Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **6**, Number **2**, May **2021**

An Improved SSD Model for Helmet Wearing Detection

Ren-Jie Song, Zi-Ming Wang*

School of Computer Science
Northeast Electric Power University
No.169 Changchun Road, Jilin, Jilin 132012, China
1939811347@qq.com
*Corresponding Author, wziming_96@163.com

Received March 2021; revised April 2021

Abstract. *Wearing a safety helmet is an important part of safe production. In view of the complex scenes of the construction site and the small size of the helmets, an adaptive feature enhancement algorithm called SKC-SSD (Selective Kernel and Channel SSD) is proposed. The self-attention mechanism is used to improve the original SSD in this paper. Different from other methods of increasing the depth and width of the network, we perform adaptive feature enhancement on the feature map from the perspective of the channel and the receptive field (RF). We analyze the performance of two self-attention feature enhanced network blocks, and find that both of them are well compatible with SSD. In addition, we also find that the two have complementary advantages. Based on the above findings, we propose SKCNet to enhance the feature maps extracted by original SSD. In order to evaluate the effectiveness of SKC-SSD, we construct multiple networks for ablation study and comparative experiment on the Safety-Helmet-Wearing-Dataset. The experimental results show that our network has excellent accuracy and robustness compared with the original SSD and other control networks. The mAP is 5.1% higher than the original SSD, and can achieve real-time detection on the GPU.*
Keywords: CNN, Helmet wearing detection, Self-attention, Feature enhancement

1. **Introduction.** Helmet is a common personal protective equipment (PPE) [1]. Relevant studies show that wearing safety helmet can significantly reduce the probability of workers suffering from brain injury and ensures job safety to a certain extent. Many serious production and construction accidents are directly related to the fact that the construction workers do not wear safety helmets. Almost all countries have statutory regulations that safety helmet must be worn on construction sites [2]. However, some workers often violate these regulations due to comfort and other reasons, which directly threatens the personal safety of them. Managers use many methods to supervise the wearing of safety helmets in the construction site. The detection methods based on hardware need additional hardware equipments, which have a higher cost and greatly reduce the comfort of the constructors wearing safety helmets. Although the accuracy rate of manual on-site supervision is high, the operation area of general construction site is large and scattered, and the monitoring managers are prone to fatigue, which usually leads to missed inspections. Therefore, it is of great significance to supervise the wearing of safety helmets on the construction site effectively and automatically. The detection methods based on image processing have the advantages of low cost, fast deployment and can realize automatic

detection all day. In recent years, These methods have attracted the attention of more and more researchers.

Because the scene of construction site and the posture of workers are generally complex, and helmets usually appear in the form of small targets in images, the detection of helmet wearing has always been a challenging research field [3]. In the past decade, related scholars have researched and developed many technologies to automatically detect the wearing of helmets. Liu et al. [4] used the skin-color detection method to locate the face area, and estimated the area above the face as the helmet area. Then, take Hu moment as the feature vector of the image, and use Support Vector Machine (SVM) to judge the wearing condition of the helmet finally. Rubaiyat et al. [5] used Histogram of Oriented Gradient (HOG) to locate the position of workers, and then use Hough transform to detect the safety helmet. Zhou et al. [6] detected the face region and extract the statistical features, local binary pattern features and fast principal component analysis features of the region. Finally, BP artificial neural network and classifier are used to classify and predict the wearing of safety helmet. However, the above methods based on traditional image processing and machine learning have some problems, such as difficult to design features manually and poor ability to adapt to environmental changes [7, 8]. Traditional target detection algorithms based on artificial feature extraction cannot meet the needs of helmet wearing detection in complex scenes.

With the continuous development and application of computer vision technology and convolutional neural network (CNN), it is possible to automatically detect the wearing of safety helmet in the construction site with complex environment. Convolutional neural network plays an important role in the field of target detection because of its advantages of self-learning and generalization ability. The emergence of excellent convolutional neural network structures, such as AlexNet [9], VGGNet [10], GoogLeNet [11], has promoted the development of target detection [12]. In recent years, target detection algorithms based on CNN are mainly divided into two categories: two-stage target detection algorithm and single-stage target detection algorithm. Two-stage target detection algorithms are represented by R-CNN [13], Fast R-CNN [14], Faster R-CNN [15], R-FCN [16], Mask R-CNN [17], etc. This kind of algorithm divides the detection into two stages, including the extraction of regions proposal and the prediction of candidate target classes and positions. Single-stage target detection algorithms are represented by SSD [18], DSSD [19], RFBNet [20], YOLO [21], etc. Different from the two-stage target detection algorithm, the single-stage target detection algorithm directly generates the class and location information of the target to be detected by regression. Compared with the two-stage target detection algorithm, it not only ensures the high accuracy relatively, but also greatly improves the detection efficiency, which makes the field of target detection get explosive development.

Target detection algorithm based on CNN has been widely used in the field of helmet detection. Yan et al. [22] used two independent CNN to extract the image features of workers and combine the traditional machine learning method with Random Forest (RF). They propose an intelligent recognition algorithm for helmet wearing detection based on DCNN and RF. Shen et al. [23] and Fan et al. [24] proposed a helmet wearing detection method based on the combination of face detection and other network models. However, because of relying on face detection to locate workers, the algorithm cannot detect workers with their backs to the monitor. More importantly, the above methods use a combination of multiple models, which makes them have high detection accuracy, but this mechanism greatly improves the complexity of the algorithm, and it is hard to achieve real-time detection. Fang et al. [25] proposed to use the improved YOLO by MobileNet [26] to detect the wearing of safety helmet, and take worker as the training label unit rather

than helmets. Although the difficulty of small target detection is reduced, it is easy to cause missing detection when the clothing color of worker is similar to the background color. What's more, there is a problem that YOLO cannot precisely locate the target [27]. SSD combines the regression idea of YOLO with anchor mechanism of Faster R-CNN, which not only ensures the detection speed but also the prediction accuracy. However, SSD still has some problems, such as low detection accuracy for small targets.

In this paper, we analyze the limitation of SSD network. SSD only adds several additional convolution layers for multi-scale detection. This paper considers that the quality of features extracted from these convolution layers is relatively poor, which affects the improvement of SSD on the accuracy of helmet wearing detection task. Previous research works [28, 29] have proved that deepening the depth and widening the width of the network can improve the performance of the network to a certain extent, but this will make network training more and more difficult, mainly because of gradient disappearance or gradient explosion. To solve this problem, we enhance the feature from the perspective of channel and receptive field. Inspired by SENet [30] and SKNet [31], we propose a novel network called SKC-SSD for helmet wear detection. We perform ablation study on the Safety-Helmet-Wearing-Dataset [32], which proves the effectiveness of our method. Experiments show that, compared with the original SSD, our method greatly improves the detection accuracy and can achieve real-time detection.

2. **Methodology.**

2.1. **SSD object detection.**

SSD is a classic single-stage object detection algorithm, compared with the two-stage detection algorithm, it has achieved a great improvement in efficiency. At the same time, the accuracy is also considered. It has been widely used in the object detection task. SSD uses VGG16 as the backbone network of feature extraction, and adds additional convolution layers to obtain more features for feature fusion. The shallow feature map has rich details, but has not enough semantic information. However, the deep feature map contains a large number of semantic feature information with translation invariance, but it loses a lot of details. Therefore, SSD not only does object detection on the final feature map, but also integrates other five feature map information. After getting six feature maps extracted by backbone feature extraction network, the multi-stage feature map is regressed and classified to generate a series of regression boxes and classification scores, and then the Non-Maximum Suppression (NMS) algorithm is used to filter the regression boxes, the prediction results are finally obtained.

SSD sets different sizes of default boxes s_k for calculation units of the k-th feature map used for regression. m is the number of feature maps, and s_{max} is the maximum proportion of the default box to the feature map (generally set to 0.9). s_{min} represents the minimum proportion of the default box to the feature map (generally set to 0.2).

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{m-1}(k-1), k \in [1, m] \tag{1}$$

In the feature map used for regression, each calculation unit generates several anchors, including two squares and several rectangles. The side lengths of the two squares in the k-th feature map are as follows:

$$\begin{cases} l_{min} = & s_k \\ l_{max} = & \sqrt{s_{k+1} \times s_k} \end{cases} \tag{2}$$

l_{max} and l_{min} represent the side lengths of the larger square and the smaller square, respectively. The length and width of the rectangle are:

$$\begin{cases} width = & \sqrt{ratio} \times s_k \\ length = & \dfrac{1}{width} \end{cases} \tag{3}$$

Among them, the number of *ratio* determines the number of rectangular anchors. Through the multi-layer feature map prediction mechanism, SSD can find a batch of anchors that are closest to the ground truth in size and position (i.e., the largest IOU) in the receptive fields with various sizes. While improving the detection speed, it maintains high accuracy. SSD has achieved good detection results in VOC data set, but when it is applied to Safety-Helmet-Wearing-Dataset, the phenomenon of missing detection of small objects often occurs. We think that the original SSD relies on the Conv4_2 and FC7 layer to detect small objects. Although the resolution is high, the feature layer is shallow and lacks deeper semantic information. With the increase of the depth of the feature layer, the network loses shallow feature information, which affects the detection effect of small objects. The network structure of SSD is shown in the Figure 1.

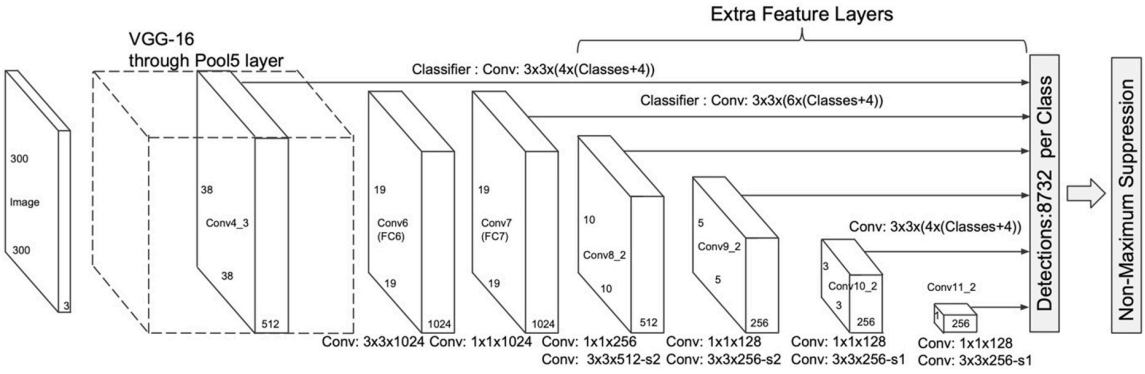


FIGURE 1. The network structure of SSD

2.2. **Adaptive channel selection strategy.** Squeeze-and-Excitation Net (SENet) aims to guide the network to learn the interdependence and importance between different channels, enhance the useful features according to the importance, and suppress the useless features for the detection task, so as to achieve better detection results. SENet mainly includes three operations: Squeeze, Exception and Reweight. The network structure is shown in the Figure 2.

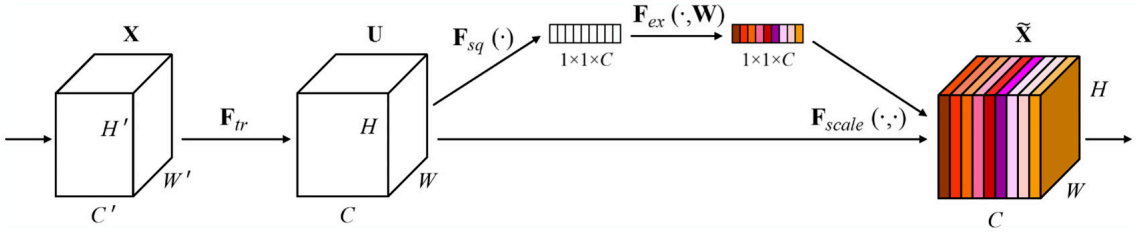


FIGURE 2. The network structure of SENet

Among them, F_{tr} represents performing convolution and pooling operations on the feature map X to obtain the feature map U with the number of channels C, and use it as the initial input of SENet.

$$F_{tr} : X \to U, X \in \mathbb{R}^{W' \times H' \times C'}, U \in \mathbb{R}^{W \times H \times C} \tag{4}$$

F_{sq} corresponds to the Squeeze operation, i.e., global average pooling is performed on the input feature layer U whose height and width are H and W respectively, and then the feature z_c whose size is 1×1 and the number of channels is C is output.

$$z_c = F_{sq}(U_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c(i, j), z \in \mathbb{R}^C \tag{5}$$

After the Squeeze operation, perform an Excitation operation F_{ex} on z_c to generate weight vector s for channels.

$$s = F_{ex}(z, W) = \sigma[W_2 \delta(W_1 z)], W_1 \in \mathbb{R}^{\frac{C}{r} \times C}, W_2 \in \mathbb{R}^{C \times \frac{C}{r}} \tag{6}$$

Among them, z is the output of the Squeeze operation, W_1, W_2 represent the Fully Connected (FC) operation, δ is the ReLU activation function, and σ is the Sigmoid activation function. In order to reduce the complexity of the model and improve the generalization ability, a bottleneck structure containing two fully connected layers is adopted here. The first FC layer plays a role of dimensionality reduction. The dimensionality reduction coefficient r is a hyperparameter (generally set to 16). Then, ReLU is used to activate, the second FC layer is used to restore the original dimension and Sigmoid function is used to get the weight of each channel. Finally, the Reweight operation F_{scale} is performed to multiply the initial input feature map and the weight vector s channel by channel to obtain the feature map $\tilde{X}$ which has been feature enhanced.

$$\tilde{X} = F_{scale}(U_c, s_c) = U_c \cdot s_c, \tilde{X} \in \mathbb{R}^{W \times H \times C} \tag{7}$$

Through the three operations of Squeeze, Exception and Reweight, the network can adaptively learn the weight coefficient of each channel, enhance the nonlinear relationship between the channels, and this attention mechanism can make the network pay more attention to the channel with the largest amount of information.

2.3. **Adaptive receptive field selection strategy.** Inspired by the fact that cortical neurons can dynamically adjust their receptive fields according to different stimulus, different convolution kernels are designed to extract features. Then the different information guided by each branch is used for feature fusion through Softmax. SKNet mainly includes three operations: Split, Fuse and Select. The structure diagram is shown in the Figure 3.

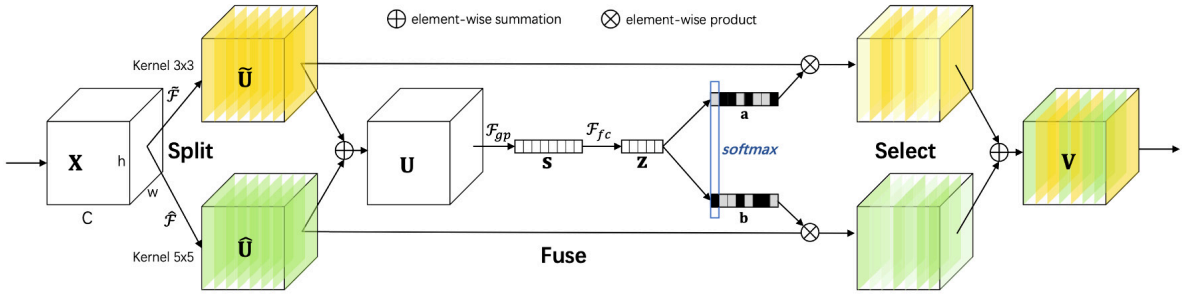


FIGURE 3. The network structure of SKNet

For any input feature map $X \in \mathbb{R}^{h\times w\times c}$, firstly the Split operation is performed, and convolute X using different sizes of convolution kernels respectively. $\tilde{F}$, $\hat{F}$ are composed of efficient grouping/depth convolution, batch normalization and ReLU activation functions, and new feature maps $\tilde{U}$, $\hat{U}$ are obtained respectively.

$$\tilde{F} : X \to \tilde{U}, \tilde{U} \in \mathbb{R}^{h\times w\times c} \tag{8}$$

$$\hat{F} : X \to \hat{U}, \hat{U} \in \mathbb{R}^{h\times w\times c} \tag{9}$$

The basic idea of Fuse is to use gates to control the information flow of multiple branches, so that these branches carry information of different scales to the neurons in the next layer. Firstly, the results of multiple branches are summed channel by channel to obtain the feature map U, and then the average pooling operation F_{gp} is used to embed the global information to obtain the vector S.

$$\begin{cases} U = & \tilde{U} + \hat{U}, U \in \mathbb{R}^{h\times w\times c} \\ S_c = & F_{gp}(U_c) = \dfrac{1}{H\times W}\displaystyle\sum_{i=1}^{H}\sum_{j=1}^{W} U_c(i,j) \end{cases} \tag{10}$$

In order to provide more accurate guidance for adaptive receptive field selection, the fully connected operation W is used to compress S into a more compact feature Z, where δ is the ReLU function, and B represents batch normalization. The number of compressed channels is d, and the attenuation rate r is used to control its value (generally set to 16), and L is the minimum value of d (generally set to 32).

$$\begin{cases} Z = & F_{fc}(s) = \delta[B(Ws)], W \in \mathbb{R}^{d\times c} \\ d = & max(\dfrac{c}{r}, L) \end{cases} \tag{11}$$

In the case of two branches, F_{fc} outputs two matrices A and B, where matrix B is redundant matrix, $B = 1 - A$. In the Select operation, $\tilde{F}$ and $\hat{F}$ are weighted by two vectors obtained by Softmax processing of matrixes A and B, and then the final feature map V is obtained by adding them. Where A_c is the c-th row of matrix A and a_c is the c-th element of a. B_c and b_c are the same as A_c and a_c respectively.

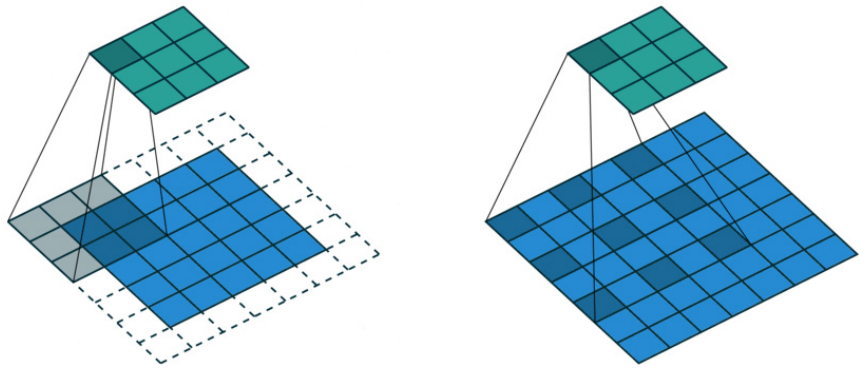
$$\begin{cases} V_c = & a_c \cdot \tilde{U}_c + b_c \cdot \hat{U}_c, V = [V_1, V_2, ..., V_c], V_c \in \mathbb{R}^{h\times w} \\ a_c = & \dfrac{e^{A_c Z}}{e^{A_c Z} + e^{B_c Z}} \\ b_c = & \dfrac{e^{B_c Z}}{e^{A_c Z} + e^{B_c Z}} \end{cases} \tag{12}$$

The essence of SKNet is to use multi-size convolution kernels in the network, and then perform multi-branch information fusion. Through this mechanism, the network can adaptively select the size of the receptive field, thereby enhancing the feature map.

2.4. **SKC-SSD safety helmet wearing detection model.** In this paper, SSD algorithm is applied to the helmet wearing detection task. Different from other tasks, the background of helmet wearing detection is generally complex, and the helmet often appears in the form of small objects in the image which brings great difficulties to the detection. The original SSD multi-feature maps fusion mechanism provides an excellent

idea for the simultaneous detection of large and small targets in the image. However, this paper argues that the feature extraction effect of original SSD on each feature layer limits the improvement of accuracy for helmet wearing detection task. Although Resnet101 [33] and other deep feature extraction networks can extract deeper features, they cannot achieve real-time detection due to the large number of network layers. Therefore, we proposed SKCNet, which enhances the feature extracted from the SSD backbone network to improve the detection performance of the original SSD for the helmet wearing detection task.

SKCNet adds two parallel convolutions after initial inputting the feature map, and both convolutions use a 3×3 size convolution kernel. The difference is that the second convolution is a dilated convolution with a dilation rate of 2. Dilated convolution not only increases the receptive field of the feature map, but also avoids the loss of feature information due to the use of pooling operations. Without increasing the number of parameters, the receptive field of a 5×5 convolution kernel is achieved. The two-branch convolution kernel we used is shown in Figure 4.



(a) (b)

FIGURE 4. The structure of kernels of parallel convolution. (a) is an ordinary convolution kernel with a size of 3×3, and (b) is a dilated convolution kernel with a size of 3×3 and dilation rate of 2. Dilated convolution greatly increases the receptive field of the input feature map without increasing the number of parameters.

In order to fuse the multi-branch feature information, we do the padding operation while doing the dilated convolution to obtain output feature maps of same size. At the same time, in order to reduce the number of parameters, we adopt a group convolution operation. The schematic diagram of group convolution in each branch of SKCNet is shown in Figure 5.

Among them, G represents the number of groups, C is the number of channels, and H and W represent height and width respectively. Two feature maps with same width and number of channels are generated from two branches. After adding the two feature maps channel by channel, global average pooling is performed, and the global receptive field information is embedded. Then the global eigenvectors are further compressed through the fully connected layer. The data with good spatial distribution makes it easier for the network to learn the classification features of each target. In order to accelerate the convergence and improve the accuracy of the model, the batch normalization regularization process is performed on each batch, and the ReLU function is used for activation. Then use fully connected operation and Softmax activation to process the compressed vector

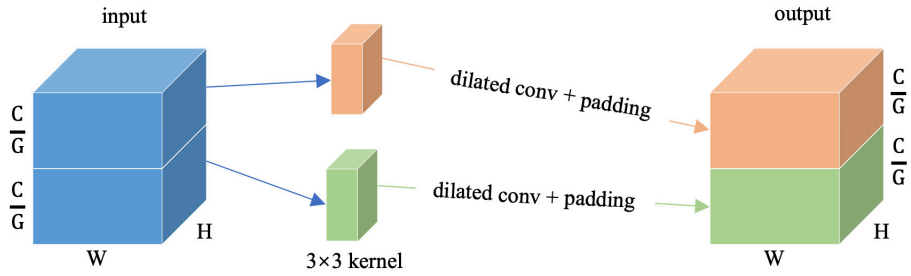


FIGURE 5. The schematic diagram of group convolution in each branch of SKCNet

into two weight vectors, and multiply them with the feature maps processed by parallel convolution and add them to realize the adaptive selection of receptive field. The feature maps obtained by the above operations are then processed by global average pooling to further embed global information. In order to make the network learn the non-linear relationship between the channels better, we perform two fully connected operations and Sigmoid activation on the obtained vectors embedded with global information to get the channel weight vectors, and multiply them with the feature map channel by channel. Finally, the feature map after the adaptive selection of receptive field and the channel is obtained. We call the above operation block SKCNet. Note that all of the above fully connected layers are implemented by convolution operations with kernels size of 1×1. The structure of SKCNet is shown in Figure 6.

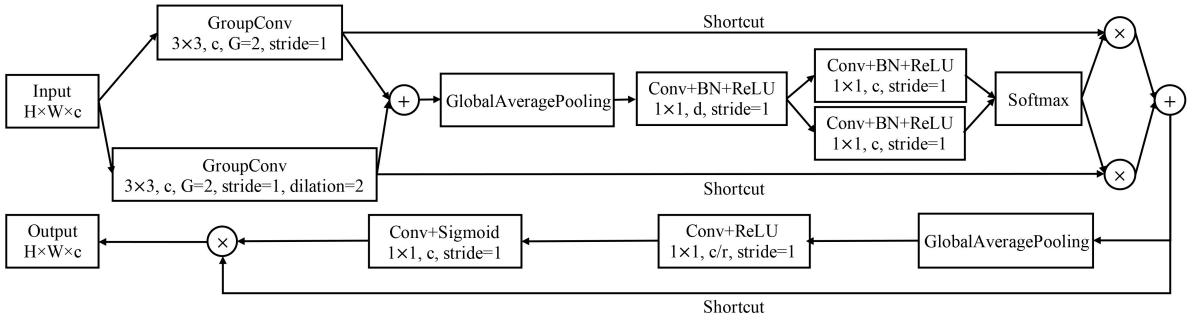


FIGURE 6. The network structure of SKCNet

This paper improves the feature extraction network on the basis of the SDD backbone network. In order to make SKCNet retain more channel information and maintain good compatibility, the input and output of SKCNet do not change the number of channels. So, we replace the Conv4_3 and Fc7 layers whose channel output is the same as their upper convolutional layer with SKCNet, and added one SKCNet after the Conv8_2, Conv9_2, Conv10_2, and Conv11_2 layers whose channel output is smaller than their upper convolutional layer respectively. On the basis of SSD multi-scale regression, feature enhancement is carried out through adaptive selection mechanism of receptive field and channel. The structure diagram of SKC-SSD is shown in Figure 7.

3. **Experiment.**

3.1. **Experimental platform and model training.** All experiments are conducted on a workstation equipped with NVIDIA Geforce GTX 1080Ti 11G for experimental training and testing. The CUDA version is 10.1, the cudnn version is 7.6.5, and the deep learning framework is Tensorflow 2.2. During the training process, online data enhancement is

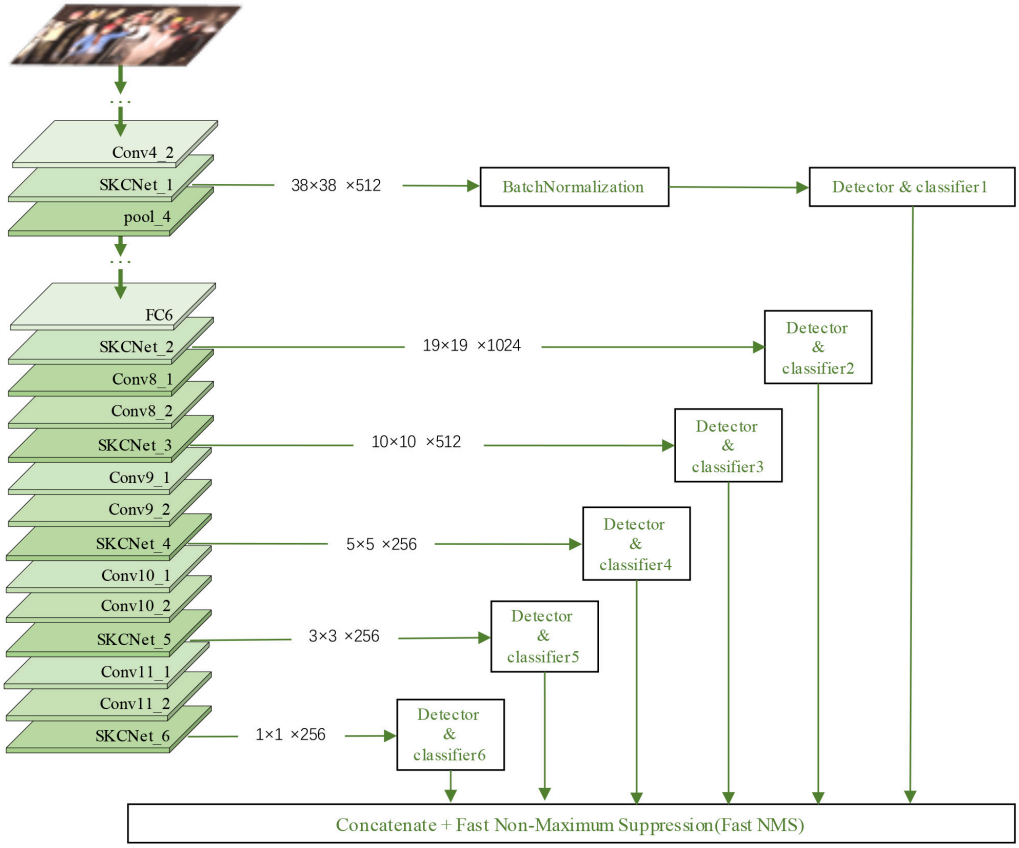


FIGURE 7. The network structure of SKC-SSD

performed on the training set, and transformations such as scaling, inversion, contrast change and brightness change are performed on the training image randomly. Set 10^{-3} as the initial learning rate lr, the model is trained for 128 epochs. Then we dropped lr to 10^{-4} and trained model for 128 epochs. Finally, set lr to 10^{-5} and perform multi-group cycle training, each group has 64 epochs. Record the optimal weight of each group, and use the early-stopping mechanism to prevent the model from overfitting during the training process.

3.2. **Ablation study and comparative analysis.** In order to verify that the adaptive selection mechanism of receptive field and channel can improve the detection accuracy of the SSD, and to prove the effectiveness of the SKCNet in the helmet wearing detection task. In addition to the original SSD, we designed three control networks for ablation study. At the same position described in Section 2.4, we add SENet, SKNet, SENet and SKNet (SENet is in front of SKNet) to form SE-SSD, SK-SSD, SE-SK-SSD, respectively. We use the same training set and test set to train and test the five network models. The results are shown in Table 1 and Figure 8.

As shown in Table 1 and Figure 8, the four methods improve the mAP of the original SSD for helmet wearing detection to a certain extent. Among them, SKC-SSD has the best performance, but the detection efficiency is reduced, mainly because the algorithm adds several SKCNet feature enhancement blocks, which leads to the increase of calculation. However, because SKCNet does not increase the network depth too much, and thanks to the group convolution and parallel acceleration of the GPU, our method can still achieve real-time detection. This verifies the feasibility of optimizing the SSD algorithm from the perspective of the channel and the receptive field in the task of wearing helmet

TABLE 1. Comparison of the effects of SSD incorporating different modules

Channel selection	RF selection	Enhanced first	Method	mAP(%)	FPS
		–	SSD	82.1	43
✓		–	SE-SSD	83.6	41
	✓	–	SK-SSD	84.8	36
✓	✓	Channel	SE-SK-SSD	86.4	35
✓	✓	Receptive field	SKC-SSD(ours)	87.2	35

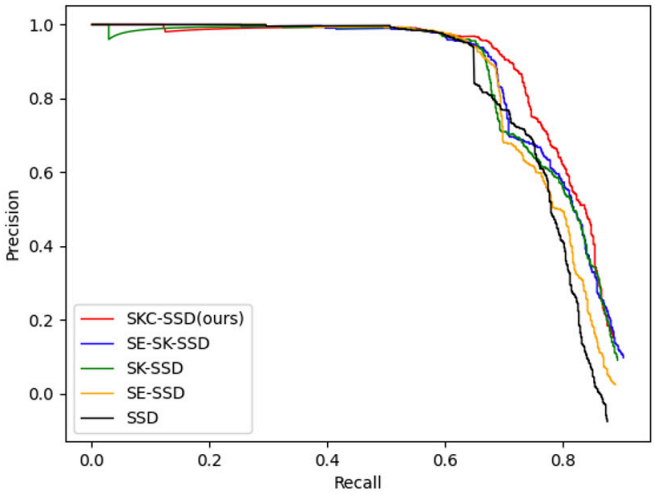
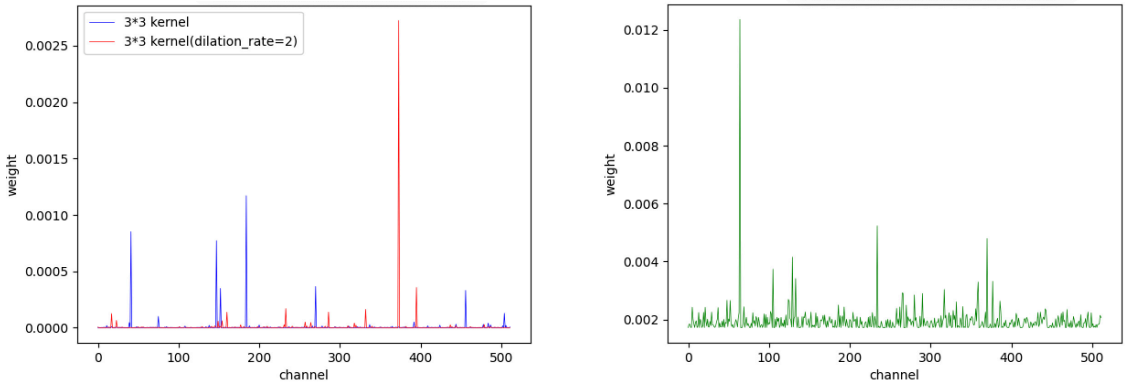


FIGURE 8. The PR curve of safety helmet wearing for different algorithms

detection. Compared with the original SSD, the mAP of SE-SSD and SK-SSD is increased by 1.5% and 2.7% respectively, indicating that in this detection task, adaptive selection of receptive fields is more effective than adaptive selection of channel. At the same time, compared with the original SSD, the mAP of SE-SK-SSD and SKC-SSD is increased by 4.3% and 5.1% respectively, which proves that SKNet and SENet have good compatibility and complementarity. More importantly, it is better to perform the adaptive selection of the receptive field first rather than the channel. In order to visually show how our proposed SKCNet works in the helmet detection task, we input the image shown in the third column of figure 11 to the network. And visualize the weights of adaptive selection of receptive field and channel of SKCNet behind Conv4_2, and the result is shown in the figure 9. (both number of output channels of adaptive selection are 512.)

As shown in Figure 9, SKCNet adaptively selects the receptive field and channel respectively on the feature map output by the Conv4_2. As shown in Figure 9(a), SKCNet has different weights distribution ranges for the kernels with two different receptive field sizes. We think that the reason for this is that the input image contains many small targets. Therefore, SKCNet makes a relatively large distribution range for the kernel with a smaller receptive field to extract more detailed information. As shown in Figure 9(b), SKCNet performs adaptive channel weights redistribution on the feature map after the adaptive receptive field selection to improve the importance of channel information which is useful for the task. To further verify our conclusion, we train several new feature extraction networks which are basically the same as backbone network of SSD. SENet, SKNet, SE-SKNet and SKCNet are added to the shallow network respectively. The reason for doing this is to show the differences of each module more intuitively by visualizing



(a) (b)

FIGURE 9. Visualization of SKCNet weights behind conv4_2 in SKC-SSD. (a) is the line chart of the weights of the adaptive selection of receptive field. (b) is the line chart of the weights of the adaptive selection of channel.

the feature map processed by these blocks. Since the pixel value of each channel after these block processing is relatively small, which is not conducive to visualization, we use standard scores (z scores) to process the pixels of each channel of the output feature maps. The pixel processing process is as equation (13).

$$\begin{cases} pix_i = & z_i \times m + b \\ z_i = & \dfrac{x_i - \bar{x}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}} \\ pix_i & = 0, if\ pix_i < 0 \\ pix_i & = 255, if\ pix_i > 255 \end{cases} \tag{13}$$

Among them, i is the i-th pixel in a channel, $\bar{x}$ is the average value of all pixels of the channel, n is the number of pixels, m is the magnification factor, and b is the bias. In the experiment, m is set to 64 and b is set to 128. Finally, all pixel values are limited to [0, 255]. The feature maps obtained by five networks processing the same picture is shown in the Figure 10.

It can be seen from Figure 10(b) that the features extracted from the original SSD backbone network are relatively fuzzy. Figure 10(c) shows that the feature map output by SENet has a lot of channel blanks or unobvious features. In addition, in some feature maps, such as the 43rd and 54th channels in Figure 10(c), are clearer than the original SSD. We think this is the result of channel selection by SENet. As shown in Figure 10(d), the features extracted by SKNet are clearer than those by SENet, indicating that more detailed features can be extracted by adaptive selection of receptive field. As shown in Figure 10(e), compared with SKNet, some channels of SE-SKNet with the addition of SENet are blurred or sharpened, which is more conducive to feature extraction. It can be seen from Figure 10(f) that our proposed SKCNet extracts a large number of features that are not extracted by other networks, such as the 14th, 23rd, and 32nd channels. By analyzing Figure 10(e) and Figure 10(f), we think that if the channel adaptive selection is performed first, the SENet will redistribute channel weights in advance according to the feature map processed by the 3×3 size convolution kernel. However, this redistribution may not be suitable for adaptive selection of receptive fields in the next step. More importantly, the feature map processed by the adaptive selection of channel may lose

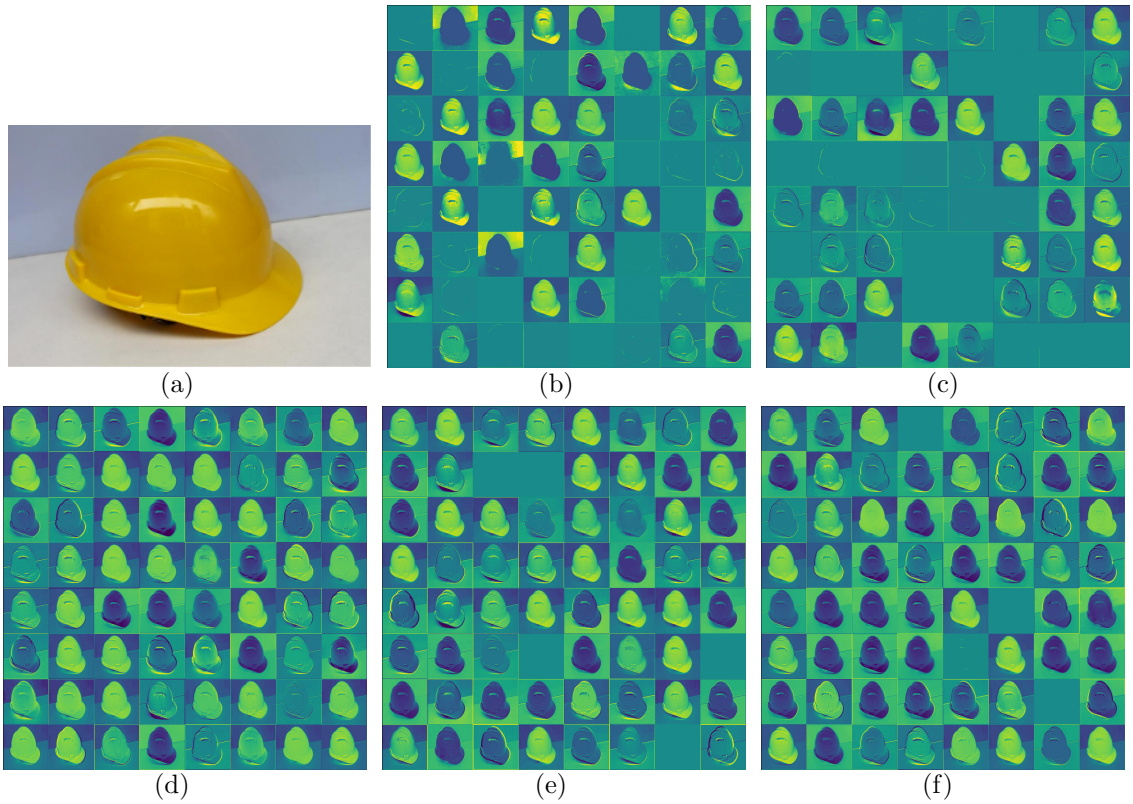


FIGURE 10. Comparison of feature maps visualization of different models (the number of output channels is 64). (a) is the input image. (b) is the output of original SSD. (c) is the output of SENet. (d) is the output of SKNet. (e) is the output of SE-SKNet. (f) is the output of our SKCNet.

some of the original features needed for adaptive selection of the receptive field. We think this is the main reason why SE-SK-SSD has worse performance than our SKC-SSD.

Figure 11 shows part of the detection results of five networks. The first column of Figure 11 contains both large and small targets relatively. The original SSD has a poor detection effect on small targets, and there are more cases of missed detection. SE-SK-SSD and SKC-SSD have good small target detection capabilities, but none of the networks detects the target on the leftmost side of the image. We think that this is caused by the few obscured target samples in the training set, and all networks have not learned the corresponding features well. Except for the original SSD, other networks have detected all the targets blocked by building materials in the second column, but SKC-SSD has significantly improved the confidence of each target. For a large number of overlapping small targets in the third column, the original SSD, SE-SSD and SK-SSD all have missed detection. SKC-SSD not only detects all targets, but also has higher confidence than other networks.

4. **Conclusions.** Aiming at the problems of complex background, large number of small targets and difficult detection of helmet wearing detection task, we proposed the SKC-SSD algorithm to improve the detection performance of SSD. Unlike most other improved algorithms, we improve the original SSD from the perspective of channel and receptive field feature enhancement. Compared with the original SSD, mAP is improved by 5.1%. The algorithm does not massively increase the network depth, so SKC-SSD not only significantly improves the accuracy of detection, but also achieves real-time detection, and

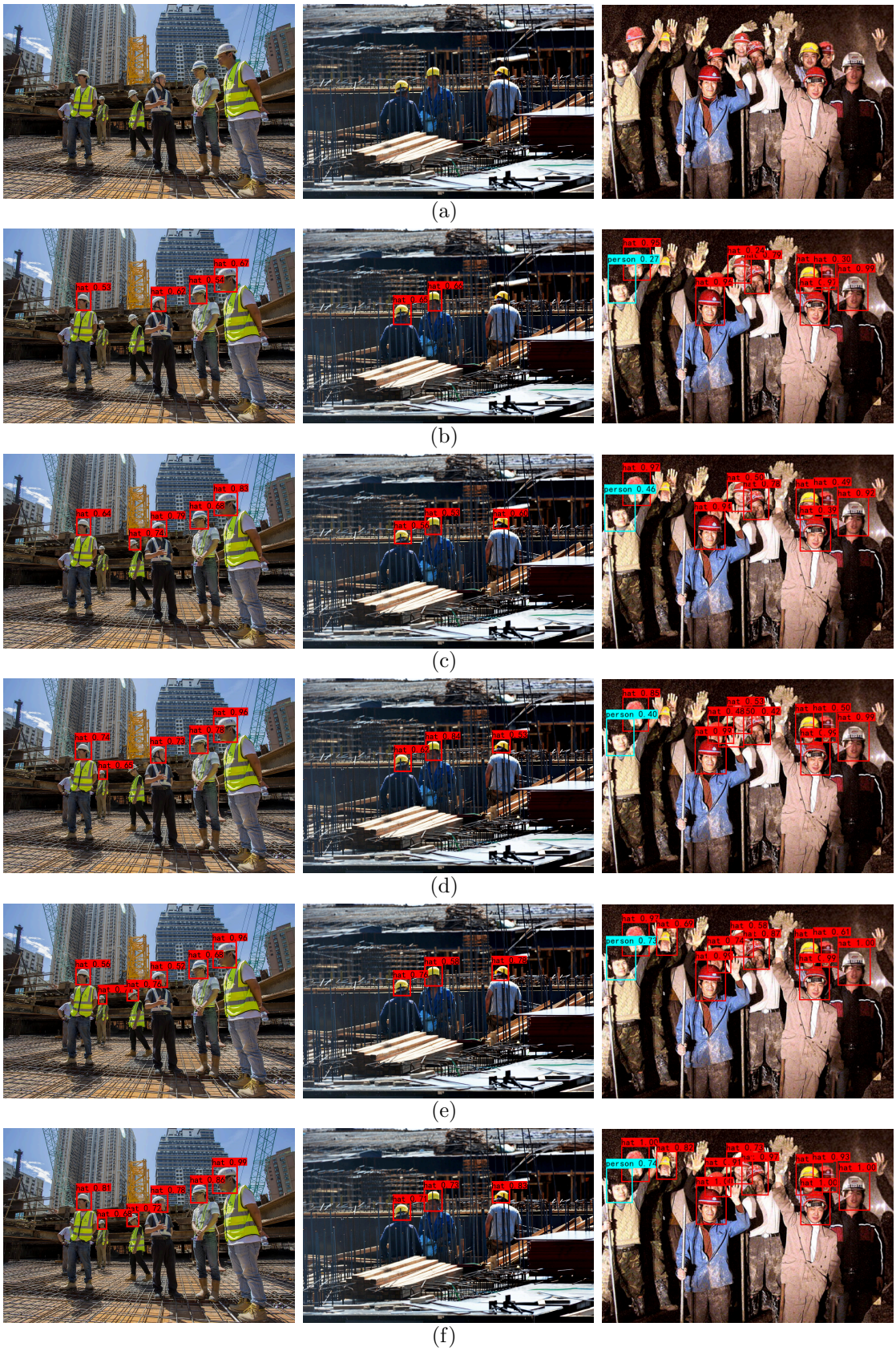


FIGURE 11. Comparison of detection results of different models. (a) is the input image. (b) is the detection result of original SSD. (c) is the detection result of SE-SSD. (d) is the detection result of SK-SSD. (e) is the detection result of SE-SK-SSD. (f) is the detection result of our SKC-SSD.

gets good detection results on the verification set. In order to further verify the effectiveness of SKCNet, we constructed new networks using SENet and SKNet, and conducted ablation study. By analyzing the experimental results, we come to our conclusion. The contributions of this paper are as follows: (a) We propose a new idea and method for helmet wearing detection, that is to improve the network from the perspective of channel and receptive field instead of increasing the network depth massively. While improving the accuracy, it also ensures the detection efficiency and can achieve real-time detection. (b) In the helmet wearing detection task, we prove that SENet and SKNet have good complementarity, and can be well compatible with the multi-feature maps fusion mechanism of the SSD to improve the detection effect. (c) Through the visualization of weight data and feature maps, we figure out the role of SENet, SKNet and SKCNet and explain how they work in the helmet wearing detection task. (d) Through experiments, we verify the different effects of relative position of SENet and SKNet on network performance when they are combined, and get the conclusion that conducting adaptive selection of receptive field first will get a better detection effect in the task of helmet detection.

Although our model has achieved good results, there are still some limitations. First of all, there will still be missed detections when detecting extremely small or occluded targets. Secondly, although our method can achieve real-time detection, compared with the original SSD, FPS still has a reduction. In future work, we will add more occluded target images to augment the data set and continue to optimize our model to solve the above problems in a targeted manner.

REFERENCES

[1] T. Wong, S. S. Man and A. Chan, Exploring the acceptance of PPE by construction workers: An extension of the technology acceptance model with safety management practices and safety consciousness, *Safety Science*, vol. 139, no.1 , pp. 105-239, 2021.

[2] R. R. Cabahug, A survey on the implementation of safety standards of on-going construction projects in Cagayan de Oro City Philippines, *Mindanao Journal of Science and Technology*, vol. 12, no. 1, pp. 12-24, 2014.

[3] F. Zhang, T. Y. Wu, J. S. Pan, G. Ding and Z. Li, Human Motion Recognition Based on SVM in VR Art Media Interaction Environment, *Human-centric Computing and Information Sciences*, vol. 9, no. 40, 2019.

[4] X. H. Liu and X. N. Ye, Application of Skin Color Detection and Hu Moment in Recognition of Safety Helmet, *Journal of East China University of Science and Technology (Natural Science Edition)*, vol. 40, no. 3, pp. 365-370, 2014.

[5] A. Rubaiyat, T. T. Toma, M. Kalantari-Khandani, Syed A. Rahman, L. W. Chen, Y. F. Ye and Christopher S. Pan, Automatic Detection of Helmet Uses for Construction Safety, *IEEE/WIC/ACM International Conference on Web Intelligence Workshops*, DOI:10.1109/WIW.2016.045.

[6] Y. Q. Zhou, H. R. Xue, X. H. Jiang, H. X. Sun and Y. Y. Xun, Low-resolution helmet recognition based on LBP statistical features, *Computer Systems & Applications*, vol. 24, no. 7, pp. 211-215, 2015.

[7] J. Su, L. Yang, H. Huang and G. D. Jin, Improved SSD Algorithm for Ship Detection of Small Targets in SAR Images, *Systems Engineering and Electronics*, vol. 42, no. 5, pp. 1026-1034, 2020.

[8] Y. J. Wang, P. P. Cao, X. S. Wang and X. Y. Yan, Research on Insulator Self Explosion Detection Method Based on Deep Learning, *Journal of Northeast Electric Power University*, vol. 40, no. 3, pp. 33-40, 2020.

[9] A. Krizhevsky, I. Sutskever and G. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017.

[10] K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, *Computer Science*, pp. 1-12, 2014.

[11] C. Szegedy, W. Liu, Y. Q. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, Going deeper with convolutions, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-9, 2015.

[12] K. Wang, C. M. Chen, MS Hossain, G. Muhammad, S. Kumar and S. Kumari, Transfer reinforcement learning-based road object detection in next generation IoT domain, *Computer Networks*, vol. 193, 108078, 2021.
[13] R. Girshick, J. Donahue, T. Darrell and J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, 2014.
[14] R. Girshick, Fast R-CNN, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1440-1448, 2015.
[15] S. Q. Ren, K. M. He, R. Girshick and J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, *International Conference on Neural Information Processing Systems*, vol. 39, no. 6, pp. 1137-1139, 2017.
[16] J. F. Dai, Y. Li, K. M. He and J. Sun, R-FCN: object detection via region-based fully convolutional networks, *Neural Information Processing Systems*, pp. 379-387, 2016.
[17] K. He, G. Gkioxari, P. Dollar and R. Girshick, Mask R-CNN, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2980-2988, 2017.
[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu and A. C. Berg, SSD: single shot multibox detector, *Proceedings of the European Conference on Computer Vision*, pp. 21-37, 2016.
[19] C. Y. Fu, W. Lin, A. Ranga, A. Tyagi and A. C. Berg, DSSD: deconvolutional single shot detector, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881-2890, 2017.
[20] S. Liu, D. Huang and Y. Wang, Receptive Field Block Net for Accurate and Fast Object Detection, *Proceedings of the European Conference on Computer Vision*, pp. 404-419, 2018.
[21] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, You only look once: unified, real-time object detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788, 2016.
[22] G. B. Yan, Q. Sun, J. Y. Huang and Y. D. Chen, Helmet Detection Based on Deep Learning and Random Forest on UAV for Power Construction Safety, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 25, no. 1, pp. 40-49, 2021.
[23] J. Shen, X. Xiong, Y. Li, W. He, P. Li and X. Y. Zheng, Detecting safety helmet wearing on construction sites with bounding-box regression and deep transfer learning, *Computer-Aided Civil and Infrastructure Engineering*, vol. 36, pp. 180–196, 2021.
[24] Z. M. Fan, C. B. Peng, L. C. Dai, F. Cao, J. Y. Qi and W. Y. Hua, A deep learning-based ensemble method for helmet-wearing detection, *PeerJ Computer Science*, vol. 6, e311, 2020.
[25] M. Fang, T. Sun and Z. Shao, Fast helmet wearing detection based on improved YOLOv2, *Optics and Precision Engineering*, vol. 27, no. 5, pp. 1196-1205, 2019.
[26] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto and H. Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, https://arxiv.org/abs/1704.04861.
[27] Y. Z. Chu, Y. Huang, X. F. Zhang and H. Liu, SSD image target detection algorithm based on self-attention, *Journal of HuaZhong University of Science and Technology (Natural Science Edition)*, vol. 48, no. 9, pp. 70-75, 2020.
[28] G. Huang, Z. Liu, L. Van Der Maaten and Kilian Q. Weinberger, Densely Connected Convolutional Networks, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261-2269, 2017.
[29] W. Li, H. L. Li, Q. B. Wu, X. Y. Chen and King Ngi Ngan, Simultaneously Detecting and Counting Dense Vehicles from Drone Images, *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9651-9662, 2019.
[30] J. Hu, L. Shen and G. Sun, Squeeze-and-excitation networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132-7141, 2018.
[31] X. Li, W. Wang, X. Hu and J. Yang, Selective Kernel Networks, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 510-519, 2019.
[32] Safety-Helmet-Wearing-Dataset, https://github.com/njvisionpower/Safety-Helmet-Wearing-Dataset/ [accessed October 28, 2020]
[33] K. He, X. Zhang, S. Ren and J. Sun, Deep residual learning for image recognition, *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016.