Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.019085

ARTICLE

An Automated Detection Approach of Protective Equipment Donning for Medical Staff under COVID-19 Using Deep Learning

Qiang Zhang[1], Ziyu Pei[1], Rong Guo[1], Haojun Zhang[2], Wanru Kong[2], Jie Lu[3] and Xueyan Liu[1,*]

[1]Department of Computer Science and Engineering, Northwest Normal University, Lanzhou, 730070, China

[2]Gansu Provincial People's Hospital, Lanzhou, 730070, China

[3]Health Statistics and Information Center of Gansu Province, Health Commission of Gansu Province, Lanzhou, 730070, China

*Corresponding Author: Xueyan Liu. Email: liuxy@nwnu.edu.cn

Received: 02 September 2021 Accepted: 25 January 2022

ABSTRACT

Personal protective equipment (PPE) donning detection for medical staff is a key link of medical operation safety guarantee and is of great significance to combat COVID-19. However, the lack of dedicated datasets makes the scarce research on intelligence monitoring of workers' PPE use in the field of healthcare. In this paper, we construct a dress codes dataset for medical staff under the epidemic. And based on this, we propose a PPE donning automatic detection approach using deep learning. With the participation of health care personnel, we organize 6 volunteers dressed in different combinations of PPE to simulate more dress situations in the preset structured environment, and an effective and robust dataset is constructed with a total of 5233 preprocessed images. Starting from the task's dual requirements for speed and accuracy, we use the YOLOv4 convolutional neural network as our learning model to judge whether the donning of different PPE classes corresponds to the body parts of the medical staff meets the dress codes to ensure their self-protection safety. Experimental results show that compared with three typical deep-learning-based detection models, our method achieves a relatively optimal balance while ensuring high detection accuracy (84.14%), with faster processing time (42.02 ms) after the average analysis of 17 classes of PPE donning situation. Overall, this research focuses on the automatic detection of worker safety protection for the first time in healthcare, which will help to improve its technical level of risk management and the ability to respond to potentially hazardous events.

KEYWORDS

COVID-19; medical staff; personal protective equipment donning detection; deep learning; intelligent monitoring

1 Introduction

As the most serious global public health emergency in 2020, Coronavirus disease 2019 (COVID-19) posed a great threat to the safety of the public [1], especially the health care personnel in the areas with the highest risk of infection [2,3]. Since SARS-CoV-2 transmits primarily through respiratory droplets and contact, proper use of personal protective equipment (PPE) can significantly reduce the risk of cross-transmission. According to the Guidelines on the Use Scope of Common Medical Protective Equipment in the Prevention and Control of Coronavirus Infection Pneumonia

(Trial) [4], a complete dataset of PPE in a medical operating environment should include a medical headcover, medical goggles, inner medical mask, outer medical mask, medical protective clothing, medical protective gloves, and medical protective foot covers. Due to the high incidence of infection caused by exposure in protective medical operations, improper use of PPE is highly likely to lead to infection and even death [5]. Therefore, medical staff in epidemic prevention and control operations must wear PPE correctly before they are allowed to enter the medical site to carry out operations.

However, working for a long time will cause a great consumption of the medical staff's energy, even after professional protection training, it is impossible to guarantee that donning multiple types of PPE is completely correct. Therefore, it is very necessary to take some measures to increase the risk response capacity of healthcare [6], such as the implementation of monitoring and management [7] of the results of medical staff donning PPE. The main problem of the PPE detection is to identify whether each body part of the medical staff is correctly donning protective equipment as required so that errors in use can be corrected in time to ensure safety. Examples of correct donning and some incorrect donning of PPE are shown in Fig. 1. Currently, there are few research results on intelligent detection of PPE using monitoring for medical staff, mainly by the personnel of the infection prevention and control group for manual inspection 24 h a day [8]. But such a monitoring model has many problems, Chen et al. used the visual monitoring system of "4G network transmission and a cloud call platform" to realize 24 h real-time monitoring, although its real-time detection is strong, the whole work is easily affected by the subjective consciousness of personnel, and long-term monitoring will bring very unstable detection results.

Figure 1: Samples of PPE donning images (a) Correct (b) Hand part error (no medical protective gloves) (c) Head, hands, and feet part errors (no medical headcover, protective gloves, and foot covers)

With the development of intelligent video, deep learning-based target detection algorithms in the field of computer vision show better performance than traditional manual methods in various practical application scenarios [9]. In particular, a one-stage algorithm known for its processing time and a two-stage algorithm known for its detection accuracy is applied according to the different requirements of detection tasks, respectively [10]. Different detection tasks need target detectors with different properties for learning [11–13]. For example, vehicle anti-collision detection task needs to use fast but low-precision detectors, while searching for parking spaces requires a slow but high-precision model.

PPE donning detection for medical staff has high requirements for detection accuracy and speed. Therefore, to propose an automated detection model to assist or even replace supervisors to monitor the situation of medical staff donning multiple PPE under COVID-19, to help that medical staff are in a safe protective state continuously, the main work and contributions of this study are as follows:

(1) A Medical Staff Dress Code Dataset (MSDCD) is constructed in a structured scenario to solve the problem of lack of data in the COVID-19 risk environment. Randomly combine different PPE classes to simulate possible donning errors for automatic detection of PPE used by medical staff. Each image in the dataset is annotated with multiple labels and bounding boxes. Data augmentation makes the data more effective and robust and prevents the model from overfitting. The protection rules are visualized by images.

(2) Considering that in the context of the epidemic, the identification task of medical staff donning PPE has dual requirements for detection accuracy and speed, especially in real-time. This paper proposes a PPE donning detection method for medical staff under COVID-19 based on the YOLOv4 network (MSPPE-YOLOv4), by simultaneously locating and classifying the PPE classes corresponding to the body parts of the medical staff in the image, the location and category information of the target can be directly obtained to determine whether the use of PPE complies with the protection rules.

(3) Different from the common tasks of automatic monitoring of PPE used by workers, this paper focuses on the healthcare field for the first time, discussing the possibility of incorrect use of multiple classes PPE, i.e., protective headcover, goggles, masks, clothing, gloves, and foot covers. Compared with the typical two-stage and one-stage target detection algorithms, the results prove that our method achieves a good balance between performance and efficiency on MSDCD, and obtains relatively accurate predictions in real-time monitoring. Furthermore, it strengthens the level of medical safety protection monitoring and improves the system's capability to respond to similar risk events at the technical level.

The following structure of the paper is organized as: Section 2 describes recent research on PPE donning automatic detection, Section 3 introduces the details of the self-built dress code dataset and the proposed PPE detection model. The experimental results and external validation are given in Section 4. Conclusions and future work are presented in Section 5.

2 Related Works

Research on the donning detection of workers' PPE in various high-risk fields is driven by the urgency of demand. The widespread use of surveillance cameras in work scenes makes personal safety protection monitoring based on computer vision instead of subjective human supervision. At present, vision-based automatic identification methods for PPE donning are mainly divided into two categories: traditional features methods and deep-learning-based methods.

In the first category of vision-based methods, some traditional manual selection features are applied to detection tasks for PPE. Park et al. [14] used the histogram of oriented gradients feature and background difference method to match the two parts of the human head part and the helmet part to detect whether the worker is donning a helmet. Shrestha et al. [15] detected whether there was an edge of a helmet in the head area of a construction worker, and used an edge detection algorithm to recognize a single helmet based on facial features. Wu et al. [16] used the difference of the skin color features and PPE color as basic information to explore the donning of equipment. However, since various classes of PPE objects worn by workers may appear in different positions in different shapes

and sizes, if a feature learning model is used, it may fail to be recognized in a complex multi-object PPE donning scene. For example, for medical staff donning the correct facial area equipment, the face is not visible and cannot be recognized using this type of method.

For vision-based deep learning methods, researchers currently mainly rely on two-stage and one-stage object detection algorithms to locate and classify PPE used by workers. Two-stage, as the name suggests, divides the entire PPE donning detection task into two stages: regional positioning and equipment classification and identification. For example, the author in [17] proposed a method based on Faster R-CNN to locate the target area of the personnel first, and then determine the use of equipment in the area, to determine the helmets and masks worn by the workers engaged in the pollution site remediation industry is it right or not. This type of algorithm can get excellent results in the accuracy of model detection. Compared with it, the one-stage algorithm has a simple structure, which concentrates positioning and classification in one stage. End-to-end learning can ensure more efficient calculation efficiency while having good detection accuracy. It is often used for some tasks with high real-time requirements. Such as face detection and public mask detection [18,19]. Guo et al. [20] proposed a detection framework based on the SSD algorithm to meet the real-time requirements of the intelligent safety supervision of the power system, it is used to judge the wearing of helmets, goggles, and other equipment by the operators. The simple network framework provides them with more real-time decision analysis.

According to the requirements of different detection tasks, some scholars study and apply one-stage and two-stage detection algorithms respectively, bringing innovations to the actual monitoring works. Nevertheless, there are still the following problems: (1) A monotonous focus area. Different from other object detection tasks, the detection of PPE used by workers has specificity and particularity in their field. Fields with urgent needs are studied first, and most of the research is devoted to the field of civil engineering. For the healthcare field where daily urgency is not high, the above research is rare. However, the outbreak of COVID-19 has given us a warning that under the background of such a big risk of urgent need, the lack of a more efficient automatic monitoring management method for PPE donning of medical staff has been exposed. (2) A small number of detected PPE objects. Although the current research on the detection of a single type of PPE, such as helmets and masks, is relatively mature, the research on multiple types of PPE detection is relatively rare. It is not a simple superposition plan of the results of many single types of PPE detection, and other factors need to be considered, for example, whether the capture of all PPE objects in the global scope is complete. Medical staff under the epidemic need to use eight classes of PPE correctly, so such a detection task is by no means as simple as the task of detecting whether workers wear safety helmets. (3) A different task requirement. Some tasks have high requirements for detection accuracy, while others pay more attention to real-time performance, which is determined according to the specific needs of different tasks. Although the two types of algorithms based on deep learning each show better results in detection accuracy and speed, as far as the task of this research is concerned, the real-time performance is higher than the detection accuracy, so the one-stage algorithm with a simple network framework is more considered by us.

3 Proposed Methods

3.1 Medical Staff Dress Code Dataset

3.1.1 Data Collection

Because there is no public PPE donning dataset for medical staff, this paper builds a medical staff dress code dataset—MSDCD in a structured scenario (preset and controlled environment). A total of 1500 images from six volunteers of different body types were collected. Different combinations of PPE are used to simulate various possible donning situations. Considering that there are many detection PPE objects in the medical staff's body parts, to enhance the robustness of the model, MSDCD is divided into Part A and Part B, in which the samples in Part A shows the whole-body images and the samples in Part B shows the local-body images. Fig. 2 shows some samples of the different PPE combinations donning images in MSDCD, including the whole-body and local-body images.

Figure 2: The different PPE combinations donning images in MSDCD (Part A is the whole-body image set, including but not limited to all dressed, unhat, ungloves, unmask, etc. Part B is the local-body image set, including but not limited to the head, mouth, hands, feet and so on)

In this paper, PPE donning is divided into 17 classes according to eight parts of the human body, which intuitively reflects the donning situation of various equipment in different parts and is convenient for model training. Based on the standard operating system of PPE for medical staff [21], this paper defines the PPE dress code for medical staff and classifies the dress situation. In other words, if and only if the PPE donning situation of 8 body parts for medical staff exactly meets class 1, medical staff can be allowed to enter the worksite, otherwise, they are deemed to be unqualified. Table 1 shows the classification of PPE usage in this paper.

Table 1: Classification of PPE usage in different body parts

Body parts	Class 1	Class 2	Class 3
Head	hat	unhat	
Eyes	glasses	unglasses	
Mouth	mask	unmask_one	unmask_two
Body	cloth	unzip	
Left hand	glove_l	unglove_l	

(Continued)

Table 1 (continued)

Body parts	Class 1	Class 2	Class 3
Right hand	glove_r	unglove_r	
Left foot	shoe_l	unshoe_l	
Right foot	shoe_r	unshoe_r	

Note: Class 1 indicates the correct donning of PPE on all human body parts. Class 2 indicates that each PPE is not or incorrectly donned in various parts of the body (for example, the protective clothing is not donned or is donned but its zipper is not fastened). Class 3 indicates the third scenario for the mask (with two masks, no outer mask, no inner and outer mask).

3.1.2 Data Preprocessing

To get a dataset with high availability, all data need to be preprocessed mainly from the following three aspects after the preliminary data collection is completed:

(1) Data cleaning: Not all images taken are valid data, so duplicate and fuzzy data that have a certain influence on model training will be deleted. In addition, to prevent the shortage or redundancy of the model training for a certain target, the number of data samples of different PPE combinations should reach equilibrium. After cleaning MSDCD, the total sample size is 1353.

(2) Data labeling: YOLOv4 uses the anchor frame as the basic detection mechanism, and the anchor frame takes the anchor point as the center to obtain different windows to detect multiple PPE objects. However, the anchor frame only refers to the effective area size of the image, when the size of the boundary box is returned. This mechanism is suitable for the object detection task with a small number of effective features in an image. Considering that under the COVID-19 environment, there are high numbers of PPE categories used by medical staff, the whole image contains a large number of effective target regions, so the dataset format needs to be converted when making accurate annotations. To obtain a reliable and accurate dataset, we adopted a three-step labeling method. First, three health care personnel who have expertise in PPE use protection management for medical staff are invited to confirm the category and label of PPE, which are manually labeled by computer professionals. By combining the whole and local features of images, the label is converted into the Pascal VOC dataset format [22]. It mainly used two parts of format, JPEGImages: image dataset $\{x_i\}_{i=1}^{N}$; Annotations: labeling dataset $\{y_i\}_{i=1}^{N}$. Second, two health care personnel carefully check the preliminary completed annotations and corrected possible mislabeling. Third, to ensure consistency of classification, a health care worker checked all annotations. In the end, high-quality annotations were obtained through this method. Fig. 3 shows the label box information of each detection target.

(3) Data augmentation: In the re-training and fine-tuning phases of the MSPPE-YOLOv4 model, data augmentation is performed based on the original dataset to support better training of the model. In particular, the original images are scaled up or down by $\pm 20\%$ randomly, rotated by a certain angle around the center of the image within the range of $[-20^{\circ}, +20^{\circ}]$, and horizontally and vertically translated by $\pm 20\%$ as the first set of expansion methods. Additionally, the color space of the image is changed, and its brightness and contrast are randomly changed within the range of $[-20\%, +20\%]$ and $[-25\%, +50\%]$, respectively, as the second set of methods to amplify the dataset. Based on the original dataset, the two sets of methods enhance the diversity of images with single and double ratios, and generate multiple versions of similar

images. The total amount of processed data is 5233 images. More data helps prevent the model from overfitting. The sample sizes of each of the 17 classes of PPE objects in the MSDCD dataset are shown in Fig. 4.

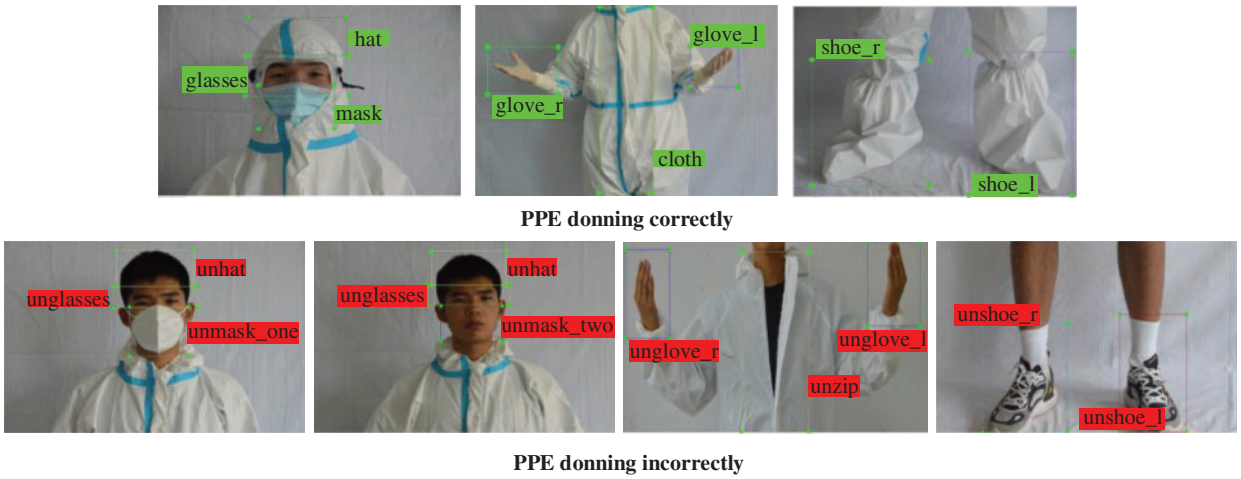

Figure 3: Examples of each labeling detection box in MSDCD (There are 8 cases of correct PPE donning, represented by green, and 9 cases of wrong PPE donning, represented by red)

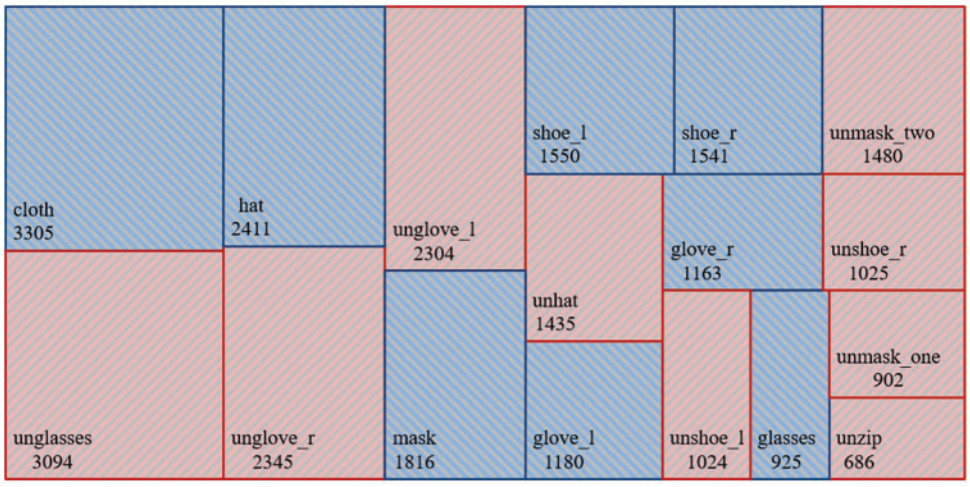

Figure 4: The number of samples for each PPE class in the MSDCD dataset

3.2 Detection Model Based on YOLOv4

In the PPE donning identification task, the convolutional neural network (CNN), as a deep learning method that uses a multi-level structure network, directly uses the collected image as the input of the network and obtains the spatial features through the receptive field, to determine whether the workers use PPE, and give the category and location information of the existing equipment. It avoids the dependence on manual feature extraction and the data reconstruction problem in traditional detection algorithms. To automatically identify PPE objects in multiple categories and scales used by medical staff in the epidemic, and at the same time consider the detection performance and efficiency

of the model, we propose a deep learning-based YOLOv4 convolutional neural network [23] detection method (MSPPE-YOLOv4), using regression analysis to directly predict the location and category of the PPE objects, pursuing higher demand real-time performance while ensuring accuracy. The overall framework is shown in Fig. 5.

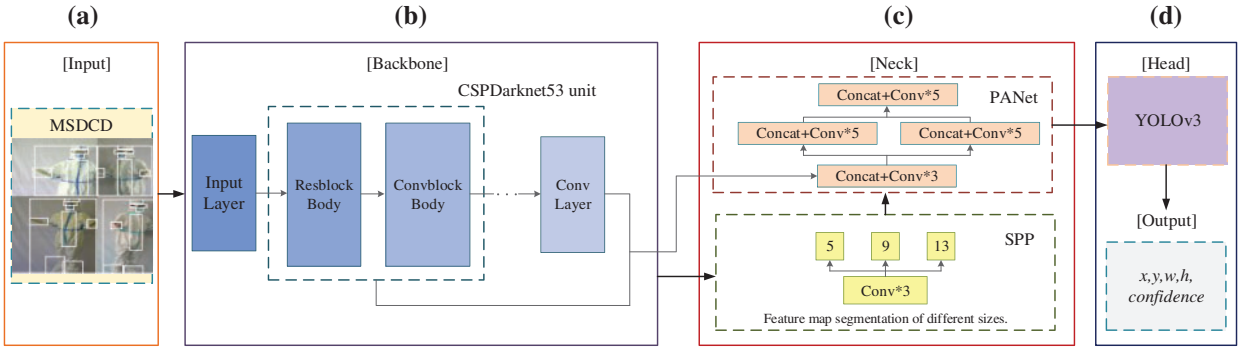

Figure 5: The overall framework of the proposed method

The automatic detection of PPE donning of medical staff has dual requirements in terms of accuracy and speed, especially in real-time. Therefore, we chose YOLOv4, a classic one-stage object detection algorithm, and applied it to MSDCD. YOLOv4 is not so much a pure algorithm, as it is a multiple sub-technology fusion. Through experiments, Bochkovskiy et al. compared multiple universal algorithms and modules and finally found a combination that can achieve the best balance between accuracy and speed. It has been used in different applications by researchers in many fields [24–26], and the verification of reliability is also an important reason why we chose it as the method of this research. It is no secret that the constant update of the algorithm makes the task have better assistants to perform. In contrast, YOLOv4 may be slightly inferior, but even so, their basic ideas are almost the same (feature extraction and bounding box regression). As far as our research is concerned, we are more focused on proposing the idea of automatic detection of PPE in the field of healthcare to arouse people's attention. Reliable, fast, and accurate YOLOv4, which combines multiple technologies, meets our mission requirements. If the experimental results prove the feasibility of this approach, multiple sub-modules in the model will be replaced and updated as needed.

Based on the PyTorch [27] dynamic learning framework, the PPE identification task based on object detection execution benchmarks are divided into five structures: input, backbone network responsible for pre-training, neck for collecting feature maps, head for prediction, and output. To make the model more flexible calculation operation, the concept of transfer learning is applied to the backbone network, and the pre-trained knowledge is transferred to MSDSD to learn new features, which are used to detect the PPE donning for medical staff. The following details the various parts of the system framework.

(a) On the input terminal, this model will randomly read four images at once from each batch. After random rotation, scaling, or color gamut changes are performed on them, the images are combined and spliced according to the four position directions. Each batch needs to repeat batch_size times of splicing operations. In this way, the basic data set is greatly enriched, and four images can be calculated at a time when calculating batch normalization, which improves the robustness of the model.

(b) To obtain a dedicated model, it is necessary to train the model from scratch based on the MSDCD, because there was no dataset centered on the PPE used by medical staff before. However, such work is very time-consuming. The introduction of transfer learning [28] overcomes this challenge. It allows the knowledge learned on some larger datasets (such as the ImageNet dataset [29] and Microsoft COCO dataset [30]) to be transferred to a new dataset (MSDCD) that is related to it according to the weight of the parameter. Train a dedicated model based on pre-learning features, optimize and accelerate the training efficiency of the model. In this study, we migrated the YOLOv4 model pre-trained on the COCO dataset to the backbone network of this model trained on the MSDCD dataset because it learned some targets (like person object) related to this research in advance.

When the input terminal sends the enhanced image to the backbone network, it first uses the mish activation function to perform convolution operations on it. Due to its low cost, smoothness, no upper bound, lower bound, etc., compared with other functions such as ReLU, it can reduce the amount of calculation while ensuring accuracy. The calculation formula of the mish function is:

$$Mish(x) = x \times \tanh(\ln(1 + e^{x})) \tag{1}$$

CSPDarknet53 is based on the DarkNet-53 [31] network and introduces Cross-Stage-Partial (CSP) [32] for optimization, which can prevent the information in the recursive computation from being reused to update different block weights. To a certain extent, it strengthens the learning ability of the network, eliminates computing bottlenecks, and reduces the hardware cost required for computing. Each unit of CSPDarknet53 is composed of blocks and layers, and each block contains k convolutional layers, and its output is determined by the connection operation between the layers and the weight of each block itself (using the backpropagation algorithm to update the weight). CSP splits the stack of the residual block into two routes: $x_0 = \{x_0', x_0''\}$ according to the network configuration parameter, x_0' is directly connected to the end of the stage after a small amount of processing and x_0'' continues to stack the blocks. Fig. 6 shows the structure of the residual block after the introduction of CSPNet.

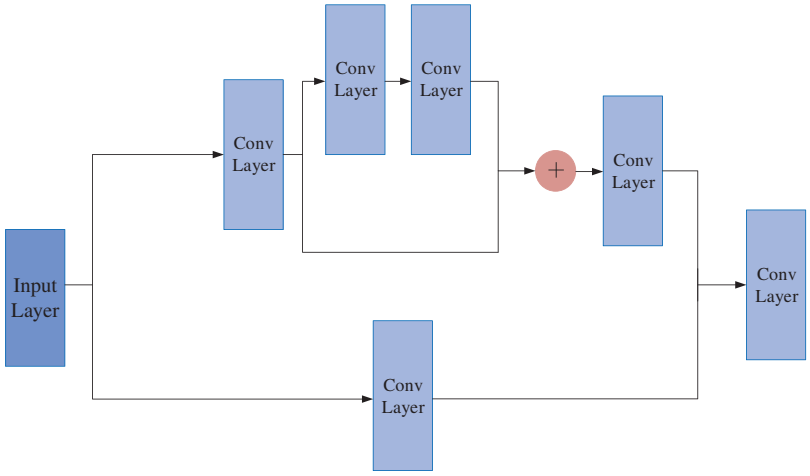

Figure 6: The structure of Resblock

(c) The detector usually consists of a pre-trained backbone and a head responsible for predicting target information. Recently, some new network layers have been inserted between the two to collect feature maps at different stages to provide more accurate feature information for subsequent prediction work, called neck. There are multiple PPE objects to be detected with different sizes in this research,

such as protective masks, protective clothing, etc. It is difficult for the fully-connected layers commonly used in ordinary CNN models to complete this task because it is limited to the classification of fixed-size images. This is contrary to our research. In this regard, we add spatial pyramid pooling (SPP) [33] to improve the receptive field. After convolution in the last feature layer of the backbone network, we use three scales to perform maximum pooling processing on the head and collect effective features at different scales.

When collecting features, some underlying information is likely to be lost due to parameter adjustments. As an instance segmentation algorithm with the ability to repeatedly extract features, Path Aggregation Network (PANet) [34] adds a path to transmit information from the lowest layer to the highest layer, which can pass information to the upper layers to protect the information flow that may be lost. Based on the feature pyramid network, it introduces adaptive feature pooling so that each proposal aggregates the information of all layers in the pyramid, which is more beneficial for the classification and location determination of multiple PPE objects. Its fusion method is replaced from the original addition operation to concatenation to better improve the accuracy of head prediction.

(d) In the prediction part of the model, we continue to use the YOLOv3 head—a one-stage detector to classify multiple PPE objects and determine the position coordinates. Taking into account the limitations of computing resources and the high real-time requirements of the task, unified classification and positioning is a task, and the region proposal stage is removed. Single testing can directly obtain the positions and types of multiple PPE used by medical staff in an image, to judge whether a medical staff donning PPE correctly according to the protection rules. The images are mapped into $n \times n$ fixed grids, and each grid is responsible for detecting the probability of an object in its center, using regression analysis. Specifically, we divide the *416 pixels* × *416 pixels* input image into *52 grids* × *52 grids*, *26 grids* × *26 grids*, and *13 grids* × *13 grids* to detect small, medium, and large PPE objects, respectively. The prediction effect of the object information is represented by the intersection over union (IoU), by considering the ground-truth bounding box (box (truth)) and the predicted bounding box (box (pred)) at the same time, the center coordinates, width and height of the object bounding box are taken as a 4-tuple (x, y, w, h) instead of independent variables for analysis. IoU represents the percentage of overlap between the box (pred) and the box (truth), its specific meaning and calculation method are shown in Fig. 7 and Eq. (2).

$$IoU = \frac{area(box(pred) \cap box(truth))}{area(box(pred) \cup box(truth))} \tag{2}$$

The ideal situation of IoU is that the two boxes completely overlap, that is, the ratio is 1. To evaluate the performance of the MSPPE-YOLOv4 by calculating accuracy, precision, recall, etc., first determine true positive (TP, the number of correctly detected PPE targets, $IoU \geq 50\%$), false positive (FP, the number of incorrectly detected PPE targets, $IoU < 50\%$), and false negative (FN, the number of targets missed by the model, $IoU = 0\%$) based on IoU. Confidence (in Eq. (3)) reflecting the reliability of the recognition of a single PPE object is returned in the output layer. Pred (object) means whether there is a PPE object falling into a certain grid of the image, if it has a value of 1, otherwise it is 0.

$$Confidence(object) = pred(object) \times IoU \tag{3}$$

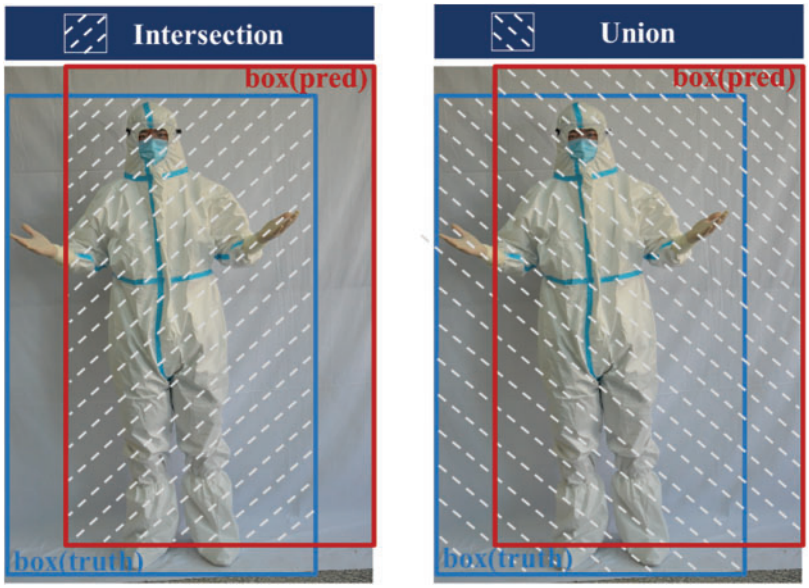

Figure 7: The meaning of IoU between the box (truth) (a blue box) and box (pred) (a red box)

4 Results and Discussion

4.1 Evaluation Indexes of Model Test

It is necessary to ensure the complete detection of multiple PPE objects used in 8 parts of the medical staff's body while maintaining a high detection accuracy rate for medical staff donning PPE. At the same time, the time cost of the model should be reduced as much as possible to pursue high-demand real-time. After the MSPPE-YOLOv4 model training is completed, we use several different performance indicators to measure the performance of the model detection, including precision (P), recall (R), accuracy (A), F1-score (F1), and processing time. The calculation methods of several evaluation indicators are as follows:

- P is a measure of how much of all the PPE objects given by the model are accurate in terms of the predicted results. The calculation method is shown in Eq. (4).

$$P = \frac{TP}{TP + FP} \tag{4}$$

- R is to evaluate whether a model algorithm is complete to identify the PPE object based on the original sample. The calculation method is shown in Eq. (5).

$$R = \frac{TP}{TP + FN} \tag{5}$$

- A represents the proportion of correctly detected PPE object types in the total predicted PPE classes. The calculation method is shown in Eq. (6).

$$A = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

- To ensure that the value of R is stable under the premise of P stability, the concept of F1 is used to make the weighted harmonic mean of P and R for unified overall evaluation. The calculation

method is shown in Eq. (7).

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{7}$$

- The model processing time is measured by the average processing time of a single image in the testing set.

In this paper, we have multiple two-category confusion matrices. The P and R of n two-category confusion matrices hope to be comprehensively evaluated. A straightforward approach is to calculate the P and R and then calculate an average value on each confusion matrix, thus obtaining "macro-P", "macro-R", and the corresponding "macro-F1". The calculations are shown in Eq. (8).

$$\begin{aligned} macro\text{-}P &= \frac{1}{\mathrm{n}} \sum_{i=1}^{n} P_i \\ macro\text{-}R &= \frac{1}{n} \sum_{i=1}^{n} R_i \\ macro\text{-}F1 &= 2 \times \frac{macro\text{-}P \times macro\text{-}R}{macro\text{-}P + macro\text{-}R} \end{aligned} \tag{8}$$

4.2 MSPPE-YOLOv4 Model Training and Results

The learning framework of model experiments is the famous deep learning platform called PyTorch. The server is configured as Intel(R) Xeon(R) Gold 5218 @ 2.30 GHz CPU, Quadro RTX 6000 GPU, and the operating system as Ubuntu64 as OS. For MSPPE-YOLOv4, the model is trained in two stages: 100 cycles (*Learning_Rate* = 0.001, *Batch_size* = 12) and 200 cycles (*Learning_Rate* $\sim=\sim$ 0.0001, *Batch_size* = 8). *Confidence* = 0.5 is selected as the recognition threshold of positive and negative cases, and the PPE objects in an image are classified as positive cases with *Confidence* $\geq$ 0.5, *Confidence* $<$ 0.5 are classified as negative cases.

The input image size is set to *416 pixels* $\times$ *416 pixels* to facilitate the detection of some small-scale PPE objects, and the minimum bounding box information of various PPE in 5233 images is labeled. In the experiment, the basic dataset MSDCD is split into two datasets Part A (the whole-body images, 1333 images) and Part B (the local-body images, 3900 images). The training set, testing set, and validation set are randomly divided at an 8:1:1 ratio of the total dataset, and the testing set is restricted to be selected only from Part A, considering that our task ultimately requires the detection of the medical staff's overall clothing. The total number of three datasets is 4185, 524, and 524, respectively. We use the validation set to dynamically adjust the training parameters in the iterative training of the model to better find the features. Table 2 is general information of the splitting datasets.

Table 2: General information of data sets

Type of dataset	Part A set	Part B set	Total
Training set	678	3507	4185
Testing set	524	0	524
Validation set	131	393	524
Total	1333	3900	5233

The 8 parts of the human body (head, eyes, mouth, body, left hand, right hand, left foot, and right foot) respectively contain 2–3 donning situations, and the sum of the donning situations corresponding

to each body part is the number of samples in the testing set. The sample size of each PPE class and their detection results (P, R, A, and F1) in 524 randomly selected images is shown in Table 3.

Table 3: Testing results of the MSPPE-YOLOv4 on testing set

No.	Classes	Number of samples	P	R	A	F1
1	hat	320	96.78%	98.69%	95.60%	97.73%
2	glasses	135	91.60%	96.77%	88.89%	94.12%
3	Mask	238	90.43%	90.87%	83.61%	90.65%
4	cloth	327	87.93%	90.11%	81.36%	89.01%
5	glove_l	206	88.24%	88.24%	81.59%	88.24%
6	glove_r	200	87.43%	87.91%	81.09%	87.67%
7	shoe_l	312	90.28%	89.66%	82.04%	89.97%
8	shoe_r	308	90.91%	91.23%	83.65%	91.07%
9	unhat	204	91.37%	96.77%	88.73%	93.99%
10	unglasses	389	92.39%	94.18%	87.40%	93.28%
11	unmask_one	102	88.42%	91.30%	84.55%	89.84%
12	unmask_two	184	89.39%	90.40%	83.18%	89.89%
13	unglove_l	318	87.84%	91.55%	81.25%	89.66%
14	unglove_r	324	87.84%	90.59%	81.08%	89.19%
15	unzip	197	86.49%	92.49%	82.08%	89.39%
16	unshoe_l	212	90.21%	88.38%	82.13%	89.29%
17	unshoe_r	216	89.74%	89.29%	82.10%	89.51%

After testing 524 images containing multiple classes of PPE, the P of MSPPE-YOLOv4 for 17 classes of PPE are all higher than 86%. Specifically, this model has the highest P and R for target 1 "hat" because it has no additional shielding and there are more targets in the testing set. On the contrary, the target 15 "unzip" has the lowest P, which is 86.49%. We speculate that this is due to the small opening of the zipper or the high similarity between the color of the volunteer's inner clothing and the outer protective clothing. But its R is 92.49%, which means that the model can capture "unzip" well. For some smaller targets, such as target 2 "glasses" and target 3 "mask", the P are 91.60% and 90.43%, respectively, which proves that it is effective to map the image to 3 grid forms in head prediction. The detected accuracy of each PPE class is above 81%. Among them, the A of targets 5, 6, 13, and 14 are 81.59%, 81.09%, 81.25%, and 81.08%, respectively. This is because the characteristics of medical gloves and hands are too similar, resulting in poor accuracy of model discrimination. In addition, we performed an average analysis of the execution time of the model on 524 randomly selected testing images, and the results proved that the processing time of a single image of MSPPE-YOLOv4 is about 42.02 ms.

Based on the comprehensive analysis of the detection experiment results of the above 17 PPE classes, the detection performances of the proposed model are calculated according to Eq. (8). The results show that the macro-P of MSPPE-YOLOv4 on MSDCD is 89.84%, the macro-R is 91.67%, and the macro-F1 is 90.74%. On the premise of ensuring a stable detection precision, the target extraction is also comprehensive. Fig. 8 visually analyzes the detection effect of MSPPE-YOLOv4 on three testing images.

Figure 8: Visualization results of MSPPE-YOLOv4. (a) A sample of correctly donning PPE images. (b) A sample of not donning protective gloves images. (c) A sample of not donning medical headcover, protective gloves, and foot cover images. (d) The detection result of (a). (e) The detection result of (b). (f) The detection result of (c)

4.3 Model Comparison and External Validation

To validate the detection performance of the MSPPE-YOLOv4 model, three deep learning-based models most commonly used in object detection tasks would be trained on MSDCD respectively, and these would be compared and analyzed five evaluation indexes of P, R, A, F1, and processing time. Among them, the two-stage typical algorithm Faster R-CNN [35] with higher detection accuracy and the one-stage algorithms YOLOv3 [30] and SSD [36] with shorter processing time are involved in the

experiment. We set up the same experimental environment and learning framework (PyTorch) for each method involved in the experiment. The experimental results are shown in Fig. 9.

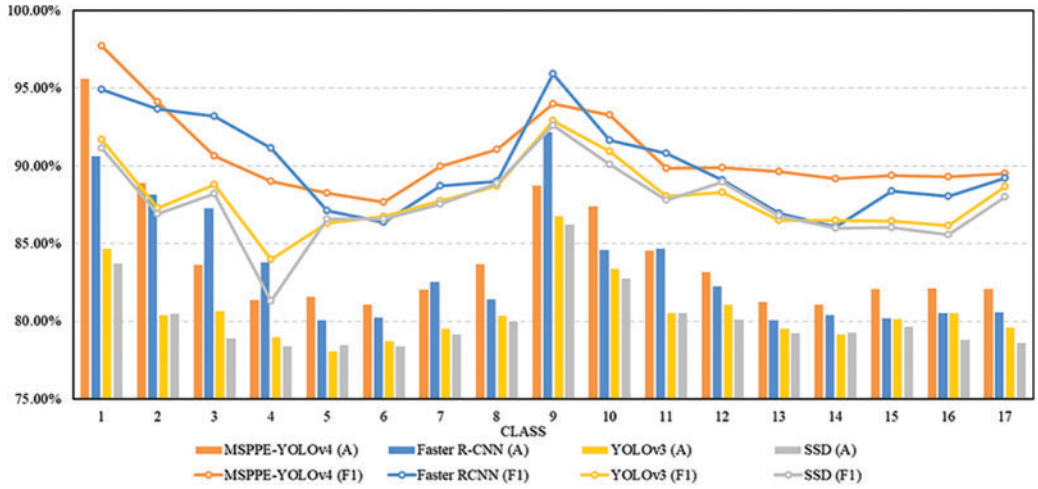

Figure 9: Comparison of the performance of four methods on MSDCD

We convert the P and R results of the methods to detect each PPE class into a comprehensive comparison of F1. The higher the F1 of the PPE object, the more stable the R is while the model has good P. Faster R-CNN has F1 of 93.21%, 91.15%, 95.90%, and 90.81% for target 3 "mask", target 4 "cloth", target 9 "unhat" and target 11 "unmask_one", respectively. The corresponding A also performed best among the four methods. Different from the other three types of one-stage methods, when Faster R-CNN detects PPE objects, it first predicts proposals in the input image, and then classifies the region, which can capture PPE object information more finely to a certain extent. However, the excellent performance of a small amount of PPE objects cannot represent the overall performance of the model, especially for the PPE detection task of medical staff. Although MSPPE-YOLOv4 is not as good as Faster R-CNN in detecting the above four targets, it is better than Faster R-CNN for F1 and A of other 13 classes of targets, such as targets 5, 6, 7, 8. MSPPE-YOLOv4 performs unbiased detection of 17 classes of targets with different sizes, with the introduction of SPP, the F1 and A of the object to be detected are above 87% and 81%, respectively. For F1 of targets 6, 8, and 14, YOLOv3 and SSD are basically the same as Faster R-CNN, but they are slightly inferior to MSPPE-YOLOv4.

According to the data in Fig. 9 and Eq. (8), the four methods of P, R, F1, and A are calculated to compare the overall performance of the model instead of the partial. In addition, the processing time of the model testing images is regarded as an important reference for judging the real-time performance of the model. The average testing time of the four methods on 524 images is calculated to obtain the processing time of each method. Table 4 is a comprehensive analysis of each method on MSDCD.

Table 4: The comprehensive analysis of the selected methods in MSDCD

Method	P	R	A	F1	Processing time
Faster R-CNN	88.91%	91.19%	83.50%	90.04%	52.88 ms
SSD	87.02%	88.30%	80.16%	87.66%	45.54 ms

(Continued)

Table 4 (continued)

Method	P	R	A	F1	Processing time
YOLOv3	87.41%	88.69%	80.72%	88.05%	43.94 ms
MSPPE-YOLOv4	89.84%	91.67%	84.14%	90.75%	42.02 ms

The results show that the A (80.16%) and F1 (87.66%) of SSD are the lowest, and the processing time reached 45.54 ms. The Faster R-CNN model has a high detection accuracy (83.50%) and F1 score (90.04%), but it is also the most time-consuming, with an average processing time of 52.88 ms. The A, F1, and processing time of YOLOv3 are relatively balanced, respectively 80.72%, 88.05%, and 43.94 ms. The F1 of MSPPE-YOLOv4 is 0.71% higher than that of Faster R-CNN, and the A is 0.64% higher. At the same time, the processing time of a single image reaches 42.02 ms, achieving the best balance between detection performance and efficiency. The end-to-end regression analysis makes the detection efficiency of the model higher.

The structure of the detection model will change due to the different requirements of detection tasks. In this study, to realize the possibility of automatic monitoring of the PPE donning of medical staff in the field of healthcare, based on the actual needs of the task, YOLOv4 is selected as the model basis, because it has a faster time under the premise of ensuring the detection accuracy deal with. Of course, if we do not pursue real-time monitoring, but conduct some offline research to analyze related issues, then Faster R-CNN will also be a good choice if we only consider the index of detection accuracy.

5 Conclusions

This research proposes a PPE donning automatic detection model for medical staff (MSPPE-YOLOv4) based on YOLOv4, which can use deep learning methods to carry out intelligent detection of multiple PPE objects. On the basis of the results of our study, this model can be used to stably and efficiently monitor the PPE donning situation for medical staff and help reduce the potential harm caused by human subjective consciousness in the management process and save medical resources. The MSPPE-YOLOv4 model is tested using the self-built dataset (MSDCD), and the detection accuracy reaches 84.14%, while the running time of processing a single image is 42.02 ms. The life safety of medical staff is the basis for fighting infectious diseases and the prerequisite for protecting public health. High-efficiency monitoring of their PPE donning is very important to overcome the challenges of future health crises and build a healthier medical team in the city. In future work, we will improve the detection accuracy and processing speed of the model through further processing the dataset and the model structure. The optimized model will be deployed in the hardware device, which can use the images collected on the spot to display the PPE donning situation for medical staff in real-time, complete the efficient real-time detection, and help to innovate the detection work of PPE in the healthcare field.

Acknowledgement: The authors wish to express their appreciation to the reviewers for their helpful suggestions which greatly improved the presentation of this paper.

Funding Statement: This research was partially supported by the grants from the Natural Science Foundation of China (No. 72161034).

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

References

1. World Health Organization (2020). Coronavirus disease (COVID-19) pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019.
2. Ge, X. Y., Pu, Y., Liao, C. H., Huang, W. F., Zeng, Q. et al. (2020). Evaluation of the exposure risk of SARS-CoV-2 in different hospital environment. *Sustainable Cities and Society, 61,* 1–7. DOI 10.1016/j.scs.2020.102413.
3. Chinese Center for Disease Control and Prevention Epidemiology Working Group for NCIP Epidemic Response (2020). The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. *Chinese Journal of Epidemiology, 41(2),* 145–151. DOI 10.3760/cma.j.issn.0254-6450.2020.02.003.
4. National Health Commission of the People's Republic of China. (2020). COVID-19 epidemic prevention and control. http://www.nhc.gov.cn/xcs/zhengcwj/202001/e71c5de925a64eafbe1ce790debab5c6.shtml.
5. Chughtai, A. A., Chen, X., Macintyre, C. R. (2018). Risk of self-contamination during doffing of personal protective equipment. *American Journal of Infection Control, 46(12),* 1329–1334. DOI 10.1016/j.ajic.2018.06.003.
6. Neto, G. C. D., Tucci, H. N. P., Godinho, M., Lucato, W. C., Correia, J. M. F. (2021). Performance evaluation of occupational health and safety in relation to the COVID-19 fighting practices established by WHO: Survey in multinational industries. *Safety Science, 141,* 1–12. DOI 10.1016/j.ssci.2021.105331.
7. Juvet, T. M., Corbaz-Kurth, S., Roos, P., Benzakour, L., Cereghetti, S. et al. (2021). Adapting to the unexpected: Problematic work situations and resilience strategies in healthcare institutions during the COVID-19 pandemic's first wave. *Safety Science, 139,* 1–9. DOI 10.1016/j.ssci.2021.105277.
8. Chen, P., Zhang, H. Y., Chen, W., Wang, H., Chen, X. E. et al. (2020). Monitoring and evaluation on medical personnel's errors in removal of personal protective equipment. *Chinese Journal of Infection Control, 19(11),* 1033–1036. DOI 10.12138/j.issn.1671-9638.20207026.
9. Lu, H. T., Zhang, Q. C. (2016). Applications of deep convolutional neural network in computer vision. *Journal of Data Acquisition and Processing, 31(1),* 1–17. DOI 10.16337/j.1004-9037.2016.01.001.
10. Jiao, L. C., Zhang, F., Liu, F., Yang, S. Y., Li, L. L. et al. (2019). A survey of deep learning-based object detection. *IEEE Access, 7,* 128837–128868. DOI 10.1109/ACCESS.2019.2939201.
11. Li, L., Xu, M., Liu, H. R., Li, Y., Wang, X. F. et al. (2020). A large-scale database and a CNN model for attention-based glaucoma detection. *IEEE Transactions on Medical Imaging, 39(2),* 413–424. DOI 10.1109/TMI.2019.2927226.
12. Shi, Z. C., Dang, H., Liu, Z. C., Zhou, X. G. (2020). Detection and identification of stored-grain insects using deep learning: A more effective neural network. *IEEE Access, 8,* 163703–163714. DOI 10.1109/ACCESS.2020.3021830.
13. Zhang, Y., Wang, J., Yang, X. (2017). Real-time vehicle detection and tracking in video based on faster R-CNN. *Journal of Physics Conference Series, 887,* 1–6. DOI 10.1088/1742-6596/887/1/012068.
14. Park, M. W., Elsafty, N., Zhu, Z. H. (2015). Hardhat-wearing detection for enhancing on-site safety of construction workers. *Journal of Construction Engineering & Management, 141(9),* 1–16. DOI 10.1061/(ASCE)CO.1943-7862.0000974.
15. Shrestha, K., Shrestha, P. P., Bajracharya, D., Yfantis, E. A. (2015). Hard-hat detection for construction safety visualization. *Journal of Construction Engineering, 2015,* 1–8. DOI 10.1155/2015/721380.
16. Wu, H., Zhao, J. S. (2018). An intelligent vision-based approach for helmet identification for work safety. *Computers in Industry, 100,* 267–277. DOI 10.1016/j.compind.2018.03.037.

17. Liu, X. Y., Zhang, B. F., Fu, Y., Zhu, J. C. (2020). Detection on normalization of operating personnel dressing at contaminated sites based on deep learning. *Journal of Safety Science and Technology, 16(7),* 169–175. DOI 10.11731/j.issn.1673-193x.2020.07.027.
18. Sethi, S., Kathuria, M., Kaushik, T. (2021). A real-time integrated face mask detector to curtail spread of coronavirus. *Computer Modeling in Engineering & Sciences, 127(2),* 389–409. DOI 10.32604/cmes.2021.014478.
19. Loey, M., Manogaran, G., Taha, M. H. N., Khalifa, N. E. M. (2021). Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection. *Sustainable Cities and Society, 65,* 1–8. DOI 10.1016/j.scs.2020.102600.
20. Guo, J. D., Li, X. L. (2020). Very low-resolution object detection algorithms for electric intelligent safety supervision. *Computer Engineering and Design, 41(11),* 3188–3192. DOI 10.16208/j.issn1000-7024.2020.11.030.
21. Fu, L., Chang, Y. Q., Chen, L. S., Li, L. (2020). Key elements of donning and doffing personal protective equipment in prevention and treatment of novel coronavirus pneumonia. *Nursing Journal of Chinese People's Liberation Army, 37(2),* 1–3 + 7. DOI 10.3969/j.issn.1008-9993.2020.02.001.
22. Everingham, M., Eslami, S. M. A., van Gool, L., Williams, C. K. I., Winn, J. et al. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision, 111(1),* 98–136. DOI 10.1007/s11263-014-0733-5.
23. Bochkovskiy, A., Wang, C. Y., Liao, H. (2020). YOLOv4: Optimal speed and accuracy of object detection. https://arxiv.org/abs/2004.10934.
24. Wu, D. H., Lv, S. C., Jiang, M., Song, H. B. (2020). Using channel pruning-based YOLOv4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. *Computers and Electronics in Agriculture, 178,* 1–12. DOI 10.1016/j.compag.2020.105742.
25. Rezaei, M., Azarmi, M. (2020). DeepSOCIAL: Social distancing monitoring and infection risk assessment in COVID-19 pandemic. *Applied Sciences, 10(21),* 1–29. DOI 10.3390/app10217514.
26. Yu, Z. W., Shen, Y. G., Shen, C. K. (2021). A real-time detection approach for bridge cracks based on YOLOv4-FPM. *Automation in Construction, 122,* 1–11. DOI 10.1016/j.autcon.2020.103514.
27. PyTorch (2021). From Research to Production. https://pytorch.org/.
28. Shin, H. C., Roth, H. R., Gao, M. C., Lu, L., Xu, Z. Y. et al. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging, 35(5),* 1285–1298. DOI 10.1109/tmi.2016.2528162.
29. Jia, D., Wei, D., Socher, R., Li, L., Kai, L. et al. (2009). ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Miami, FL, USA. DOI 10.1109/CVPR.2009.5206848.
30. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *European Conference on Computer Vision (ECCV),* vol. 8693, pp. 740–755. DOI 10.1007/978-3-319-10602-1_48.
31. Redmon, J., Farhadi, A. (2018). YOLOv3: An incremental improvement. *Computer Science,* https://arxiv.org/abs/1804.02767.
32. Wang, C. Y., Liao, H. Y., Wu, Y. H., Chen, P. Y., Hsieh, G. W. et al. (2020). CSPNet: A new backbone that can enhance learning capability of CNN. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1571–1580. Seattle, WA, USA. https://arxiv.org/abs/1911.11929.
33. He, K. M., Zhang, X. Y., Ren, S. Q., Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9),* 1904–1916. DOI 10.1007/978-3-319-10578-9_23.

34. Liu, S., Qi, L., Qin, H. F., Shi, J. P., Jia, J. Y. (2018). Path aggregation network for instance segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8759–8768. The Calvin L. Rampton Salt Palace Convention Center, Salt Lake City, Utah. https://arxiv.org/abs/1803.01534.
35. Ren, S. Q., He, K. M., Girshick, R., Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6),* 1137–1149. DOI 10.1109/TPAMI.2016.2577031.
36. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. et al. (2016). SSD: Single shot multibox detector. *European Conference on Computer Vision (ECCV),* vol. 9905, pp. 21–37. DOI 10.1007/978-3-319-46448-0_2.