Higher Accuracy Yolov5 Based Safety Helmet Detection

Zizhen Wang(✉), Yuegong Sun, Zhening Wang, and Ao Li

School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China

wangjackson2022@163.com

Abstract. For the construction site with high-risk possibility, object detection based on safety helmet and reflective clothing will greatly reduce the risk of workers. At present, the algorithm based on deep learning is the mainstream algorithm of object detection. Among them, the YOLO algorithm is fast and widely used in real-time safety helmet detection. However, for the problems of small objects such as safety helmets and relatively dense detection scene objects, the detection effect is not ideal. For these problems, this paper proposes an improvement of the safety helmet detection algorithm based on YOLOv5s. The DenseBlock module is used in the improved algorithm to replace the Focus structure in the backbone network, which has an improved feature extraction capability for the network; secondly, Soft-NMS is used to retain more category frames when removing redundant frames. After the experiments, it is shown that the accuracy is improved on the homemade safety helmet dataset, which indicates the effectiveness of the improved algorithm.

Keywords: Deep Learning · Object Detection · YOLOv5

1 Introduction

On the construction site, safety helmet can be regarded as a necessary product for workers to protect their lives, which for the fall of objects from height to produce a certain buffer effect, can reduce the damage caused by the accident. Traditional safety helmet detection on construction sites adopt the manual supervision, but this approach is likely to fail to monitor all workers at the construction site, which can easily cause errors, and also cannot monitor the staff at all times.

To solve these problems, the algorithm based on deep learning have become the mainstream approach. Such detection algorithms, which can be deployed to mobile terminal is convenient and accurate, significantly reducing costs and being more efficient than manual monitoring, and it is also more important to develop highly accurate and high-performance detection models [1].

There are usually two types of object detection. One of the algorithms usually includes R-CNN, Fast R-CNN, etc. These algorithms first generate candidate frames

A. Li et al. (Eds.): 6GN 2022, LNICST 504, pp. 89–99, 2023.
https://doi.org/10.1007/978-3-031-36011-4_8

through a specialized module and then further classify them through the generated candidate frames. R-CNN [2], the earliest of this series of algorithms, also reveals many problems. Fast R-CNN [3] was proposed to address these problems, which reduces the large amount of redundancy in feature extraction, speeds up the algorithm, and the end of the grid uses different fully connected layers to achieve better results that output classification results and window regression results simultaneously, enabling end-to-end multi-task training and eliminating the need for additional storage. The RPN candidate frame extraction module of the generation network further improves the speed of the algorithm.

YOLO [5] was proposed as a classical single-stage detection algorithm by Redmon J et al. The most important feature of YOLOv1 is that it uses only one convolutional neural network to implement object detection end-to-end, and its algorithm is to divide the image into multiple grids and predict two bounding boxes for each grid, which differs from R-CNN to avoid the operation of completing the detection task in two steps and improves the detection speed. Later, YOLOv2 [6] was proposed, which introduced anchor frames based on YOLOv1, increasing the number of grids and prediction frames each grid is responsible for. YOLOv3 [7] uses a prediction head that produces three scales of output, with high scale predicting small objects and low scale predicting large objects, alleviating the loss of small, medium and large objects in YOLOv1 and YOLOv2. YOLOv4 [8] uses a large number of tuning techniques based on YOLOv3 to make the model more accurate, however, YOLOv4 has a large model size and slower inference speed. YOLOv5 uses a backbone network with adjustable model size to enable the algorithm to achieve better detection results while maintaining detection speed [9].

At present, many scholars have carried out relevant research on helmet detection, most of which adopt the method of deep learning. For example, Sun et al. added a self-attention layer to the two-stage object detection algorithm Faster R-CNN to extract multi-scale global information of the object and enhanced the training of small objects by improving the anchor frame. The final improved algorithm has better robustness for helmet detection in various general scenarios. Kai Xu et al. improved the detection of small objects by adding feature maps to the single-stage object detection algorithm YOLOv3, and then used K-means clustering to select suitable anchor frames, after which GIOU Loss was used instead of IOU Loss to calculate the border loss, and Focal Loss was added to balance the positive and negative samples. Cheng Rao et al. Used the depth separable convolution and residual (SR) module to replace the original convolution layer of yolov3 tiny algorithm, reducing the amount of parameters and calculation, and improved the spatial pyramid pooling (SPP) module that extract more features, and finally introduce CIOULoss as the border loss function to improve the regression accuracy. Improved the regression accuracy and significantly improved all metrics [10]. Zhang Jin et al. Added a multispectral attention module in the neck of yolov5 network, which improved the generalization ability of the model. Xu Chuanyun et al. proposed a scene enhancement-based data augmentation algorithm based on the YOLOv4 algorithm to improve the performance of the model in detecting small objects.

In this paper, two types of object, whether the construction workers wear safety helmets or not, are taken as detection tasks, and more than 6000 pictures are collected from the open-source safety helmet data set for preprocessing, and the data set of this

experiment is constructed. In this paper, yolov5s model is selected for training, and an improved method is proposed for the detection of safety helmet based on yolov5s. Firstly, using the DenseBlock module to improve the Focus structure in the backbone network to be able to extract features better, and secondly, improving YOLOv5s algorithm for removing redundant frames and using Soft-NMS to solve the problem that the NMS algorithm retains only the highest confidence prediction frames of its kind when objects overlap in height The accuracy of safety helmet detection is further improved by fusing the two improvements. The experimental results show that the mAP (Mean Average Precision) of the improved YOLOv5 algorithm is significantly improved and can meet the requirements of detection in construction scenarios.

2 Related Work

2.1 YOLOv5s

Yolov5 provides four object detection networks, including yolov5s, yolov5m, yolov5l and yolov5x. Among these networks, yolov5s network has the smallest depth and the smallest width of the feature map. The other three types are continuously deepened and widened on this basis.The network structure of YOLOv5s is shown in Fig. 1.

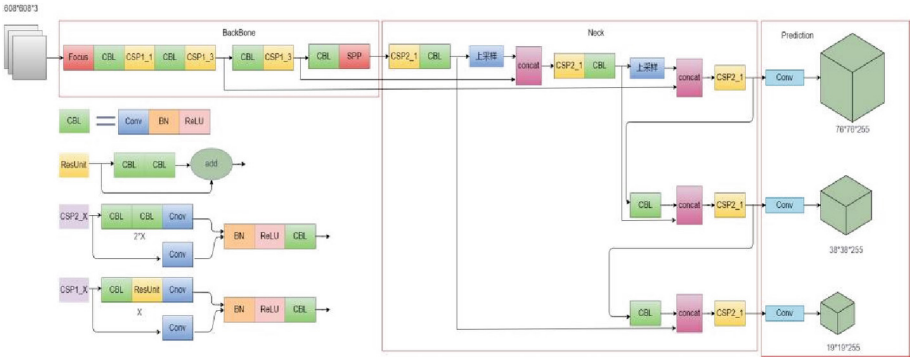

Fig. 1. Network structure of YOLOv5s

YOLOv5s network mainly consists of Backbone, Neck, and Head. Yolov5s uses mosaic data enhancement at the input end, which is the same as YOLOv4, adaptive anchor frame calculation, and self-use picture scaling; Focus structure and CSP (Cross Stage Partial network) are used at Backbone, which was not available in the previous generations of YOLOv5; The neck is a feature fusion network with a combined top-down and bottom-up feature fusion approach, which better fuses multi-scale features [11]. The current neck of YOLOv5 adopts the FPN + PAN structure as in yolov4. However, when yolov5 first came out, only the FPN structure was used, and the PAN structure was added later. In the post-processing of object detection, NMS operation is required for the screening of object frames. In yolov5s, weighted NMS is adopted, which has some

improvements for some occluded overlapping objects. Compared with yolov4, the focus structure is added in the backbone network of yolov5, which is mainly used for slicing.

For feature extraction, Yolov5 will have a size of $3 \times 608 \times 608$ general image input network, converted to a feature map of size $12 \times 304 \times 304$ by a Focus slicing operation, and then converted to a feature map of size $32 \times 304 \times 304$ by a normal convolution operation with 32 convolution kernels. YOLOv5 algorithm designs two new CSP structures, which are different from yolov4, which only uses CSP structure in the backbone network. As can be seen from the yolov5s network structure above, the backbone network adopts CSP1_1 structure and CSP1_3 structure, CSP2_1 for neck structure to strengthen the feature fusion between networks.

3 The Proposed Method

3.1 DenseBlock

DenseBlock is an important part of the DenseNet network [12], and its main idea is that for each layer, the feature mappings of all the previous layers are used as the input of the current layer, while their own feature mappings are the input of the subsequent layers, forming a full mutual link. The feature mappings extracted from each layer are available for the subsequent layers. The advantages are that it can enhance the feature propagation, alleviate gradient disappearance and reduce the number of parameters. The structure diagram shown in Fig. 2.

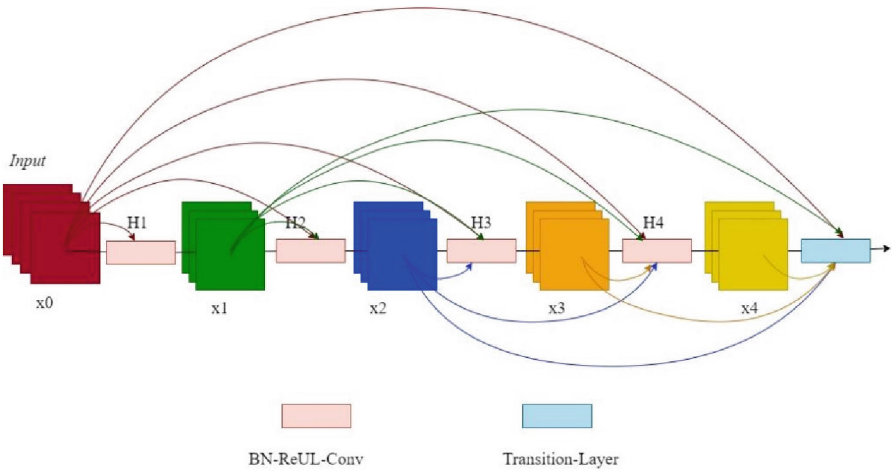

Fig. 2. Structure diagram of Densenet

Many deep learning networks choose to use ResNet, but DenseNet's dense connection mechanism is more prominent: as can be seen from the above structural diagram, each layer will have the input of all the previous layers, that is, the layers are connected and used as the input of the next layer. Assuming that there is an L-layer network,

DenseNet contains $L(L+1)/2$ connections, which will form a more dense connection. The DenseNet network is a strategy of outputting feature maps for stitching before performing nonlinear transformations, i.e., doing stacking between channels, rather than summing pairs of values [13, 14], moreover, DenseNet directly connects feature maps from different layers, which allows features to be reused and improves efficiency. This feature is DenseNet's greatest advantage over ResNet.

This network has fewer parameters than traditional convolutional networks, due to the fact that the network does not have to relearn redundant feature maps. The DenseNet architecture makes a clear distinction between information added to the network and information saved. The DenseNet layers are very narrow, only a small group of feature maps are added to the network, and the remaining feature maps are kept unchanged. The final classifier makes decisions based on all feature maps in the network.

For DenseNet, in addition to better parameter efficiency, its advantage is its improved information flow and gradient in the network, which will be easier to train. The gradients of the loss function and the original input signal are directly available for each layer, leading to an implicit deep level of supervision. This helps to train deeper network architectures.The output of the traditional network at l layer is shown in (1):

$$x_l = H_l(x_{l-1}) \tag{1}$$

While for ResNet, the identity function from the input of the previous layer is shown in (2):

$$x_l = H_l(x_{l-1}) + x_{l-1} \tag{2}$$

In DenseNet, all previous layers are connected as inputs is shown in (3):

$$x_l = H_l([x_0, x_1, \cdots, x_l]) \tag{3}$$

The first layer of the original YOLOv5s backbone network is the Focus structure. In this structure, slicing is critical. The input image of 640 × 640 size with 3 channels is passed into the slicing structure and then convolved into a feature map of 320 × 320 × 2 by a convolution operation of 11 size with 32 channels, which reduces the dimensionality and computation. This reduces the dimensionality and computation, as shown in Fig. 3.

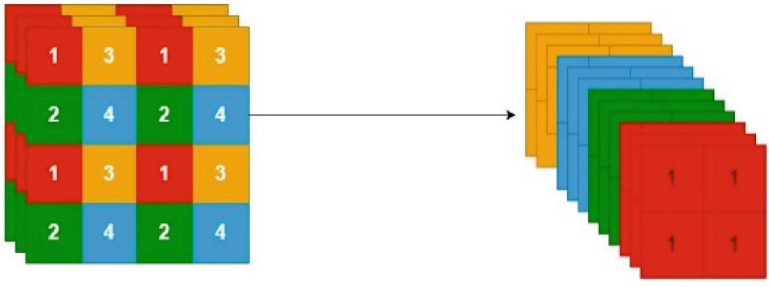

Fig. 3. Focus structure diagram operation diagram

In this paper, we propose to replace the Focus structure by DenseBlock, which can accomplish twice the downsampling to extract feature information at more scales and because this unique connection does not require learning a large amount of useless feature map information, it can improve the gradient information of the network to avoid gradient explosion, which may reduce the appearance of overfitting phenomenon. While increasing the number of finite parameters and computational effort, it improves the feature extraction capability of the network and enables the network to enhance the training of small object feature information. Furthermore, since the final output incorporates the feature outputs of all previous layers, it can retain the feature information of small objects as much as possible during the downsampling process and reduce the loss of their information [15], so that the subsequent convolution operation can extract more information of small objects. Thus, it is beneficial to enhance the training of the model for small objects and improve the performance of the model in detecting small objects.

3.2 Soft-NMS

The idea of NMS algorithm is to search the local maximum and suppress the maximum. NMS algorithm can be specific in different applications. Although its implementation methods are different, they are all the same idea. Non maximum suppression is widely used and plays an important role in edge detection, face detection and other object detection tasks. It removes the redundant frames by multiple iterations, selects the frame with the highest confidence in each iteration, then calculates the intersection ratio (IOU) between the highest confidence preselected frames and the remaining frames, removes the frames with IOU greater than the intersection ratio threshold, obtains the remaining highest confidence preselected frames, and repeats the above process.

The main drawback of non-extreme value suppression is that when objects are highly overlapping, only the prediction boxes with the highest confidence of the same kind are retained, which leads to the possibility that similar but different objects may be mistakenly deleted. Soft-NMS does not directly delete the remaining boxes that generate high intersection ratio with the highest confidence box [16], but reduces the confidence of the remaining boxes and retains more prediction boxes to avoid the mistaken deletion of overlapping objects from occurring in the extreme value suppression. The original NMS algorithm can be expressed as the following (4):

$$s_i = \begin{cases} s_i, & iou(M, b_i) < N_t \\ 0, & iou(M, b_i) \geq N_t \end{cases} \tag{4}$$

Soft-NMS algorithm is shown in (5):

$$s_i = \begin{cases} s_i, & iou(M, b_i) < N_t \\ s_i(1 - iou(M, b_i)), & iou(M, b_i) \geq N_t \end{cases} \tag{5}$$

where N_t is the threshold value, s_i is the new confidence generated by the currently selected object box and the highest confidence box, and $IOU(M, b_i)$ is the intersection ratio of the current highest confidence prediction box M and the remaining i-th prediction

box bi. The image and the enhanced image are fed into the model at the same time, and all the prediction frames obtained by Soft-NMS often contain some useful information, so that when the highest confidence frame is not accurate enough, the information of the remaining frames can be used to correct the prediction frames.When the highest confidence preselected box is not accurate enough, the information of the remaining boxes can be used to correct the preselected boxes.

4 Experiments

4.1 Dataset

In order to carry out experiments related to deep learning object detection, there must be sufficient data sets. The only open source helmet dataset available is SHWD (SafetyHelmetWearing-Dataset), where the SCUT-HEAD dataset is a surveillance image or a photo taken by students in a classroom scene, therefore, the data set is not a picture in the scene of the construction site, and cannot meet the requirements under the background of safety helmet detection. In order to solve this problem, this paper makes a self-made data set that meets the requirements of helmet detection in a construction site.

The data set of this experiment has 6057 pictures about safety helmets, of which 4845 pictures are the training set and 1212 pictures are the validation set. The dataset was downloaded from the web, and the annotation tool was used to mark the object type as well as the coordinates of each photo, and the dataset was divided into two categories, one for the unwearing safety helmet marked as 0, and one for the safety helmet marked as 1. The objects for the unwearing safety helmet included the face as well as the head, and the safety helmet objects were the face and the safety helmet as a whole. The size of the overall dataset photos varies.

4.2 Metrics

Precision (P), recall (R) and mean accuracy (mAP) are used as the relevant metrics for model performance evaluation. The precision rate is used to measure the accuracy of model detection, i.e., the accuracy rate. Recall rate is used to assess the comprehensiveness of model detection, i.e., the check-all rate [17]. The single-category accuracy (AP) is calculated using the integration method to calculate the accuracy and recall curves as well as the area enclosed by the coordinate axes. By summing the single-category AP values and then dividing them by the number of categories, the mAP value can be obtained. mAP values are generally calculated when IOU = 0.5, i.e., mAP@0.5, where IOU is the intersection ratio and is an important function for calculating mAP [9], as shown in the following equation.

$$IOU = \frac{A \cap B}{A \cup B} \tag{6}$$

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$AP = \int_0^1 P(r)dr \tag{9}$$

$$mAP = \frac{\sum_{i=1}^{C} APi}{C} \tag{10}$$

where A and B are the prediction frame and the true frame, respectively, the denominator is the intersection of the two frames, and the numerator is the concatenation of the two frames. TP is true positive; predicting positive object as positive; false positive TP; incorrectly predicting negative object as positive; false negative FN; incorrectly predicting positive object as negative. $P(r)$ is the smoothed accuracy and recall curve, for which the integration operation is to find the area occupied by the smoothed curve. C is the number of categories, and APi denotes the accuracy rate of the i th category, where i is the ordinal number, and the number of categories in this paper is 2.

4.3 Process and Results

The yolov5 improved algorithm experiment proposed in this paper is based on the code provided by the official. The original $3 \times 608 \times 608$ image input Focus structure in the original YOLOv5s, using a slicing operation, first becomes a $12 \times 304 \times 304$ feature map, and after convolution, it finally becomes a $32 \times 304 \times 304$ feature map. For the method proposed in this paper, firstly, the DenseBlock module is used to replace the first layer of Focus structure at the input side, which can pass more information and reduce the gradient between networks by using its dense connection and unique feature map information transfer, thus effectively reducing the loss of small object information in the process of downsampling at the input side. Next, the NMS model is improved and the prediction frame is enhanced using the Soft-NMS algorithm.

Experiments will record the unimproved model parameters, as well as the MAP indices using the two improved methods and after fusing the two methods, respectively.

Since the dataset images vary in size, the trained images are resized to 640 size. The model is trained for 100 rounds, and the learning rate is set to 0.02. Finally, the last trained model and the best one are obtained, and the obtained model is tested to get the results of the unimproved model trained by mAP, and the results of the unimproved model are shown in Table 1.

Table 1. Results of the original model

	P%	R%	mAP@ 0.5%	mAP@ 0.5: 0. 95%
All	92.2	84.0	89.1	55.4
Person	93.6	88.0	92.7	46.6
Hat	90.7	80.0	85.4	64.1

In order to compare the results of different models more intuitively, the model with DenseBlock is denoted as DN-YOLOv5s, the model with Soft-NMS algorithm is denoted as S-YOLOv5s, and the method incorporating both algorithms is denoted as DNS-YOLOv5s. The parameters such as the number of training rounds are not changed and the comparison of the improved model after the safety helmet set is shown in Table 2.

Table 2. Results of the improved model on the dataset

	P%	R%	mAP@ 0.5%	mAP@ 0.5: 0. 95%
YOLOv5s	92.2	84.0	89.1	55.4
DN-YOLOv5s	92.7	86.0	90.4	57.1
S-YOLOv5s	93.0	87.3	91.6	58.6
DNS-YOLOv5s	93.2	87.5	92.2	59.8

In order to further measure the performance of the algorithm for helmet detection in this paper, Table 3 was drawn with reference to the model data of other papers for safety helmet detection, which can be more intuitive to compare the data obtained from the experiments done in this paper. After the comparison, the feasibility of the improved model is again verified.

Table 3. Performance of other object detection algorithms

	mAP@ 0.5%	mAP@ 0.5: 0. 95%	Fps/(frame)
YOLOv3	74.39	44.67	17.36
YOLOv3-tiny	72.87	43.12	21.48
YOLOv3-spp	75.78	47.15	16.58
YOLOv4	84.16	53.92	16.03
YOLOv5s	81.37	52.26	24.01
YOLOv5m	84.93	54.42	19.94
YOLOv5l	86.58	56.31	17.87
YOLOv5x	87.12	56.96	16.07
YOLOX-L	86.89	57.17	17.23
PP-YOLOv2	86.73	57.03	17.12
Pre Model	89.1	55.4	23.12
Our Method	92.20	59.80	23.35

4.4 Result Analysis

After the results of this experiment, it is obvious that the improved model has some optimization. For the D-YOLOv5s model with the addition of the DenseBlock block, there is an increase of 1.3% points on mAP@ 0.5% and 1.7 percentage points on mAP@ 0.5: 0. 95% compared to the original model. The other method also increased by 2.5% and 3.2%, respectively, compared to the original model. The model after incorporating the two improved methods increased by 3.1% on mAP@ 0.5% and 4.4% on mAP@ 0.5: 0. 95% compared to the initial model. In summary, from the two results in Table 1 and Table 2, and compared with other object detection algorithms in Table 3, it is clear that the use of the two improved methods proposed in this paper does effectively increase the accuracy of the model detection, and there is a significant improvement from the data.

5 Conclusion

In this paper, we proposed an improved method of safety helmet detection based on YOLOv5s algorithm. Aiming at the problems such as small safety helmet targets and dense detection scene targets, we proposed to use denseblock module to replace the focus structure in the network in YOLOv5, which improved the feature extraction ability of the network. Secondly, we adopted soft NMS algorithm to retain the effective prediction frame. Then, the relevant experiments are designed. Compared with the original model, the results of the experiments show that our method has a certain improvement in the accuracy of helmet detection, and the data after the experiments also prove the effectiveness of the improved algorithm, so as to obtain a more accurate helmet detection model.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grant 62071157, National Key Research and Development Programme 2022YFD2000500 and Natural Science Foundation of Heilongjiang Province under Grant YQ2019F011.

References

1. Yan, G., Sun, Q., Huang, J., et al.: Helmet detection based on deep learning and random forest on UAV for power construction safety. J. Adv. Comput. Intell. Intell. Inform. **25**(1), 40–49 (2021)
2. Girshick, R., Donahue, J., Darrell, T., et al.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
3. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
4. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, p. 28 (2015)

5. Redmon, J., Divvala, S., Girshick, R., et al.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
6. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
7. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
8. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
9. Liu, Y., Lu, B.H., Peng, J., et al.: Research on the use of YOLOv5 object detection algorithm in mask wearing recognition. World Sci. Res. J. **6**(11), 276–284 (2020)
10. Cheng, R., He, X., Zheng, Z., et al.: Multi-scale safety helmet detection based on SAS-YOLOv3-tiny. Appl. Sci. **11**(8), 3652 (2021)
11. Zhu, L., Geng, X., Li, Z., et al.: Improving yolov5 with attention mechanism for detecting boulders from planetary images. Remote Sens. **13**(18), 3776 (2021)
12. Wang, Y., Hao, Z.Y., Zuo, F., et al.: A fabric defect detection system based improved YOLOv5 detector. J. Phys. Conf. Ser. **2010**(1), 125–134 (2021)
13. Li, G., Zhang, M., Li, J., et al.: Efficient densely connected convolutional neural networks. Pattern Recogn. **109**, 107610 (2021)
14. Albahli, S., Ayub, N., Shiraz, M.: Coronavirus disease (COVID-19) detection using X-ray images and enhanced DenseNet. Appl. Soft Comput. **110**, 107645 (2021)
15. Jung, E., Chikontwe, P., Zong, X., et al.: Enhancement of perivascular spaces using densely connected deep convolutional neural network. IEEE Access **7**, 18382–18391 (2019)
16. Bodla, N., Singh, B., Chellappa, R., et al.: Soft-NMS–improving object detection with one line of code. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5561–5569 (2017)
17. Hsu, W.Y., Lin, W.Y.: Adaptive fusion of multi-scale YOLO for pedestrian detection. IEEE Access **9**, 110063–110073 (2021)