Computers and Electrical Engineering 95 (2021) 107458

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Method based on the cross-layer attention mechanism and multiscale perception for safety helmet-wearing detection

Guang Han [a,*], Mengcheng Zhu [a], Xuechen Zhao [a], Hua Gao [b]

[a] Engineering Research Center of Wideband Wireless Communication Technique, Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing, China
[b] Zhejiang University of Technology, Hangzhou, China

ARTICLE INFO

Keywords:
Object detection
Safety helmet detection
Attention mechanism
Feature pyramid
Multiscale perception

ABSTRACT

To solve the problem of low accuracy in existing safety helmet detection methods, a novel object detection algorithm based on Single Shot Multibox Detector (SSD) is proposed in this paper. The algorithm uses the spatial attention mechanism for low-level features and the channel attention mechanism for high-level features, this cross-layer attention mechanism can further refine the feature information of the object region. The proposed detection algorithm introduces a feature pyramid and multiscale perception module to improve its robustness to object scale change. In addition, an effective anchor box adaptive adjustment method is designed to adaptively adjust the scale distribution of each layer of the anchor boxes based on the feature map size. Experiment results demonstrate that our detection model has mean Average Precision (mAP) of 88.1% and 80.5% on helmet dataset and VOC 2007 dataset respectively, which is better than baseline by 15.65% and 3.4%.

1. Introduction

As increasing attention is given to construct additional infrastructure, such as buildings, the safety of construction workers is increasingly prominent. In construction sites, high-altitude falling objects occasionally injure workers in areas of mining, electric power, chemical industries and other work areas. Thus any person entering these relevant work areas must wear a safety helmet to ensure their safety. However, there are still many problems in the safety management of these industries. First, comprehensive coverage of safety education is difficult from the perspective of construction personnel, there are always construction workers that have a low-risk awareness and don't wear safety helmets as required. Second, most enterprises and supervision departments still rely on special personnel to supervise whether construction workers wear safety helmets, which has a low supervision efficiency and poor timeliness. On the problem of helmet wearing detection, computer vision technology can greatly reduce labor costs while effectively supervising.

Object detection is an important problem to be solved in the field of computer vision, it also is the basic task in video surveillance technology. Object detection is widely used in industrial inspection, robot navigation, aerospace and other fields, it is of great practical significance to reduce the consumption of human capital. A recent detection problem is whether or not an individual is wearing a safety helmet in construction sites. In recent years, there has been in-depth research in the field of object detection [1–4]. Park and Brilakis

This paper is for regular issues of CAEE. Reviews processed and approved for publication by the co-Editor-in-Chief Huimin Lu.

* Corresponding author.
E-mail addresses: hanguang8848@njupt.edu.cn (G. Han), ghua@zjut.edu.cn (H. Gao).

https://doi.org/10.1016/j.compeleceng.2021.107458
Received 28 March 2021; Received in revised form 3 September 2021; Accepted 7 September 2021
Available online 17 September 2021
0045-7906/

[1] used the HOG features to detect human bodies in the effective range and used the color histogram to detect whether there is a helmet on the identified human body area. Rubaiyat et al. [2] first combined the frequency domain information of an image with a human detection algorithm to detect construction workers, the color and circular Hough transform feature extraction method was used to detect whether individuals were wearing safety helmets. Du et al. [3] proposed a combination of image processing and machine learning methods for helmet detection in video sequences. The works were mainly divided into three categories. The first is face detection based on Haar-like features. The second is motion detection and skin color detection to reduce false positives in the face. The third is to use color information on the face area for helmet detection. Cai and Qian [4] constructed standard images of helmets, and the features were extracted along four directions. They finally used Gaussian functions to model the distribution of these features, the local images were divided into either helmet or non-helmet. Although the above research has achieved good results, there are still some defects and challenges in the detection of safety helmets. Most of these studies use traditional object detection methods, which require image processing to first perform grayscale, normalization, erosion, expansion processing of the image and other related operations. The extraction of image features relies primarily on manual designs, pattern recognition classification algorithms are used to detect objects. These hand-designed features are relatively unstable in complex construction environments. In addition, object scale change are less robust, the recognition process is complicated with a slow detection speed and other issues, it cannot reach real-time detection and high accuracy requirements in production environments. These problems are well-solved through the improvement and innovation of object detection algorithms based on deep learning.

Object detection algorithms based on deep learning are generally divided into two categories: method based on region proposal and method based on object regression. The former is also called two-stage detection. Its main idea is to first generate a series of sample candidate boxes and classify these samples through a convolutional neural network. Some classic algorithms include the Faster R-CNN [5] and R-FCN [6]. The latter is also called one-stage detection. Unlike the two-stage detection, one-stage detection does not necessarily generate sample candidate boxes, but directly converts the object location into regression problem, such as SSD [7], YOLO-v3 [8] and other algorithms. Due to the RPN structure, two-stage method represented by the Faster R-CNN has a high detection accuracy but a slow speed, which makes it difficult to reach real-time processing requirements for some scenes. One-stage method can achieve the shared features of a single training, the speed can be significantly improved while keeping a certain accuracy. However, both the one-stage and two-stage methods have the problem that the robustness to object scale change is poor, especially when detecting small objects.

To solve these problems, an object detection algorithm based on the cross-layer attention mechanism and multiscale perception is proposed with the SSD algorithm as baseline. The algorithm uses a spatial attention mechanism for low-level features and a channel attention mechanism for high-level features to further refine the feature information of the object area, which can generate more effective object features. Subsequent experiments prove that this method of using different attention mechanism networks for different sizes of feature maps can effectively improve the accuracy of the detection algorithm. In addition, our algorithm introduces a feature pyramid and multiscale perception module to help the network capture target information of different scales to enhance the network's robustness to target scale changes. In the multiscale perception module, it adopts feature adaptive fusion to instead of the simple proportional addition of features. Through learning the weight parameters, this strategy can effectively fuse the features of different layers according to their importance. In general, safety helmets are considered as detection task of small objects to a certain extent. Thus a fast and effective anchor box method with adaptive adjustment is proposed. This approach can be adaptive to adjust the scale distribution of each layer of the anchor boxes based on the size of the feature map, it is helpful to detect small objects on low-level feature maps.

The remainder of this paper is organized as follows. The next section presents the recent work on small object detection and main contributions of our algorithm. Section III gives the network framework of proposed CASP detection algorithm and the technique details, including the cross-layer parallel attention network, multiscale perception module, feature adaptive fusion and anchor box allocation strategy. We present the comparison of experiment result on two public datasets and ablation experiment analysis in Section IV. Finally, the paper is concluded in Section 5.

2. Related work

In actual scenes, object detection based on deep learning is sensitive to scale change of objects, especially when detecting small objects. According to the definition of the COCO dataset [9], targets with pixel areas less than 32×32 are considered as small objects, those with pixel areas between 32×32 and 96×96 are considered medium objects, and those with pixel areas greater than 96×96 are considered large objects. The main reasons for the difficulty of detecting small objects have the following three points. (1) Small objects contain fewer fine-grained features, such as edge or gray-scale information. As the training network deepens, these fine-grained features are easily lost, and there is less high-level semantic information. In addition, there may be some noise information in the image that misleads the training network to obtain incorrect features. (2) Due to the limitation of receptive field, it is difficult to extract the global semantic information contained in the background by using small receptive field to focus the features of the object itself. However, a larger receptive field is used to focus on the semantic information in the background, it maybe lose the features of small objects. (3) The convolutional neural network implements discrete feature extraction, making it difficult to achieve sub-pixel accuracy. In the neural network, each pixel in the deep layer feature may corresponds to 8, 16 or more pixels in the shallow layer feature. While this has little impact on larger objects, it significantly affects small object detection. These considerations have urged some experts to propose corresponding solutions along the following directions.

First, image pyramids of different resolutions are used for multiscale sliding window detection. A classifier with a fixed input resolution is used to detect the objects by sliding in each layer of the pyramid, the purpose is to detect small objects at the bottom of the

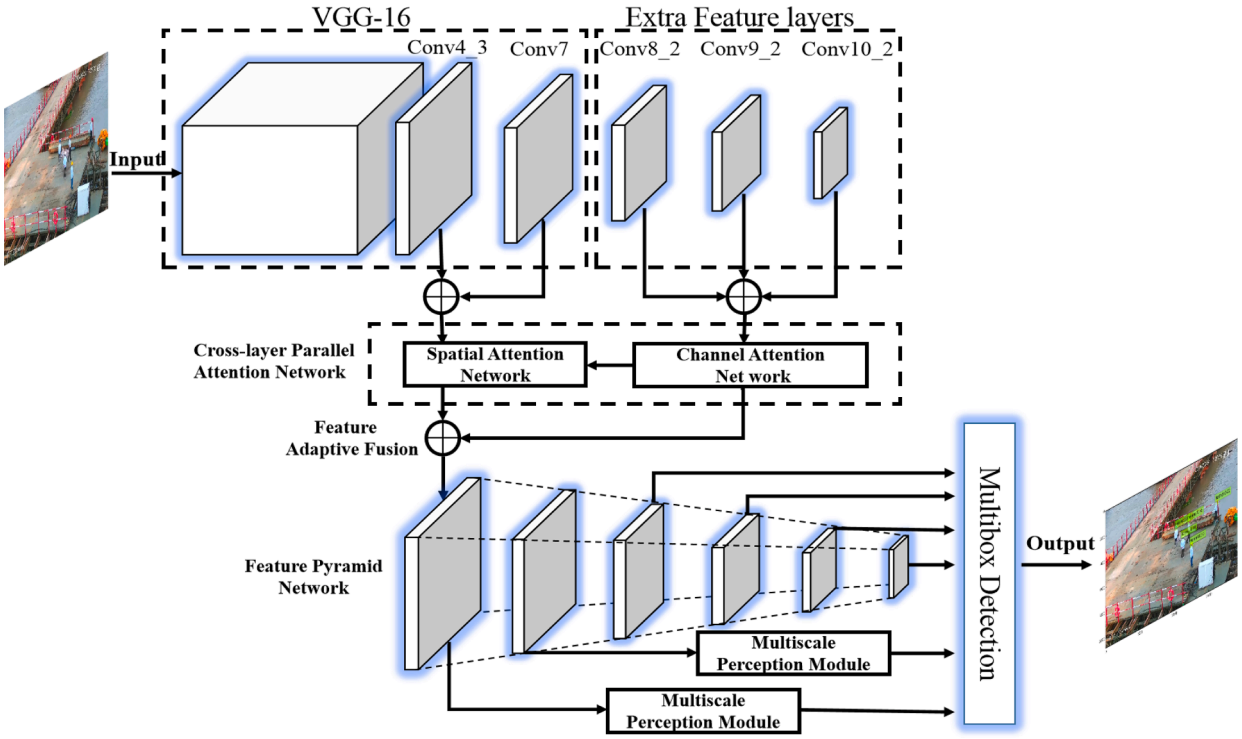

Fig. 1. The network framework of CASP detection algorithm proposed in this paper.

pyramid. For example, the face detector MTCNN [10] used the image pyramid to detect faces of different resolutions. The SNIP [11] proposed by Bharat Singh also adopted the idea of image pyramid. However, as all the images with different resolutions need to go through a convolutional neural network, this results in a low efficiency of the image pyramid, thus there are also some works from the perspective of features, such as FPN [12] and DSSD [13].

Second, the idea of multiscale feature fusion is adopted to fuse feature maps with different resolutions, it can improve the richness and information content of features and make it easily detect the objects of different sizes, such as the FPN proposed by Lin et al. Most of the original object detection algorithms only use deep features for prediction, but the FPN algorithm performs prediction independently for different feature layers, where deep-level features are fused with shallow-level feature by the up-sampling. DSSD also adopted a similar idea, fused the up-sampling feature map with the feature map of the same size during the down-sampling process, which can recover part of lost semantic feature (loss caused by the up-sampling). In addition, some methods also used the idea of feature fusion, such as RefineDet [14] and YOLO-v3.

Third, data augmentation is used, and the scale and distribution of anchor boxes is adjusted. For example, Kisantal et al. [15] thought that the low accuracy of small object detection had two primary reasons. 1) Few images contained small objects. 2) Even if some images contained small objects, they appeared less frequently. The authors proposed two simple methods to solve it: re-sampling the image with small objects, copying and pasting the small objects that appear in the image to other positions, this can increase the number of small object matching anchor boxes to improve the weight of small object training. Wang et al. [16] used generative adversarial networks to generate a super-resolution feature for small objects, and it was superimposed on the feature map of original small object to enhance its feature expression. In addition, some face detection methods considered changing the anchor box setting strategy to improve the detection effect of small objects, such as Face Boxes [17].

Fourth, some scholars have considered the influence of receptive fields on object detection, such as RFBNet [18] and TridentNet [19]. Both believed that the size of the receptive fields was crucial to detect small objects. In addition, some scholars also considered improvements on the feature alignment and ROI Pooling alignment. For example, Google adopted the Aligned Xception in DeepLabV3+, which had good results in the object detection and segmentation. Mask R-CNN [20] used the RoIAlign to solve the region mismatch caused by two quantization in the ROI Pooling operations.

This paper proposes an object detection algorithm based on a cross-layer attention mechanism and multiscale perception. Its main contributions are provided as follows.

- A cross-layer parallel attention network is proposed. By adopting the channel attention mechanism for the high-level fusion features and spatial attention mechanism for the low-level fusion features, the semantic information of the high-level fusion features and the fine-grained information of the low-level fusion features are further refined.

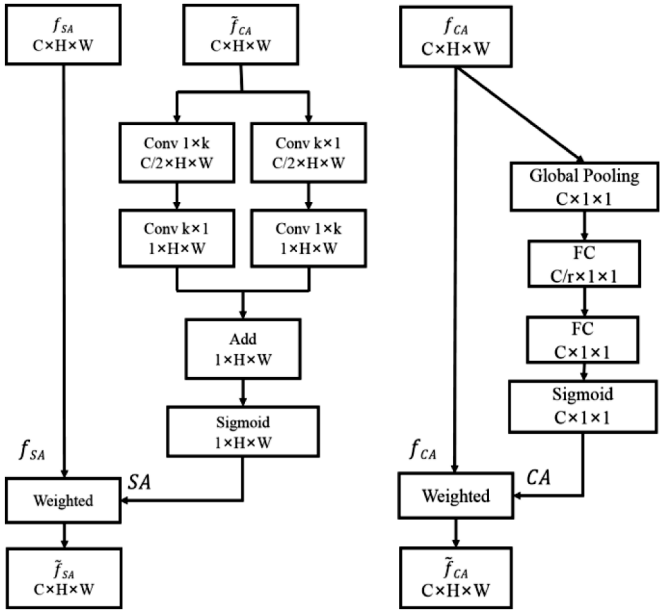

Fig. 2. Spatial attention network (left figure) and channel attention network (right figure).

- A multiscale perception module is proposed to improve the robustness of the network to object scale change and small objects. The internal branches of the module adopt dilated convolutions with different dilation rates to ensure that each branch of the receptive fields with different sizes can help the algorithm detect objects at different scales.
- An effective anchor box allocation strategy is proposed for the detection of safety helmets. This strategy can adaptively adjust the scale distribution of the anchor box for each layer based on the size of the feature map. It can achieve a finer division of the scale range for small objects. This helps detect small objects at the low-level feature maps.
- The experiment results show that our model can achieve mAP of 88.1% on the open-source helmet dataset GDUT-HWD [21], it is an increase of 12.8% compared to [21]. In order to verify the performance of the model on the general detection dataset, we also test it on the PASCAL VOC 2007. The mAP of our model is 80.5%, which is better than baseline by 3.4%.

3. Architecture of proposed network

Fig. 1 shows the network framework of proposed CASP detection algorithm, which is composed primarily of three parts: the cross-layer parallel attention network, multiscale perception module and a feature pyramid. Distinct from U-Net, CBAM [22] and other algorithms, CASP does not simply adopt the up-sampling and down-sampling, but uses different attention mechanisms for high-level and low-level features. Specifically, the characteristics of different feature levels allow using the channel attention to focus on high-level features and spatial attention to select effective low-level features. Spatial attention is not used for high-level features because high-level features contain highly abstract semantic information, there is no need to filter spatial information. There are almost no semantic differences between the channels of low-level features, thus there is also no need to use channel attention for low-level features. The channel attention network is primarily used to enhance the high-level semantic features. The spatial attention network is used to enhance the spatial structure features of object and retain its edge and other fine-grained information, which is conducive to detecting small objects. Multiscale perception module and feature pyramid are used to improve the robustness of the algorithm for object scale change. At the same time, the low-level feature map of the feature pyramid is also conducive to small object detection. In addition, unlike previous methods that use element-by-element addition or cascading multi-layer features, our model uses feature adaptive fusion methods in parallel attention networks and multiscale perception modules. The fusion of feature maps and the fusion weights at various scales are both obtained through network training.

3.1. Cross-layer parallel attention network

Cross-layer parallel attention network consists of the channel and spatial attention network, as shown in Fig. 1. For the channel attention network, given the input feature map X, the feature compression is firstly performed along the spatial dimensions using a compression operation, which converts each two-dimensional feature channel into a real number. This real number represents the global receptive field to some extent, and the dimensionality of the output matches the number of input feature channels. This represents the global distribution of the responses on the feature channel, it is allowed that the layers close to the input can also obtain global receptive fields. The weight is generated for each feature channel through the parameter W, which explicitly models the correlation between the feature channels through learning. The final step is the feature weighting operation, which weights the previous

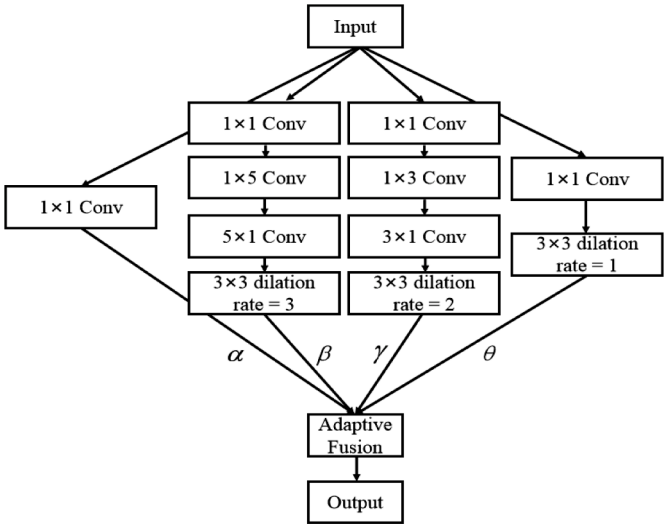

Fig. 3. Multiscale perception module and feature adaptive fusion.

features channel by channel via multiplication to complete the weighting of the original features in the channel dimension. The weight of the activation output reflects the importance of each feature channel, as shown on the right figure in Fig. 2.

Specifically, three high-level feature maps of Conv8_2, Conv9_2 and Conv10_2 are first fused. The fused feature map is expressed as $f_{CA} \in R^{C \times H \times W}$ and sent to the channel attention network. Global pooling is used to compress it to generate a *C*-dimensional feature vector. Two consecutive fully connected layers are then combined to form a bottleneck structure for modeling the correlation between channels. The action taken here is to firstly reduce the feature dimensionality to $1/r$ at the input, and then go back to the feature dimensionality at the input through a fully connected layer after using the relu activation layer. The advantages of doing this over direct use of a fully connected layer are (1) better nonlinearity, which can better fit the complex correlation between channels, (2) significant reduction of the number of parameters and calculations amount.

Then a sigmoid gate function is then used to map the values in the feature vector to the range of [0, 1]. Finally this feature vector is multiplied by the input feature map along the channel dimension for weighting. To describe these steps more intuitively, the above process is expressed mathematically as (1) and (2).

$$CA = F(v,\ W) = \sigma_1(fc_2(\delta(fc_1(v,\ W_1)),\ W_2)) \tag{1}$$

$$\widetilde{f}_{CA} = \ CA \cdot f_{CA} \tag{2}$$

where *W* represents the weight parameter that needs updating in the attention mechanism module,*v*is the *C*-dimensional feature vector, σ_1 is the sigmoid activation operation, *fc* is a fully connected layer, δ is the relu activation function, and $\widetilde{f}_{CA}$ is the weighted feature map.

For the spatial attention network, two low-level feature maps Conv4_3 and Conv7 are first fused, and the generated feature map is expressed as $f_{SA} \in R^{C \times H \times W}$, as shown on the left figure in Fig. 2. Similar to the method adopted by Zhao and Wu [23], feature map output from the channel attention network is first passed through two parallel asymmetric convolutional layers with kernels sized at $k \times 1$ and $1 \times k$. While the receptive field obtains more global information, it also reduces the network parameters and calculation amount. The outputs of the two asymmetric convolutional layers are added along the channel direction to generate a single-channel feature map. Finally, sigmoid gate function maps the feature values in the generated feature map to the range of [0, 1], the feature map is used to weight the input feature map f_{SA}. This process is expressed in mathematical form as (3)–(6).

$$C_1 = conv_2\left(conv_1\left(\widetilde{f}_{CA},\ W_1^{\ 1}\right),\ W_1^{\ 2}\right) \tag{3}$$

$$C_2 = conv_1\left(conv_2\left(\widetilde{f}_{CA},\ W_2^{\ 1}\right),\ W_2^{\ 2}\right) \tag{4}$$

$$SA = F(Y,\ W) = \sigma_2(C_1 + C_2) \tag{5}$$

$$\widetilde{f}_{SA} = SA \cdot f_{SA} \tag{6}$$

where *W* refers to the convolution kernel parameters, $conv_1$ and $conv_2$ refer to the convolution layers with kernel sizes of $k \times 1$ and $1 \times k$ respectively, $\widetilde{f}_{CA}$ is the output feature map of the channel attention network, $\sigma_1$is the sigmoid operation, and $\widetilde{f}_{SA}$ is the feature map

obtained by weighting the input feature map with the *SA*.

3.2. Multiscale perception module

The receptive field is an important concept for object detection, it refers to the area of a single pixel on the output feature map corresponding to the input layer. It is understood that the receptive field determines the number of pixels mapped to the original image in each pixel on the final feature map. If the receptive field of the output feature map is too small, there will not be enough elements corresponding to the original pixels in the feature map. This will cause the output feature map to contain less information and have poor detection results. Similarly, if the receptive field of the output feature map is too large, it is difficult to retain the fine-grained information, such as the spatial structure of small object, which can result in poor small object detection. Although the SSD uses high-resolution low-level feature maps to detect small objects and low-resolution high-level feature maps to detect large objects respectively, it does not consider that the high-resolution feature maps have too little semantic information, which does not improve the detection effect for small objects. Thus the proper design of a suitable network receptive field is important.

Similar to the ASPP [24], RFBNet, TridentNet and other networks, a multiscale perception module is used to improve the detection effect for small objects, as shown in Fig. 3. This module samples the given input feature map with dilated convolutions of different dilation rates, which is equivalent to capturing the context information of an image with multiple scales. Distinct from the above algorithms, such as ASPP, the proposed multiscale perception module adopts the feature adaptive fusion method to combine the feature maps with different receptive fields, it can effectively fuse the features of different layers based on their importance by learning the weight parameters.

Specifically, the fusion feature map obtained from the attention network is used as an input of the multiscale perception module. Each branch firstly reduces the dimensionality of the input feature through a 1×1 convolution layer. The middle two branches then pass through 3×3 and 5×5 convolutional layers. To reduce the network parameters and calculation amount, the cascaded convolutional layer pairs at 1×3 and 3×1 along with 1×5 and 5×1 are used instead of 3×3 and 5×5 convolutional layers respectively. Feature maps with different receptive field sizes are then generated from the dilated convolutions with dilation rates of 1, 2 and 3. Finally, features of each branch are fused by using feature adaptive fusion.

3.3. Feature adaptive fusion

In existing object detection algorithms, making full use of the semantic information of high-level features and the fine-grained features of the bottom layer requires adopting more feature fusion methods, such as the FPN architecture. These are generally direct cascading feature layer or adding the corresponding elements, such as YOLO-v3 and RetinaNet. However, feature fusion of FPN and other similar structures may cause scale inconsistency between feature maps with different receptive field sizes. Thus a novel feature adaptive fusion method is proposed here. The main idea is to determine the weight parameters of the feature fusion for different layers by network learning. Previous methods are directly cascading the different feature layers without any differentiation, or the corresponding elements are added directly. However, advantage of the proposed method is that the important features will automatically be given more weight to effectively integrate the features of different layers based on their importance.

Taking the feature map in Fig. 3 as an example, when the output feature maps $X^l (l = 1, 2, 3, 4)$ for different receptive field sizes are fused, the features from different branches can be multiplied by their corresponding weights, α_{ij}, β_{ij}, γ_{ij}, θ_{ij}, and summed to finally generate the fusion feature map Y_{ij}. As shown in (7) and (8), ij in the lower right corner represents the coordinate (i, j) on the feature map. In this example, as the size of the feature map and the number of channels to be fused are the same, no deconvolution operation is required. If feature maps with different sizes are fused, those should firstly be adjusted to be the same size by using the interpolation or deconvolution before fusion. If the number of channels in the feature map to be fused is different, this can be made consistent through a convolutional layer with a 1×1 kernel.

$$Y_{ij} = \alpha_{ij} \cdot {X^1}_{ij} + \beta_{ij} \cdot {X^2}_{ij} + \gamma_{ij} \cdot {X^3}_{ij} + \theta_{ij} \cdot {X^4}_{ij} \tag{7}$$

$$\alpha_{ij} = \frac{e^{\lambda_{\alpha_{ij}}}}{e^{\lambda_{\alpha_{ij}}} + e^{\lambda_{\beta_{ij}}} + e^{\lambda_{\gamma_{ij}}} + e^{\lambda_{\theta_{ij}}}} \tag{8}$$

Softmax function can make α_{ij}, β_{ij}, γ_{ij} and θ_{ij} in the range of [0,1] with $\alpha_{ij} + \beta_{ij} + \gamma_{ij} + \theta_{ij} = 1$. λ_α, λ_β, λ_γ and λ_θ are single-channel feature maps generated by performing 1×1 convolutional dimension reduction on the input feature maps $X^l (l = 1, 2, 3, 4)$. The network can learn the weights via gradient back propagation.

3.4. Anchor box allocation strategy

The conventional SSD algorithm uses (9) to assign anchor boxes of different scales to different feature layers. The basic concept is to assign small-scale anchor boxes to low-level feature maps and large-scale anchor boxes to high-level feature maps. Anchor boxes are assigned at the same interval between levels of the feature maps. However, when there are many small- and medium-sized objects in the dataset, this anchor box allocation strategy is not suitable. This is because the anchor box scale allocation at the same interval between the feature maps may cause the feature layers to ignore objects of certain scales. In addition, the top-level feature maps may not detect objects because there are either no or fewer large objects that can be matched with large-scale anchor boxes, and

Table 1
Distribution of anchor box sizes when using (9) and (10) respectively.

Layer Index	Scale	S_l	$\overline{S}_l$
1	Conv4_3: 38 × 38	0.10	0.10
2	Conv7: 19 × 19	0.26	0.13
3	Conv8_2: 10 × 10	0.42	0.21
4	Conv9_2: 5 × 5	0.58	0.35
5	Conv10_2: 3 × 3	0.74	0.58
6	Conv11_2: 1 × 1	0.90	0.90

Table 2
Comparison of the quantitative experiment results on the GDUT-HWD testing set, where bold numbers indicate the highest mAP in each column. Helmet and Warning columns indicate that the individual wears a safety helmet or not respectively.

Method	Input Size	Backbone	mAP	Helmet	Warning
Faster R-CNN	300 × 500	VGG16	67.32	70.20	63.72
ResNet + FPN	384 × 640	ResNet-50	74.21	76.31	69.26
SSD	300 × 300	VGG16	72.46	77.48	66.57
DSSD	321 × 321	Residual-101	76.85	80.12	72.48
RFBNet	300 × 300	VGG16	77.68	81.75	73.62
YOLOv3	416 × 416	Darknet-53	78.37	81.26	76.17
CASP (Ours)	300 × 300	VGG16	**88.11**	**89.62**	**86.85**

unnecessary network parameters are added.

The safety helmet detection dataset GDUT-HWD contains a large number of small and medium objects (about 90% of the total dataset). To more effectively detect small objects like safety helmets, (10) is proposed to assign anchor boxes to different feature layers at different scale intervals. Specifically, from the low- to high-level feature map, the scale intervals of the anchor box of each layer are from small to large. Thus the range of the small-scale object can be more finely divided, which helps detect small objects of various scales in the low-level feature map.

$$S_l = S_{min} + \frac{S_{max} - S_{min}}{m-1}(l-1),\ l \in [1,\ m] \tag{9}$$

$$\overline{S}_l = S_{min} + \frac{W_{m+1-l}}{W_{min}} \frac{S_{max} - S_{min}}{W_{max}},\ l \in [1,\ m] \tag{10}$$

where $S_{min} = 0.1$, $S_{max} = 0.9$, m = 6, l is the index of the feature layer, W_{min} and W_{max} are the scale of the smallest and largest feature maps respectively, W_{m+1-l} is the scale of the $m + 1 - l$ layer feature map used for detection. S_l and $\overline{S}_l$ are the anchor box size distributions obtained using (9) and (10) respectively. Table 1 shows that the intervals of the former are the same, they are 0.16. While the intervals between anchor box scales of the latter are from small to large, it changes with the scale of the feature map to better cover the scale range of small objects. This helps match the anchor box to more small objects and improve its recall rate.

4. Experiments

4.1. Experiment introduction and parameter setting

4.1.1. Datasets and evaluation indicators

The network model is evaluated on the open-source safety helmet dataset GDUT-HWD, which has a total of 3174 images. The training set and the testing set include 1587 images respectively, which contain a total of 18,893 objects. All object instances are divided into either blue, white, yellow and red based on the color of the helmet or none category, indicating the individual is not wearing a helmet. According to the definition of the COCO dataset, the numbers of corresponding small-, medium- and large- objects in the GDUT-HWD are 8950, 7924 and 2019 respectively. It is seen that the number of small- and medium-scale objects in the dataset account for approximately 90% of the total objects, this indicates it is a challenge to detect safety helmets. In practice, it is of primary concern to detect whether a person is wearing a helmet. Thus four color categories of helmet are combined into one category, the dataset is re-divided into the two categories (total object number is unchanged) of Helmet and Warning. As a fair comparison with state-of-the-art methods, model evaluation is also performed on the public dataset PASCAL VOC 2007, which contains 9963 images divided into training, validation and testing sets. A total of 5011 images are used as the training set. The standard evaluation indicators in the PASCAL Challenge Protocol are used to evaluate the model, they are Average Precision (AP) and mAP.

4.1.2. Experimental details

The pre-trained VGG-16 model is used to initialize the weight parameters of the underlying shared convolutional layer. The other newly added network layer parameters are initialized by using a Gaussian distribution with a zero mean and variance of 0.01. The

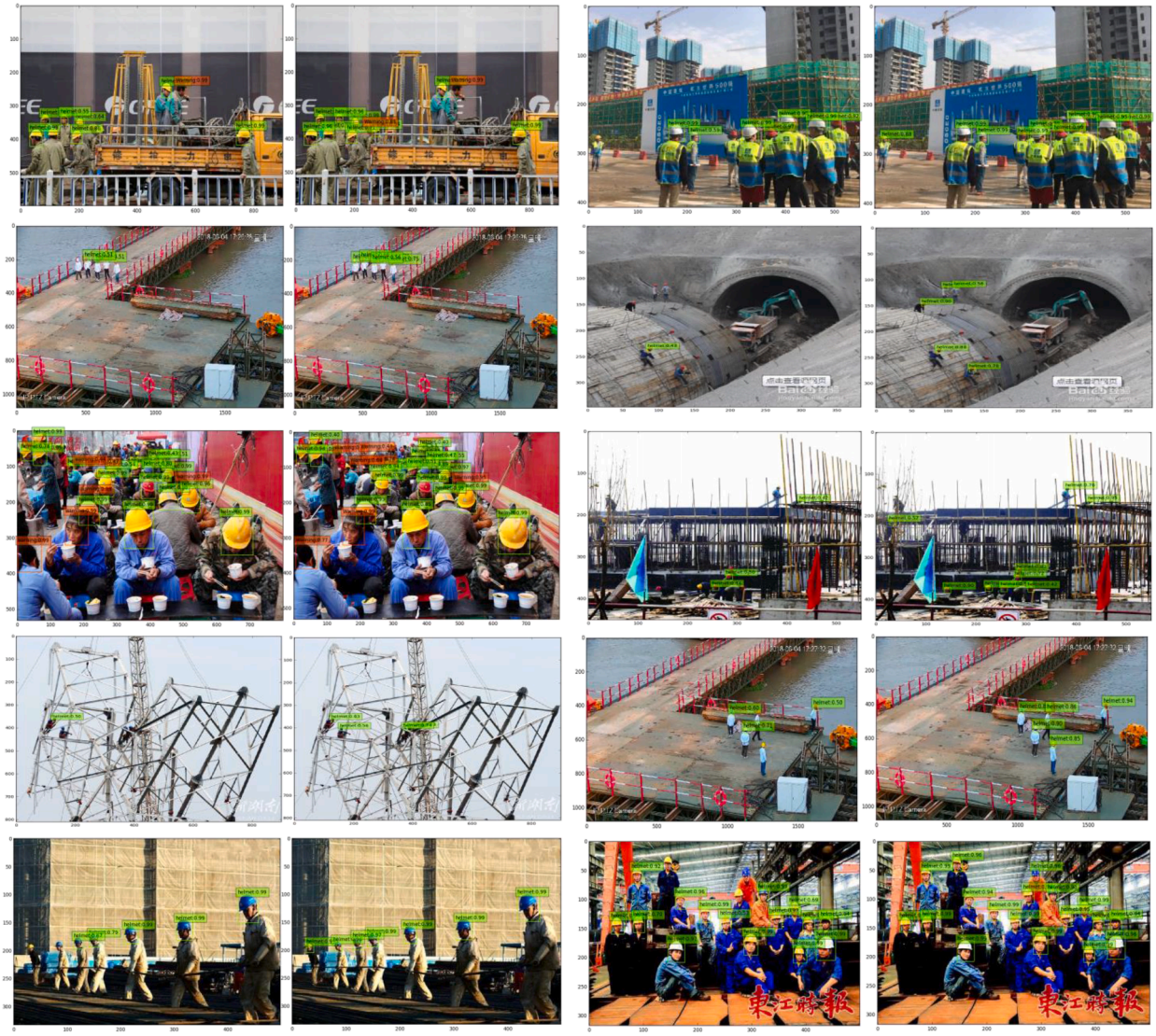

Fig. 4. Comparison of the qualitative experiment results of SSD and the CASP proposed in this paper on the GDUT-HWD testing set. (a) and (b) are the detection results of SSD. (a') and (b') are the detection results of CASP.

image batch size used during the training is 32 and the initial learning rate is 0.004. All other training parameter setting is consistent with the SSD algorithm. The model is trained with a total of 250 epochs, a learning rate warm-up is set to 5 epochs. The learning rate is decreased to 0.0004 and 0.00004 at 180 and 220 epochs respectively. The hyperparameter r used for dimensionality reduction in the channel attention is set to 4, the hyperparameter k for the asymmetric convolution kernel in the spatial attention is set to 5. The model is trained on two NVIDIA GeForce GTX 1080Ti graphics cards.

4.2. Performance results on GDUT-HWD

4.2.1. Performance comparisons

Table 2 gives the comparison of the quantitative experiment results on the GDUT-HWD dataset. The proposed CASP detection algorithm and the Faster R-CNN, ResNet + FPN, SSD, DSSD, RFBNet, YOLOv3 are compared here. The mAP of CASP can achieve 88.11%, which is an improvement compared with the other detection methods. Compared with the SSD and Faster R-CNN, the mAP is improved by 15.65% and 20.79% respectively, it is also increased by 9.74% compare with YOLO v3. For object detection, small objects and scale change both are challenging attributes. YOLO v3, SSD, Faster R-CNN and other algorithms that perform well on general object detection datasets have difficulty in achieving a high detection accuracy for the dataset of small- and medium-scale objects. The selective combination of attention mechanism, multi-scale perception module, feature pyramid and other improved technologies can

Table 3
Ablation experiment results on the GDUT-HWD testing set, where bold numbers indicate the highest mAP, "√" in each column indicates that the leftmost component is used in the model, and the last number in each column represents the mAP obtained using the corresponding component.

Component						
Feature Pyramid Network	Baseline	√	√	√	√	√
Cross-layer Parallel Attention Network			√	√	√	√
Anchor Assignment				√	√	√
Multiscale Perception Module					√	√
Feature Adaptive Fusion						√
GDUT-HWD mAP	72.46	77.13	80.66	85.37	87.37	**88.11**

effectively improve the accuracy of the helmet wearing detection algorithm. To ensure real-time detection, the input image of the network in the experiment uses a smaller resolution (300 × 300 pixels) to reduce the network parameters and improve the detection speed. If the input image is replaced with the larger resolution image (512 × 512 pixels), the detection accuracy is improved to a certain extent but the processing speed will slow down. Fig. 4 gives the comparison of the qualitative experiment results of the proposed CASP and the basic SSD on the GDUT-HWD testing set. It is seen that the proposed algorithm has the much better detection ability than the SSD for small objects at longer distances, which further validates the effectiveness of our algorithm.

4.2.2. Ablation studies

The effect of different components of the proposed algorithm is further studied in the ablation experiments. As a fair comparison, except for the parameters of the newly added module, the remaining parameters, such as the input image size and related hyper-parameters, are all the same in the ablation experiment. As shown in Table 3, the mAP is only 72.46%, when original base algorithm (baseline) only uses hierarchical prediction without any feature fusion. The low-level features retain fine-grained information that is helpful to detect small objects but lack semantic information for classification. Thus small objects detected at the low-level features may also cause classification errors due to insufficient semantic information.

Similarly, high-level features retain good semantic information, but most of the fine-grained information is missing. This makes it difficult for high-level features to detect small objects, it is prone to miss detection. After the low- and high-level features merge, the generated feature map contains both certain semantic information and some fine-grained information. A feature pyramid is constructed from the feature map along the bottom to the top, and predictions are performed on each layer. As a result, the mAP is raised to 77.13% compare with baseline, it shows that the feature pyramid is an effective way to solve the scale change and small object detection. To extract more effective high- and low-level features, the mAP is raised to 80.66% by using the cross-layer parallel attention network, which are used to refine the spatial structure features and weight the semantic information in the channel with different weights.

The anchor allocation is one idea to address the scale change and small object detection. The original anchor scale allocation at each layer is readjusted by using the proposed Eq. (10), which helps the mAP raise to 85.37%. Based on RFBNet, we know that different receptive field sizes also have significant influence on the detection effect of objects with different scales. The larger the receptive field is, the better the detection effect of large objects is. However, this does not help better detect small objects. Conversely, a smaller receptive is helpful to detect small objects, but it is not conducive to the detection of large objects. Thus the size of the receptive field is proportional to the scale of object to be detected. A novel scale-aware module is adopted here to help the network reduce its sensitivity to objects with various scales. This module uses parallel branches with different receptive field sizes to solve the above problems. At the same time, the high-level feature map of the pyramid becomes increasingly smaller, making it difficult to match the sizes of different convolution kernels in the scale perception module. Therefore, the proposed module is only applied in the bottom two layers of pyramid with the larger resolution. This obtains some improvement and help the mAP raise to 87.37%.

Finally, in order to more effectively fuse features, a novel feature adaptive fusion method is used to perform the fusion of feature layers based on their importance. The mAP of the best model trained on the GDUT-HWD is 88.11%. The results of the ablation experiments show that the poor robustness is caused by unsatisfactory detection effect of small objects and scale change. The combination of different modules, the use of relevant optimization methods, such as anchor box scale allocation and feature adaptive fusion, all of these help reduce the sensitivity to object scale change and improve the detection performance of small objects.

4.3. Performance results on PASCAL VOC 2007

4.3.1. Performance comparisons

Table 5 gives the comparison of the quantitative experiment results on the VOC 2007 dataset. The proposed CASP algorithm and recently published detection algorithms (RFBNet, STDN [25] and DSSD) are compared here. The models in the experiments are trained on the VOC 2007 and the VOC 2012 training sets, they are tested on the VOC 2007 testing set. For a fair comparison, all the models in the experiments use the same hyperparameters.

It is seen from Table 5 that the mAP of CASP algorithm on 20 categories is 80.5%, which indicates better detection results than other state-of-the-art algorithms. The mAP of CASP is improved by 7.3% and 2.0% compared with the two-stage detection algorithm Faster R-CNN and R-FCN respectively. The mAP of CASP is improved by 3.4%, 2.2% and 1.2% compared with one-stage detection algorithm SSD300, DSSD300 and STDN320 respectively. In addition, both the one- and two-stage algorithms have poor detection effect on small

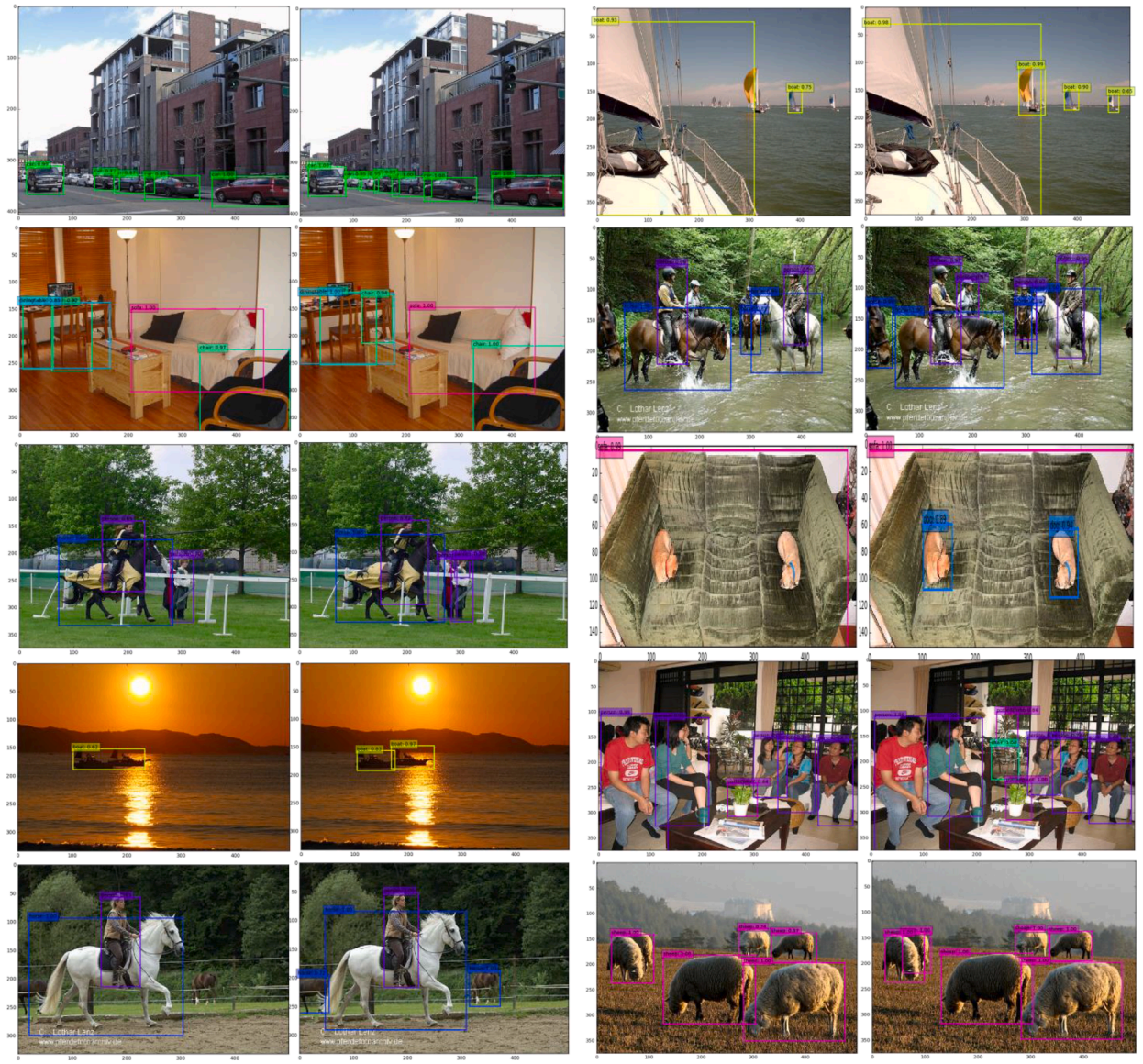

Fig. 5. Comparison of qualitative experimental results of SSD and the CASP algorithm proposed in this paper on the VOC 2007 testing set. (a) and (b) are the detection results of SSD. (a') and (b') are the detection results of CASP.

Table 4
Ablation experiment results on the VOC 2007 testing set, where bold numbers indicate the highest mAP, "√" in each column indicates that the leftmost component is used in the model, and " × " indicates that the leftmost component is removed. The last number in each column represents the mAP obtained using the corresponding component.

Component						
Feature Pyramid Network	Baseline	√	√	√	√	√
Cross-layer Parallel Attention Network			√	√	√	√
Anchor Assignment				√	×	×
Multiscale Perception Module					√	√
Feature Adaptive Fusion						√
GDUT-HWD mAP	77.1	78.3	79.5	78.4	80.3	**80.5**

Table 5
Comparison of the quantitative experiment results on the VOC 2007 testing set, where bold numbers indicate the highest mAP in each column.

Method	*mAP*	*aero*	*bike*	*bird*	*boat*	*bottle*	*bus*	*car*	*cat*	*chair*	*cow*	*table*	*dog*	*horse*	*mbike*	*Person*	*plant*	*sheep*	*sofa*	*train*	*tv*
Faster R-CNN	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
OHEM	74.6	77.7	81.2	74.1	64.2	50.2	86.2	83.8	88.1	55.2	80.9	73.8	85.1	82.6	77.8	74.9	43.7	76.1	74.2	82.3	79.6
R-FCN	79.5	79.9	87.2	78.5	72.0	65.8	86.8	87.1	**89.8**	**67.0**	86.1	74.5	**87.8**	**90.6**	79.9	81.2	53.7	**81.8**	80.5	85.9	79.9
SSD300	77.1	79.5	83.9	76.0	69.6	50.5	87.0	85.7	88.1	60.3	81.5	77.0	86.1	87.5	83.9	79.4	52.3	77.9	79.5	87.6	76.8
DSSD300	78.3	81.9	84.9	**80.5**	68.4	53.9	85.6	86.2	88.9	61.1	83.5	78.7	86.7	88.7	86.7	79.7	51.7	78.0	**80.9**	87.2	79.4
RFBNet300	80.1	83.5	86.7	77.8	73.9	61.8	87.8	**87.5**	87.9	64.8	86.8	75.2	85.4	87.6	86.5	81.6	56.8	80.9	78.9	88.3	80.5
STDN320	79.3	81.2	**88.3**	78.1	72.2	54.3	87.6	86.5	88.8	63.5	83.2	**79.4**	86.1	89.3	87.0	77.3	52.5	80.3	80.8	86.3	**82.1**
CASP(Ours)	**80.5**	**84.6**	88.1	79.5	**75.8**	**66.6**	**88.3**	86.8	88.1	64.7	**87.3**	77.4	86.6	89.1	**87.5**	**81.7**	**57.2**	78.6	79.6	**87.7**	79.7

objects. For example, the mAPs of bottle and plant are 52.1% and 38.8% for Faster R-CNN, while those for SSD are 50.5% and 52.3% respectively. The proposed CASP can achieve good detection effect for these two types of small objects, the mAP is raised to 66.6% and 57.2% respectively. It is verified that the CASP can detect small-scale objects well. In addition, CASP also can achieve the best mAP in other types of object detection, such as aero, persons and buses. This also validates that the proposed CASP not only can obtain improved results in special detection tasks, such as wearing a helmet, but also show good generalization performance for general object detection tasks, especially for small- and medium-scale objects. Fig. 5 gives the comparison of the qualitative experiment results of the proposed CASP and the basic SSD on the VOC 2007 testing set. It is seen that the detection effect of CASP is better, especially for small objects, which further verifies its effectiveness.

4.3.2. Ablation studies

Ablation experiments are conducted on the VOC 2007 dataset to test whether each component of the proposed algorithm is effective when performing general object detection tasks, as shown in Table 4. The mAP is raised to 78.3% by using reconstructed feature pyramid. Then the mAP is raised to 79.5% by using the cross-layer parallel channel and spatial attention networks. However, when introducing the anchor box adaptive scale allocation strategy proposed in this paper, the mAP is dropped to 78.4%. Although this is greater than the baseline, it is not optimal. This could be because the proportion of small- and medium-scale objects in the GDUT-HWD dataset is relatively large, it is necessary to design an anchor box size distribution strategy that is more conducive to detect small objects, which can improve the match of the anchor box and ground truth to increase recall rate. However, the proportion of small and medium objects in the VOC 2007 dataset is not large. If the anchor box scale allocation strategy designed for small objects is still used, although the recall rate of small objects will be improved, it will have a negative impact on the detection of large-scale objects to a certain extent. Therefore, the adaptive scale allocation strategy is removed during the subsequent ablation experiments, the anchor box allocation strategy of SSD is used instead. After multiscale perception module and feature adaptive fusion is introduced, the mAP of our algorithm is raised to 80.5%. The ablation experiment results indicate that the components used here can still improve the detection effect when performing general object detection tasks.

5. Conclusion

In this paper, we propose a novel detection network model that combines the attention mechanism and the multiscale perception module. The cross-layer parallel attention network proposed in this paper is used to further refine the fine-grained information of the low-level fusion features and enhance the semantic information of the high-level fusion features. To better improve the robustness of the proposed algorithm, a multiscale perception module is used to increase the detection performance of the object with different scales. This module uses the feature adaptive fusion method to fuse feature maps with different receptive field sizes. In addition, in order to improve the detection accuracy of helmets, an effective anchor box scale allocation strategy is designed. Finally, the proposed algorithm is evaluated on the general object detection dataset PASCAL VOC 2007 and the helmet-wearing detection dataset GDUT-HWD. The mAP of our network model on these two datasets is 80.5% and 88.1% respectively, which reach state-of-the-art detection accuracy. In the future, we plan to integrate the object detection with the object tracking and pedestrian re-recognition, at the same time carry out lightweight research (such as pruning the network) to promote its development.

Funding

This work was supported by the Natural Science Foundation of China NSFC under Grants 61871445, 61302156; Key R & D Foundation Project of Jiangsu Province under Grant BE2016001–4; Public Welfare Technology Application Research Plan Project of Zhejiang Province under Grant LGG18F030002.

Declaration of Competing Interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled, “Method based on the cross-layer attention mechanism and multiscale perception for safety helmet-wearing detection”.

References

[1] Park M, Brilakis I. Construction worker detection in video boxes for initializing vision trackers. Autom Constr 2012;28(15):15–25. vol.no.pp.
[2] Rubaiyat AHM, Toma TT, Kalantari-Khandani M, Rahman SA, Chen L, Ye Y, Pan CS. Automatic detection of helmet uses for construction safety. In: Proceedings of the IEEE/WIC/ACM international conference on web intelligence workshops; 2016. p. 135–42. pp.
[3] Du S, Shehata M, Badawy W. Hard hat detection in video sequences based on face features, motion and color information. In: Proceedings of the 3rd international conference on computer research and development. IEEE; 2011. p. 25–9. pp.
[4] Cai L, Qian J. A method for detecting miners based on helmets detection in underground coal mine videos. Min Sci Technol 2011;21(4):553–6. Chinavolnopp.
[5] Ren S, He K, Girshick R, Jian S. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2015;39:1137–49. volvol.6pp.
[6] Dai J, Li Y, He K, Sun J. R-fcn: object detection via region-based fully convolutional networks. In: Proceedings of the advances in neural information processing systems; 2016. p. 379–87. pp.

[7] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C, Berg A. Ssd: single shot multibox detector. In: Proceedings of the European conference on computer vision; 2016. p. 21–37. pp.
[8] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," Tech Report, Computer Vision and Pattern Recognition (cs.CV), arXiv:1804.02767 [cs.CV], 2018.
[9] Lin T, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: common objects in context. In: Proceedings of the European conference on computer vision; 2014. p. 740–55. pp.
[10] Zhang K, Zhang Z, Li Z, Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 2016;23(10): 1499–503. volnopp.
[11] Singh B, Davis LS. An analysis of scale invariance in object detection snip. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 3578–87. pp.
[12] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2117–25. pp.
[13] C.Y. Fu, W. Liu, A. Ranga, A. Tyagi, A.C. Berg, "DSSD: deconvolutional single shot detector," Computer Vision and Pattern Recognition (cs.CV), arXiv: 1701.06659 [cs.CV], 2017.
[14] Zhang S, Wen L, Bian X, Lei Z, Li S. Single-shot refinement neural network for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 4203–12. pp.
[15] Kisantal M, Wojna Z, Murawski J, Naruniec J, Cho K. Augmentation for small object detection. In: Proceedings of the 9th international conference on advances in computing and information technology; 2019.
[16] Wang C, Xu C, Wang C, Tao D. Perceptual adversarial networks for image-to-image transformation. IEEE Trans Image Process 2018;27(8):4066–79. volnopp.
[17] Zhang S, Zhu X, Lei Z Z, Shi H, Wang X, Li S. Faceboxes: a CPU real-time face detector with high accuracy. In: Proceedings of the IEEE international joint conference on biometrics; 2017. p. 1–9. pp.
[18] Liu S, Huang D. Receptive field block net for accurate and fast object detection. In: Proceedings of the European conference on computer vision; 2018. p. 385–400. pp.
[19] Li Y, Chen Y, Wang N, Zhang Z. Scale-aware trident networks for object detection. In: Proceedings of the IEEE International conference on computer vision; 2019. p. 6054–63. pp.
[20] He K, Gkioxari G, Dollar P, Girshick R. Mask r-cnn. In: Proceedings of the international conference on computer vision; 2017. p. 2980–8. pp.
[21] Wu J, Cai N, Chen W, Wang H, Wang G. Automatic detection of hardhats worn by construction personnel: a deep learning approach and benchmark dataset. Autom Constr 2019;106:102894. volpp–.
[22] Woo S, Park J, Lee YJ, Kweon SI. Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision; 2018. p. 3–19. pp.
[23] Zhao T, Wu X. Pyramid feature attention network for saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2019. p. 3085–94. pp.
[24] Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 2017;40(4):834–48. volnopp.
[25] Zhou P, Ni B, Geng C, Hu J, Xu Y. Scale-transferrable object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 528–37. pp.

Guang Han received the B.S. degree from Shandong University of Technology in 2004, and M.S. and Ph.D. degrees from Nanjing University of Science and Technology, in 2006 and 2010, respectively. Since 2010, he has been with Nanjing University of Posts and Telecommunications. His-current research interests include pattern recognition, video analysis, computer vision and machine learning.

Mengcheng Zhu received the B.S. degree from Chaohu University in 2018. Currently, he is pursuing the M.S. degree in Nanjing University of Posts and Telecommunications. His-current research interests include computer vision and deep learning.

Xuechen Zhao is pursuing the B.S. degree in Nanjing University of Posts and Telecommunications. Her current research interests include single object tracking based on deep learning, video analysis and computer vision.

Hua Gao received the B.S. degree from Weihai Department of Shandong University in 2005, then received M.S. and Ph.D. degrees from Nanjing University of Science and Technology, in 2007 and 2013, respectively. Since 2013, he has been with Zhejiang University of Technologies. His-current research interests include artificial intelligence, pattern recognition and image retrieval.