature maps.

2.2.2 Attention Mechanism

The attention mechanism aims to imitate humans to pay more attention to important information and ignore irrelevant information. In practice, it makes networks only focus on essential parts or objects and suppress other uninterested information as humans do.

Various models of attention mechanisms have been proposed. At first, the spatial transformer network [20] is proposed to extract important spatial information from the input image. It maps the input image to another space and uses the spatial attention module (SAM) to extract the feature information of the spatial position coordinates. Then, SENet [21] is proposed to highlight important channel information via the channel attention module (CAM). Compared with SAM, CAM focuses more on the channel weights of each convolution layer and uses global information to calculate the weights of different channels. Later, DANet [22] combines spatial and channel attention via residual attention learning. It takes feature tensors before and after mask processing as the input of the next layer. So it can get more features and pay more attention to them. Recently, a lightweight module, the convolutional block attention module (CBAM) [23] has emerged. It combines spatial and channel attention. So, it has both advantages of them.

However, there is still much room for improvement in attention mechanisms. For example, how attention techniques developed in one area can be applied to other areas is an interesting topic. Specifically, how to select and apply an attention mechanism to helmet-wearing detection is an interesting problem.

2.3 YOLO

YOLO was first proposed by Redmon et al. [12]. It first presents a real-time end-to-end one-stage method for object detection, which sets the stage of the subsequent advances of the YOLO family. YOLO series and their variants [24–26] have become the current mainstream real-time object detectors till now. Among them, YOLOv5 achieves better trade-offs between accuracy and efficiency compared to the latest versions of YOLO [25]. In addition, the latest versions of YOLO are mainly based on YOLOv5. For example, YOLOv8 enhances capabilities, such as instance segmentation, pose estimation, and classification based on YOLOv5 [25], and YOLOv9 [26] is trained on YOLOv5 codebase implementing programmable gradient information. So, YOLOv5 is selected as the base for developing the vision-based helmet wearing detection method.

YOLOv5 consists of four parts: input, backbone, neck, and prediction. The input part introduces mosaic data augmentation, adaptive anchor box calculation, and adaptive image scaling to perform data augmentation. The backbone part uses the CSP of YOLOv4 and introduces the Focus layer to improve the network performance. The neck part of YOLOv5 is the same as that of YOLOv4, where the feature pyramid network (FPN) and path aggregation network (PAN) structure, i.e., the FPN+PAN structure, is used. In the prediction part, the Generalized Intersection over Union (GIOU) Loss instead of the Complete Intersection over Union (CIOU) Loss is used as the loss function to improve the prediction performance. Finally, the anchor box will be applied to the output feature map to generate the final output vector with the class probability, the confidence score, and the bounding box.

Although YOLOv5 achieves good performance on object detection, there is still a gap to meet the need of applications of workspaces due to severe occlusion, scale variation, and perspective distortion in surveillance video/images caused by complex, cluttered, and large-scale scenes. In addition, YOLOv5 is the base of various newly emerging YOLO. So, it is selected as the baseline of the vision-based helmet wearing detection.

3 Proposed Method

3.1 Improved YOLO Network

As shown in Fig. 1, the proposed HWD-YOLO has a structure similar to YOLOv5. It can also be divided into four parts: input, backbone, neck, and prediction. The input part is the same as that of YOLOv5, where mosaic data enhancement, adaptive anchor box calculation, and adaptive image scaling are also used to improve learning efficiency and generalization capability. The whole structure of the proposed HWD-YOLO network follows the Reverse form of the U-net (R-Unet) [27]. This structure combines the local feature information extracted from the backbone part with up-samplings in the neck part. Similarly, the prediction structure is also combined with the neck part, which can improve prediction accuracy via feature information fusion. As shown in Fig. 1b, the main improvements of the proposed HWD-YOLO are as follows. (1) A new multi-scale contextual aggregation module (MSCAM) is proposed in the backbone part. (2) A new detection block (DetecBlock) is proposed in the prediction part. (3) Some newly emerged modules are also incorporated into the proposed network to enhance the safety helmet-wearing detection performance: the Ghostnet [28] is used in the backbone to modify the block constituting of convolution (Conv), batch normalization (BN) and Sigmoid weighted linear unit (SiLU) layers (CBS) to make the network more lightweight, the multi-head self-attention mechanism [29] of Transformer (MHSA) is applied in C3 layer in the neck to fine-tune and enhance the aggregated feature information, and prior knowledge of implicit information [30] is incorporated to speed up the detection speed.

3.2 Multi-Scale Contextual Aggregation Module

With the increase of network depth, some feature information diminishes or even vanishes. To remedy this information loss, a multi-scale context aggregation module is proposed. Fig. 2a shows the structure of MSCAM, where the dilate convolution [31] is denoted by Dconv.

First, MSCAM exploits the attention mechanism, i.e., CBAM as shown in Fig. 2b, to extract more features from space and channel dimensions. So, it can increase the probability of successfully detecting small objects by capturing important feature information on a global scale. Secondly, MSCAM takes a structure similar to ASPP to deal with the edges and regions with details while retaining the maximum pooling structure of SPP to highlight their feature information. The proposed MSCAM combines dilate convolution and maximum pooling to aggregate multi-scale feature information within a larger

receptive field. It can consider small objects in the global scope. So, the human head and the worn safety helmet, which are distant from the camera, can be detected as a whole. This is one reason why the proposed network can perform well on safety helmet-wearing detection. Finally, the deep features are further extracted from the feature maps with the cascade structure and cross-stage strategy to alleviate shortcomings of feature mapping via concatenation.

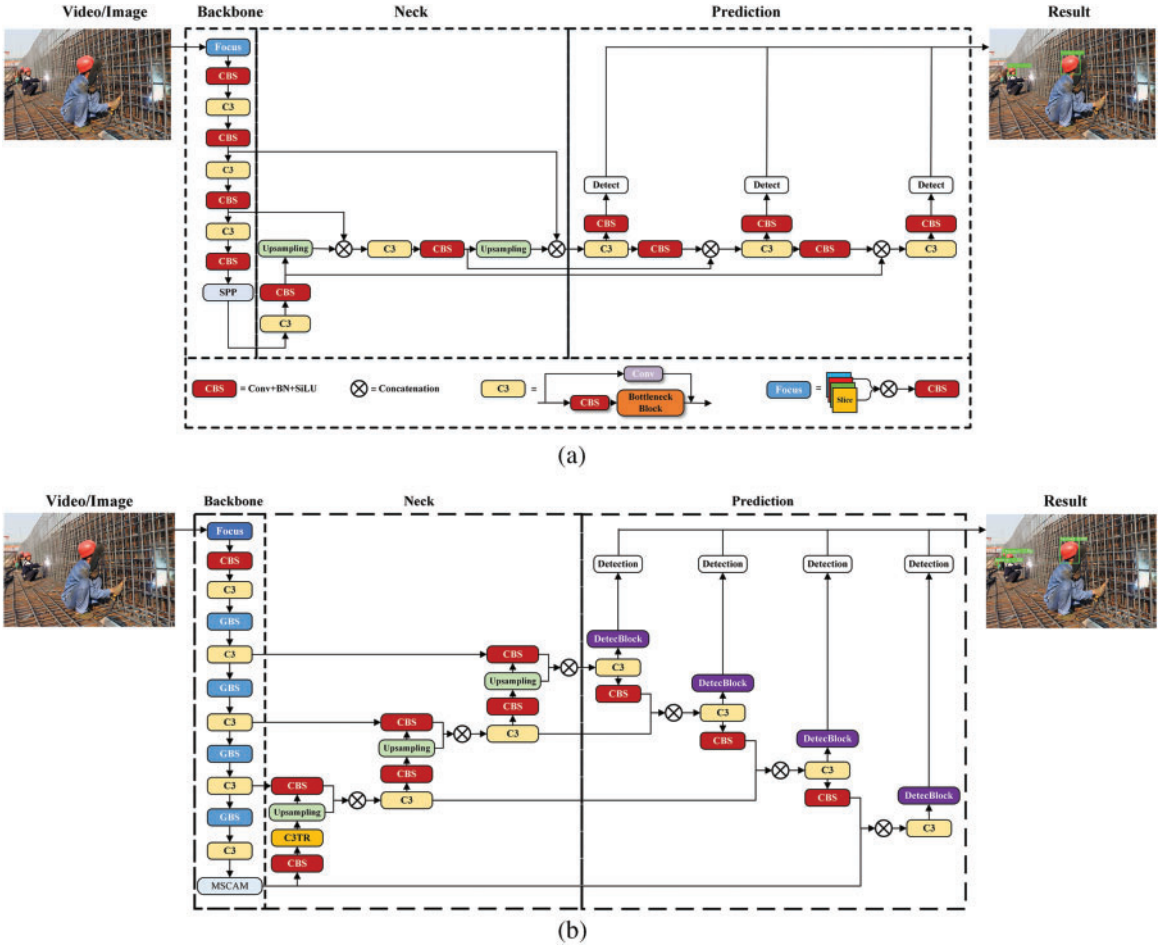

Figure 1: Scheme of YOLOv5 and the proposed HWD-YOLO. (a) YOLOv5. (b) The proposed HWD-YOLO

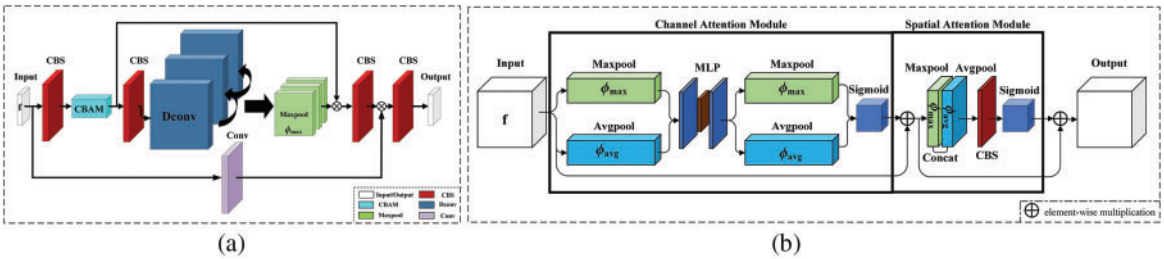

Figure 2: Structures of MSCAM and CBAM. (a) MSCAM. (b) CBAM

The calculation of the MSCAM can be expressed as

$$\begin{aligned}\text{Output} = \text{CBS}(\text{Concat}(\text{Conv}(\mathbf{f}); \text{CBS}(\text{Concat}(\text{CBAM}(\text{CBS}(\mathbf{f}));\\ \phi_{\max}(\text{Dconv}_1(\text{CBS}(\text{CBAM}(\text{CBS}(\mathbf{f}))))));\\ \phi_{\max}(\text{Dconv}_2(\text{CBS}(\text{CBAM}(\text{CBS}(\mathbf{f}))))));\\ \phi_{\max}(\text{Dconv}_3(\text{CBS}(\text{CBAM}(\text{CBS}(\mathbf{f}))))),\end{aligned} \tag{1}$$

where CBS denotes the block constitutes of convolution (Conv), batch normalization (BN) and Sigmoid weighted linear unit (SiLU), i.e., CBS = Conv + BN + SiLU, Conv $(\cdot)$ represents the convolution operation with the kernel size of 1×1 and stride of 1, Concat $(\cdot)$ represents the concatenation operation, $\phi_{\max}$ denotes the max pooling operation, $\text{Dconv}_i(\cdot)$ denotes the dilate convolution with an expansion rate of i and kernel size of 3×3, CBAM represents the convolutional block attention module shown in Fig. 2b, and $\mathbf{f}$ represents the input feature map.

The CBAM module can obtain important contextual global feature information from space and channel dimensions. It consists of Channel Attention Module (CAM) and Spatial Attention Module (SAM), as shown in Fig. 2b. The CAM compresses the feature map in the spatial dimension and pays attention to the more critical feature information, as shown in Fig. 3. Firstly, the input feature information passes through the average pooling and max pooling separately and then is aggregated and fed into the multi-layer perceptron (MLP) network. After that, the output separately passes through the average pooling and max pooling again. Finally, a ReLU is applied to the output. The calculation equation of the CAM can be expressed as

$$\text{CAM}(\mathbf{f}) = \text{Sigmoid}(\phi_{\text{avg}}(\text{MLP}(\phi_{\text{avg}}(\mathbf{f}) \oplus \phi_{\max}(\mathbf{f}))) \oplus \phi_{\max}(\text{MLP}(\phi_{\text{avg}}(\mathbf{f}) \oplus \phi_{\max}(\mathbf{f})))), \tag{2}$$

where $\oplus$ denotes pixel-wise multiplication, $\phi_{\max}$ and ϕ_{avg} denotes the max pooling and average pooling operation, respectively, whose calculation equation can be expressed as

$$\phi_{\max} = \max_{c=\{1,\cdots,C\}} \phi^{\text{c}}(\text{h}, \text{w}), \tag{3}$$

$$\phi_{\text{avg}} = \frac{\sum_{c=1}^{C}(\phi^{\text{c}}(\text{h}, \text{w}))}{C}, \tag{4}$$

where c represents the number of channels of the feature map, C represents the total number of channels in the feature map, ϕ^{c} represents the c-th channel of the feature map, and (h, w) represents the position in the space.

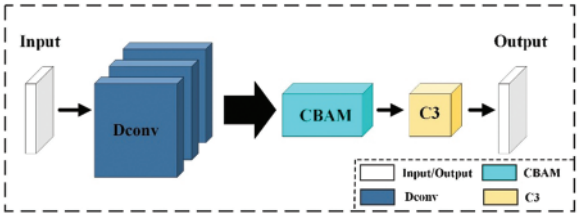

Figure 3: Structure of DetecBlock

The SAM compresses the channel information and aggregates the detailed feature information of the smaller objects, as shown in Fig. 2b. Average pooling and max pooling are used for aggregation

in the channel dimension. Then, their outputs are combined to obtain spatial attention feature information. The calculation equation is

$$\mathrm{SAM}(\mathbf{f}) = \mathrm{Sigmoid}\left(\mathrm{Conv}\left(\left[\phi_{\mathrm{avg}}(\mathbf{f}); \phi_{\mathrm{max}}(\mathbf{f})\right]\right)\right). \tag{5}$$

Although small objects, such as distant helmets, have smaller scales in the feature map, and there are differences in the representation of small objects in different scales of the feature map, the proposed MSCAM applies the CBAM to enhance the attention weight of the concerned objects before aggregating multi-scale features. This makes the module pay more attention to the detected concerned small objects during subsequent aggregation. In addition, the dilate convolutions [31] with the kernel size of 3×3 and dilation rates of 1, 2 and 3 are used to solve the problem of small or distant objects' features diminishing by expanding the perceptual field without loss of resolution or coverage. All of these make the proposed MSCAM effectively enhance the feature extraction ability for concerned objects and reduce the interference of uninterested objects.

3.3 Detection Block

Similar to YOLOv5, the backbone and neck parts formulate the FPN structure, and the neck and the prediction parts compose the PAN structure in the proposed HWD-YOLO. Generally, some feature information may diminish or even vanish after passing through FPN and PAN. For example, some features of small or distant objects vanish. Especially, to detect helmets far from the camera, the prediction part of the proposed network contains an additional detection branch in compare with that of YOLOv5. All of these make the final detection result poor. In order to retain deep-level feature information and increase the detection accuracy, a new detection block, i.e., DetecBlock, is proposed. The structure of DetecBlock is shown in Fig. 3, which consists of the dilate convolution, attention mechanism, and C3 layer.

Dilate convolution [31] mainly solves the problem of small or remote objects' features vanishing by expanding the perceptual field without loss of resolution or coverage. Here, dilate convolutions with the kernel size of 3×3 and dilation rate of 1, 2, 3 are used. When the feature information is relatively complete and includes some concerned and important small objects, the attention mechanism can assign more weight to the important information of concerned objects from both spatial and channel dimensions. Small object detection needs to pay attention not only to important information in space but also to important information in the channel dimension. The CBAM module can obtain important contextual global feature information from both dimensions. Then, a CBAM is incorporated into the DetecBlock following the dilate convolution. The C3 layer is a combination of the Bottleneck module based on DarkNet [32] and CSP, which uses CSP to learn residual features to obtain deep feature information. Thus, the DetecBlock can effectively detect deep features with global information and improve the detection performance. It also ensures that the human head and the wearing safety helmet are detected as a whole. This is the other reason why the proposed network can perform well on safety helmet-wearing detection.

Both DetecBlock and MSCAM contain the dilate convolution and attention mechanism. The differences between them are as follows. The MSCAM aims to fuse multi-scale features and emphasize the concerned features. So, it first uses the mixed domain attention mechanism to assign the weights of important feature information and then uses multi-scale dilate convolution to aggregate features in a large range of receptive fields. Thus, the MSCAM can effectively ensure that feature information contained in both space and channel dimensions is fused. On the contrary, the purpose of the DetecBlock in the prediction part is to predict. So, it first uses the dilate convolution to select a large

range of receptive fields, and then in these fields, the concerned objects are paid more attention via the mixed attention mechanism. Thus, the prediction results become complete without missed detection.

3.4 More Improvements

The proposed network also contains other improvements, such as the convolution layer based on Ghostnet (GBS) in the backbone part, the C3-based Transformer (C3TR) in the neck part, and the implicit information guided network (i.e., Implicit).

3.4.1 GBS

By analyzing the feature maps of YOLOv5, it is found that extracted feature maps in a multi-scale pyramid have many redundant features. These redundant features may cause a waste of unnecessary computation and training resources. So, the convolution layer based on Ghostnet [28], i.e., GBS, is introduced to replace the basic convolution layer of CBS in the backbone part. The cheap transformation of Ghostnet can greatly reduce the network calculation amount. In the proposed network, the cheap transformation is fulfilled by a 3×3 linear convolution kernel. Since it reserves the strong feature extraction ability of the original CBS and reduces the redundant features and calculation, GBS can significantly improve detection efficiency.

3.4.2 C3TR

The self-attention mechanism [29] can make the neural network pay more attention to the important features and global context information. However, each self-attention mechanism can only obtain feature information in a single representation space. The multi-head self-attention mechanism (MHSA) can jointly attend to information from different representation subspaces [29]. So, the C3 module with a Bottleneck Block in YOLOv5 is modified to the C3TR module incorporating a Transformer Block. Fig. 4 shows the structures of C3 and C3TR. C3TR exploits the MHSA mechanism [29] to divide the model into multiple branches to form multiple subspaces, allowing it to focus on different aspects corresponding to concerned objects in different sizes. Each attention mechanism function is only responsible for a subspace of the final output, which can not only prevent overfitting due to the over-mining of a single feature but also be aggregated to ensure better feature extraction performance. Details of MHSA can be found in [29].

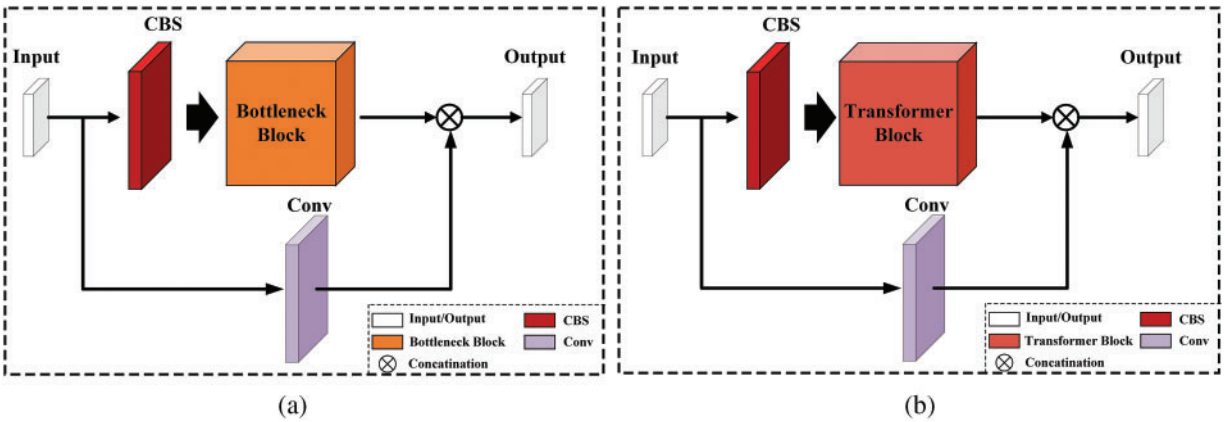

Figure 4: Structure of C3 and C3TR. (a) C3. (b) C3TR

3.4.3 Implicit Information

The implicit information [30] is introduced into the proposed network in three aspects: FPN feature alignment, prediction refinement, and multi-task learning.

For the network model of object detection with varying scenes and object sizes, it is hard to align the kernel space during model training, which usually leads to poor training results. To align each kernel space of the neural network, feature pyramid maps can be added and multiplied with the implicit representation in the FPN of the neck part, which corresponds to the rotation and scaling of the kernel space. So, the kernel space alignment problem can be effectively solved.

For the safety helmet-wearing detection in workspaces, the size of the safety helmet is small in images/videos due to the large-scale scenes and the remote distance to the camera. It is hard to accurately estimate the bounding box center of a small object at one time. So, implicit information is incorporated into the prediction part of the proposed network to refine prediction results. The adding operation can refine the center's coordinates of the predicted bounding box. The multiplication operation is used to refine anchor points' hyperparameters.

In addition, safety helmet-wearing detection involves multiple tasks, including the classification task of helmets and head, the prediction task of the bounding box's position, and the calculation task of the bounding box's confidence. Interaction between different tasks often occurs in the training process, resulting in poor training results. So, implicit information is introduced into each task to guide the training of the proposed network, which can improve the representation ability of each task and balance multiple tasks. This improvement can further improve the detection performance on the accuracy and speed of the proposed network.

4 Experiments

To validate the proposed HWD-YOLO, extensive experiments are performed on the open dataset, Safety Helmet Wearing Dataset [33]. This dataset contains 7581 images with and without safety helmets including cluttered scenes with different lighting conditions. The dataset includes a training set with 6064 images and a test set with 1517 images. Some randomly selected images from it are shown in the first row of Fig. 5.

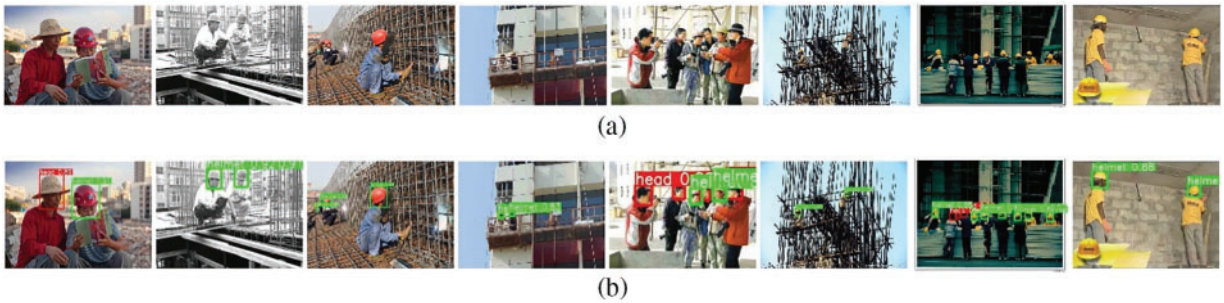

Figure 5: Some detection results of the proposed network. (a) Some randomly selected images. (b) Corresponding detection results of the proposed network

4.1 Benchmark Results

The proposed HWD-YOLO is evaluated on the open dataset with the mean average precision (mAP) and Frames Per Second (FPS) metrics. For comparison, some state-of-the-art methods, such

as Faster-RCNN [7], SSD [34], YOLOv3 [33], improved YOLOv3 [5], improved YOLOv4 [8], YOLOv5l [35], YOLOv5x [35], improved YOLOv5 [1,16], YOLOv5-BEH [36] and YOLOv7 [24] are also operated on Safety Helmet Wearing Dataset. The results are shown in Table 1, where the best results are shown in bold and the second-best results are shown in underlined.

Table 1: Comparison of objective results on open dataset

Method	mAP (%)	FPS (f/s)
Faster R-CNN [7]	87.56	4
SSD [33]	85.87	26
YOLOv3	88.13	22
AT-YOLO [5]	**96.5**	25.7
Improved YOLOv4 [8]	92.98	<u>43</u>
YOLOv5l	90.6	37
YOLOv5X	92.9	17
YOLO_CA [1]	93.6	**119**
Improved YOLOv5 [16]	95.94	29
YOLOv5-BEH [36]	95.9	\
YOLOv7-E6 [24]	95.3	26
Proposed	<u>96.3</u>	33

As shown in Table 1, the proposed HWD-YOLO significantly improves the safety helmet-wearing detection performance, which balances accuracy and speed well. It has the second-best mAP of 96.3% and the FPS of 33 f/s. Although the FPS of the proposed HWD-YOLO is not the best, it is greater than 30 f/s. So, it can operate in real-time on surveillance videos that generally have a frame rate of 30 f/s. Compared with the baseline, YOLOv5x, the mAP increases by 3.4%, and the FPS increases from 17 to 33 f/s. It should be pointed out that the best mAP, 96.5% of AT-YOLO, is reported in [5] on their private, larger, and cleaned dataset. The dataset of AT-YOLO extends 7581 images of the Safety Helmet Wearing Dataset to 13,620 images, and the number of hat instances has increased from 9031 to 27,236. More importantly, AT-YOLO cannot discern a common hat from a safety helmet. In addition, AT-YOLO only has a FPS of 25.7 f/s. The FPSs of the improved YOLOv5 [16] and YOLOv7-E6 [24] are also less than 30 f/s, which has the third-best mAP of 95.94% and the fourth-best mAP of 95.3%, respectively. That is, they could not deal with surveillance videos in real time. It is also worth noting that YOLOv7 [24] is the state-of-the-art network for general object detection in YOLO series, which has the same mAP as YOLOv9 [26] and that YOLOv9 has the same mAP as the latest YOLOv10 [37]. Compared with YOLOv7-E6, which has the best mAP among YOLOv7 variants [24], the mAP increases by 1.0%, and the FPS increases by 7 f/s. So, the proposed HWD-YOLO improves helmet-wearing detection performance and has a better trade-off between accuracy and efficiency.

In order to further subjectively evaluate the proposed HWD-YOLO, Fig. 5b shows the detection results of the proposed HWD-YOLO on the images shown in Fig. 5a, which are randomly selected from the Safety Helmet Wearing Dataset. In Fig. 5b, the bounding box in green shows the helmet-wearing detection results, and the bounding box in red shows the detected head without helmet-wearing. In addition, the confidence ratio of detection results is also shown in the output images. It can be seen that the proposed HWD-YOLO performs well on construction sites despite the severe

occlusion, scale variation, and perspective distortion in images caused by complex, cluttered, and large-scale scenes. It should be pointed out that the proposed HWD-YOLO cannot only succeed in safety helmet-wearing detection but also overcome the interference of the common hats and unworn helmets. As shown in Fig. 5b, the straw hat on the left of the first image and the fisherman bucket hat in the right of the fifth image are correctly classified as without safety helmet-wearing. In addition, the safety helmets left on a table in the middle left and the center of the seventh image, and in the left bottom of the last image do not confuse the proposed HWD-YOLO.

Fig. 6 further shows the comparison results of the proposed HWD-YOLO and YOLOv7 on some self-captured images. The original images are shown in Fig. 6a, and the helmet-wearing detection results of YOLOv7 and the proposed HWD-YOLO are shown in Fig. 6b,c, respectively. It can be seen that YOLOv7 sometimes fails to detect the helmet and misjudges some similar objects as a helmet. Especially in the first and third images, the helmets on the top of the image fail to be detected. Moreover, in the first, second, and third images, the head of a traffic cone and an ordinary hat are misjudged as helmets. On the contrary, the proposed HWD-YOLO succeeds in detecting the helmets on the top of the first and third images. The ordinary hat is successfully detected as a head without wearing a helmet in all four images, and the traffic cone in the images is not misjudged as a head or a helmet. So, the proposed HWD-YOLO is obviously superior to the state-of-the-art model, YOLOv7.

Figure 6: Detection results comparison between YOLOv7 and the proposed HWD-YOLO. (a) Original images. (b) Detection results of YOLOv7. (c) Detection results of the proposed HWD-YOLO

4.2 Ablation Study

A series of ablation experiments are performed to verify each proposed module's contribution to the proposed network. Table 2 shows the results of the ablation study.

In Table 2, the baseline network is YOLOv5x, corresponding to the first row. The Basic represents the basic network of the proposed HWD-YOLO without any proposed module, as shown in Fig. 1b, where the Basic network has more layers than the original YOLOv5 shown in Fig. 1a. Implicit

represents the implicit information introduced into the network to guide the training and refining of the whole network. GBS represents the convolutional layer based on GhostNet instead of a part of CBS. MSCAM represents the Multi-scale contextual aggregation module. DetecBlock represents the detection network module. C3TR represents the transformer-based C3 module. It can be seen that each proposal has significant performance improvement. The Basic network can improve both the accuracy and operation speed. Especially, the Basic dramatically improves the accuracy from 92.9% to 94.6%, and the Implicit significantly boosts the operation speed from 17 to 39 f/s. The proposed MSCAM, DetecBlock, and C3TR improve the accuracy further with a slight speed drop. Especially, the MSCAM boosts the mAP from 95.3% to 95.9%. The GBS can boost the speed with a slight decrease in accuracy. Moreover, the GBS can significantly reduce the number of parameters.

Table 2: Results of the ablation study

Module						Accuracy	Speed
Basic	Implicit	GBS	MSCAM	DetecBlock	C3TR	mAP (%)	FPS (f/s)
×	×	×	×	×	×	92.9	17
√	×	×	×	×	×	94.6	18
×	√	×	×	×	×	93.7	39
√	√	×	×	×	×	95.4	36
√	√	√	×	×	×	95.3	38
√	√	√	√	×	×	95.9	36
√	√	√	√	√	×	96.1	35
√	√	√	√	√	√	96.3	33

The position of the DetecBlock in the proposed network affects the detection accuracy. In order to get a better position of the DetecBlock, experiments are performed on different networks with the DetecBlock at different positions. Experimental results are shown in Table 3, where the baseline network, i.e., the first row, is the Basic network of the proposed HWD-YOLO only with Implicit information, corresponding to the fourth row of Table 2. DetecBlock represents the module incorporated into the Prediction part of the baseline network, where the position is shown in Fig. 1b. Bone DetecBlock represents the module being added after each C3 layer in the Backbone part of the Basic network. It can be seen that the mAP of the Bone DetecBlock is slightly reduced. This result may be because excessive consideration of global information when strengthening the feature weight of small objects in the Backbone part may lead to some irrelevant noise interference, which results in a decrease in network performance. DetecBlock is only integrated globally when all features of small objects have been enhanced in the prediction part. That is, the prediction part can see the concerned small objects globally. Therefore, the DetecBlock can improve accuracy slightly without reducing speed.

The Ghostnet can speed up detection by reducing the number of network parameters. There are two ways to apply Ghostnet to replace the basic convolution layer. One is that only the basic convolution layers of CBS layers in the Backbone part are replaced as we do in the proposed network, and another is that Ghostnet replaces all basic convolution layers of CBS in the whole network. So, experiments are performed to verify which is the better. Results are shown in Table 4, where the baseline network, i.e., the first row, is the basic network of the proposed HWD-YOLO only with

Implicit, which is the same as the first row of Table 3, the BGhostnet denotes the first way, and the AllGhostnet denotes the second way. It can be seen that with the incorporation of Ghostnet, the network speed and the number of parameters and calculation operations are significantly improved while the accuracy decreases. To balance the accuracy and the speed, the BGhostnet, i.e., only the basic convolution layers of CBS in the Backbone part being replaced by Ghostnet, is taken, which can ensure a good detection effect while reducing the computation.

Table 3: Results of different DetecBlock location

Module		Accuracy	Speed
DetecBlock	Bone DetecBlock	mAP (%)	FPS (f/s)
×	×	95.40	36
×	√	95.17	34
√	×	95.43	34

Table 4: Results of different number of convolution layers based on GhostNet

Module					
BGhostnet	AllGhost	mAP (%)	FPS (f/s)	Parameters (M)	FLOPS (M)
×	×	95.40	36	175.1	288.3
×	√	94.75	41	157.7	189.0
√	×	95.31	38	163.4	192.6

From these extensive ablation experiments, it can be seen that each improvement module can effectively improve the performance of safety helmet-wearing detection. So, the proposed improved safety helmet-wearing detection algorithm is validated, which has achieved excellent performance in both accuracy and speed.

4.3 Generalization Ability and Portability

Due to the limitations of the dataset availability and the complexity of the overall network structure, we only verified the generalization ability and portability of two modules, the MSCAM and DetecBlock. These two modules are directly transplanted to the YOLOv7 network without changing structure. Experimental results are shown in Table 5. It can be seen that the detection performance of YOLOv7 is greatly improved, where the MSCAM increases the mAP by 0.5% and the DetecBlock increases the mAP by 0.3% further. This experiment validates the generalization ability and portability of the proposed two improved modules.

4.4 Application

To further verify the proposed HWD-YOLO, we applied it to real applications. An intelligent helmet-wearing detection system has been developed and used on construction sites. This system can fulfill real-time high-precision safety helmet-wearing detection on complex construction sites. As shown in Fig. 7, the system can display the input video/image and the detection results. It also provides

the function of playback and save of the input video and detection results video for users to check later. Long-time successful operation of the developed intelligent safety helmet-wearing detection system also validates the proposed method.

Table 5: Generalization ability of MSCAM and DetecBlock

Module		mAP (%)	FPS (f/s)
MSCAM	DetecBlock		
×	×	95.3	26
√	×	95.8	24
√	√	96.1	24

Figure 7: Visual interface of the developed system

5 Conclusion

This paper presents a new vision-based safety helmet-wearing detection method based on the improved YOLO, HWD-YOLO. First, a new multi-scale context aggregation network module (MSCAM) is proposed in the backbone of HWD-YOLO to extract global information. Then, a new detection block (DetecBlock) is proposed to solve the feature vanish problem of small or remote objects. Some newly emerged modules, such as implicit information, Transformer and Ghostnet, are also incorporated into the proposed HWD-YOLO to balance the detection accuracy and speed. Extensive experiments on the open dataset validate the proposed HWD-YOLO. It reaches the state-of-the-art results of safety helmet-wearing detection on workspaces. The proposed improved modules

and strategy could be used in and guided the design of deep neural network for small object detection in computer vision systems.

Acknowledgement: The authors would like to thank the editors and reviewers.

Funding Statement: This work was supported in part by National Natural Science Foundation of China under Grant No. 61772050, Beijing Municipal Natural Science Foundation under Grant No. 4242053 and Key Project of Science and Technology Innovation and Entrepreneurship of TDTEC (No. 2022-TD-ZD004).

Author Contributions: Study conception and design: Licheng Sun, Heping Li, Liang Wang; data collection: Licheng Sun; analysis and interpretation of results: Licheng Sun, Heping Li, Liang Wang; draft manuscript preparation: Licheng Sun, Liang Wang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The Safety Helmet Wearing Dataset used in this article is available at https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset (accessd on 01 December 2023). The code that supports the findings of this study is available at https://github.com/Sun-Licheng/YOLOG (accessed on 07 May 2024).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] X. Wu, S. Qian, and M. Yang, "Detection of safety helmet-wearing based on the YOLO_CA model," *Comput. Mater. Contin.*, vol. 77, no. 3, pp. 3349–3366, Dec. 2023. doi: 10.32604/cmc.2023.043671.

[2] S. H. Kim, C. Wang, S. D. Min, and S. H. Lee, "Safety helmet wearing management system for construction workers using three-axis accelerometer sensor," *Appl. Sci.*, vol. 8, no. 12, Nov. 2018, Art. no. 2400.

[3] S. Barro-Torres, T. M. Fernández-Caramés, H. J. Pérez-Iglesias, and C. J. Escudero, "Real-time personal protective equipment monitoring system," *Comput. Commun.*, vol. 36, no. 1, pp. 42–50, Dec. 2012. doi: 10.1016/j.comcom.2012.01.005.

[4] M. W. Park and I. Brilakis, "Construction worker detection in video frames for initializing vision trackers," *Autom. Constr.*, vol. 28, no. 4, pp. 15–25, Dec. 2012. doi: 10.1016/j.autcon.2012.06.001.

[5] Q. Zhou, J. Qin, X. Xiang, Y. Tan, and N. N. Xiong, "Algorithm of helmet wearing detection based on AT-YOLO deep mode," *Comput. Mater. Contin.*, vol. 69, no. 1, pp. 159–174, Jun. 2021. doi: 10.32604/cmc.2021.017480.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017. doi: 10.1109/TPAMI.2016.2577031.

[7] H. Son, H. Choi, H. Seong, and C. Kim, "Detection of construction workers under varying poses and changing background in image sequences via very deep residual networks," *Autom. Constr.*, vol. 99, no. 4, pp. 27–38, Mar. 2019. doi: 10.1016/j.autcon.2018.11.033.

[8] J. Chen *et al.*, "Lightweight helmet detection algorithm using an improved YOLOv4," *Sensors*, vol. 23, no. 3, Jan. 2023, Art. no. 1256.

[9] A. S. Talaulikar, S. Sanathanan, and C. N. Modi, "An enhanced approach for detecting helmet on motor-cyclists using image processing and machine learning techniques," in *Adv. Comput. Commun. Technol.*, Singapore: Springer, 2019, vol. 702, pp. 109–119.

[10] J. Chiverton, "Helmet presence classification with motorcycle detection and tracking," *IET Intell. Transp. Syst.*, vol. 6, no. 3, pp. 259–269, Sep. 2012.
[11] H. Wu and J. Zhao, "An intelligent vision-based approach for helmet identification for work safety," *Comput. Ind.*, vol. 100, pp. 267–277, Sep. 2018. doi: 10.1016/j.compind.2018.03.037.
[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779–788.
[13] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 6517–6525.
[14] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 89–95.
[15] J. Wang *et al.*, "YOLO-DD: Improved YOLOv5 for defect detection," *Comput. Mater. Contin.*, vol. 78, no. 1, pp. 759–780, Jan. 2024. doi: 10.32604/cmc.2023.041600.
[16] L. Sun and L. Wang, "An improved YOLO V5-based algorithm of safety helmet wearing detection," in *Proc. Chin. Control. Decis. Conf. (CCDC)*, Hefei, China, 2022, pp. 2030–2035.
[17] C. Y. Wang *et al.*, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE Conf. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, Seattle, WA, USA, 2020, pp. 1571–1580.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015. doi: 10.1109/TPAMI.2015.2389824.
[19] L. C. Chen *et al.*, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018. doi: 10.1109/TPAMI.2017.2699184.
[20] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Montreal, QC, Canada, Dec. 2015, pp. 2017–2025.
[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 7132–7144.
[22] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 6450–6458.
[23] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Euro. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 3–19.
[24] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, 2023, pp. 7464–7475.
[25] J. R. Terven and D. M. Cordova-Esparaza, "A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond," 2023, *arXiv:2304.00501*.
[26] C. Y. Wang, I. H. Yeh, and H. Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," 2024, *arXiv:2402.13616*.
[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent., (MICCAI)*, Munich, Germany, Oct. 2015, pp. 234–241.
[28] J. Han, Y. Wang, and C. J. Xu, "GhostNets on heterogeneous devices via cheap operations," *Int. J. Comput. Vis.*, vol. 130, no. 4, pp. 1050–1069, Apr. 2022. doi: 10.1007/s11263-022-01575-y.
[29] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2021, pp. 6000–6010.
[30] C. Wang, I. H. Yeh, and H. Liao, "You only learn one representation: Unified network for multiple tasks," 2021, *arXiv:2105.04206*.
[31] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.
[32] J. Redmon, "Darknet: Open source neural networks in C," 2016. Accessed: Nov. 02, 2023. [Online]. Available: http://pjreddie.com/darknet

[33] Njvisionpower, "Safety-helmet-wearing-dataset," 2019. Accessed: Dec. 01, 2023. [Online]. Available: https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset
[34] W. Liu, D. Anguelov, and D. Erhan, "SSD: Single shot multibox detector," in *Proc. Euro. Conf. Comput. Vis. (ECCV)*, Amsterdam, Netherlands, Oct. 2016, pp. 21–37.
[35] H. G. Woong, "Smart construction," 2023. Accessed: Dec. 07, 2023. [Online]. Available: https://github.com/PeterH0323/SmartConstruction
[36] C. Shan, H. Liu, and Y. Yu, "Research on improved algorithm for helmet detection based on YOLOv5," *Sci. Rep.*, vol. 13, no. 1, Oct. 2023, Art. no. 18056. doi: 10.1038/s41598-023-45383-x.
[37] A. Wang *et al.*, "YOLOv10: Real-time end-to-end object detection," 2024, *arXiv:2405.14458*.