Computer Modeling in
Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2023.026933

ARTICLE

ISHD: Intelligent Standing Human Detection of Video Surveillance for the Smart Examination Environment

Wu Song[1], **Yayuan Tang**[2,3,*], **Wenxue Tan**[1] **and Sheng Ren**[1]

[1]School of Computer and Electrical Engineering, Hunan University of Arts and Sciences, Changde, 415000, China

[2]School of Information Engineering, Hunan University of Science and Technology, Yongzhou, 425119, China

[3]School of Computer Science and Engineering, Central South University, Changsha, 410083, China

*Corresponding Author: Yayuan Tang. Email: tangyayuan@huse.edu.cn

Received: 05 October 2022 Accepted: 04 January 2023

ABSTRACT

In the environment of smart examination rooms, it is important to quickly and accurately detect abnormal behavior (human standing) for the construction of a smart campus. Based on deep learning, we propose an intelligent standing human detection (ISHD) method based on an improved single shot multibox detector to detect the target of standing human posture in the scene frame of exam room video surveillance at a specific examination stage. ISHD combines the MobileNet network in a single shot multibox detector network, improves the posture feature extractor of a standing person, merges prior knowledge, and introduces transfer learning in the training strategy, which greatly reduces the computation amount, improves the detection accuracy, and reduces the training difficulty. The experiment proves that the model proposed in this paper has a better detection ability for the small and medium-sized standing human body posture in video test scenes on the EMV-2 dataset.

KEYWORDS

Deep learning; object detection; video surveillance of exam room; smart examination environment

1 Introduction

Generally, standardized tests are divided into several stages (empty examination room, exam preparation and closing, candidates' admission, examinations, candidates leaving, etc.). In the automatic invigilation system, the main research is conducted in the middle stage. In the middle stage of the examinations, with the exception of the invigilator, other personnel with standing postures in the surveillance video are generally in abnormal circumstances. Therefore, it is important to detect the standing human posture in the scene frame of exam-room video surveillance for further research and realization of an automatic invigilation system and smart examination rooms, which is also an important part of the smart campus.

With the rapid development of cloud computing, IoT, and artificial intelligence technologies [1], cloud video surveillance (CVS) has attracted considerable interest [2,3], video surveillance systems have become an important part of security systems in various industries, and it is a comprehensive system with strong preventive capability [4]. In the examination supervision scenario, video surveillance provides indispensable research data, which provides a reliable guarantee for subsequent image and video research in various examination-based scenarios. This paper focuses on the object detection model based on the examination room video. The main claim of video abnormal behavior detection is to monitor the abnormal behavior of certain objects appearing in the video in real time. Object detection is a very mature technology and can fulfill this requirement very well. In recent years, object detection methods based on deep learning have rapidly developed and are increasingly linked to IoT data [5–7], empowering sensing data for infrastructures and information modeling for smart cities [8], and video surveillance data in smart examination environment are one type of IoT data. The method of object detection based on deep learning has performed better than the traditional computer vision method [9,10] and can be divided into two categories [11,12]. The first category is a two-stage detection algorithm: in the first stage, the region proposals are generated, and in the second stage, these proposals are put into classification and location neural networks [13–15]. Typical representations of such algorithms are R-CNN, Fast R-CNN, and Faster R-CNN. The second category is a one-stage detection algorithm, which does not require candidate region stages and directly generates the category probability and position coordinate values of objects [16–18]. Typical algorithms are YOLO and SSD (Single Shot MultiBox Detector).

The R-CNN model proposed by Girshick et al. [19] uses selective search [20], instead of the sliding window method, to generate approximately 2000 candidate regions and then puts these into a CNN, which greatly improves the detection speed and accuracy. Girshick [21] proposed Fast R-CNN, borrowing the concept of the SPP layer [22] and proposing the ROI pooling layer to share the feature extraction layer and to simultaneously output the fully connected layer. Fast R-CNN is 9 times faster than R-CNN training. The proposed Faster R-CNN [23] used the RPN (Region Proposal Network), instead of the selective search algorithm, to make the object detection algorithm end-to-end, which makes the detection speed reach 5 fps and the detection accuracy increase to 73.2%. YOLO proposed by Redmon et al. [24] converts the detection task to a regression task, which increases the detection speed to 45 fps. Liu et al. [25] followed the idea of YOLO, combining the anchor mechanism with Faster R-CNN, and the SSD network was proposed, which combines the regression idea in YOLO and the anchor mechanism in Faster-RCNN, and uses multi-scale regions at each position of the full map for regression, which has great speed and high detection accuracy. However, improvement is still needed in the detection of small objects [26–28]. In recent years, Kim et al. [29] introduced a simple and effective data enhancement method Mix/UnMix (MUM) for SSOD (semi-supervised target detection) framework of blended image block unmixing into feature blocks, which can significantly improve the performance of SSOD method. Alairaji et al. [30] determined students' cheating behavior by detecting their face and hand movements in the examination room surveillance video.

The emergence of the deep learning method makes the detection task increasingly accurate, but all methods have different focuses. When faced with different application scenarios and different datasets [31–34], the model still needs to be further optimized and trained to better meet different detection tasks [35–37]. In this paper, the deep learning method is applied to the examination scene, and the SSD model is improved according to the morphological characteristics of the standing human posture in the examination scene. The network model proposed in this paper retains the advantages of both SSD and MobileNet networks, which simultaneously improve the detection speed and detection accuracy. The network model also reduces the difficulty of end-to-end training modes and simple network structures. By reasonably selecting multiple feature maps of different sizes and fusing multiple antecedent feature maps, the representation and detection of small and medium-sized objects in specific scenes of examination room video surveillance can be effectively improved. Merging a priori knowledge of standing humans reduces redundant calculations and improves the accuracy of target detection, which effectively mitigates the problem of insufficient exam proctoring datasets by introducing migration learning. The experiment shows that the ISHD model proposed in this paper has a stronger detection ability for small and medium-sized objects, such as the standing human posture in the examination scene. This process is fully automated and forms a very important part of the smart examination environment.

The contributions of this paper are presented as follows:

ISHD replaces the baseline network VGG16 with a lighter MobileNet network, and its new detection layer adopts the same depth-level separable convolutional layer [38] as MobileNet, which greatly reduces the network parameters and training difficulty while maintaining the consistency of the network.

In the process of selecting the size of the detection layer feature map, the ISHD model moves the detection layer forward because the standing human body in the examination room has a smaller picture ratio to reduce the receptive field of the first feature map layer. In addition, by reasonably selecting multiple feature maps of different sizes and fusing multiple front feature maps, the ISHD model can effectively improve the expression and detection ability of small and medium-sized objects with the specific scene of exam-room video surveillance.

In the process of generating the a priori boxes, the ISHD model removes the disproportionate a priori boxes according to the feature of the aspect ratio of the standing human body in the exam room, especially the aspect ratio of the sitting body, which reduces the resulting redundant calculations and improves the accuracy of object detection.

The training process of the ISHD network model adopts a transfer learning strategy. Through learning pedestrian characteristics and fine-tuning, the expression and detection ability of the standing human body in the examination room can be specifically improved, effectively solving the problem of insufficient data in the video dataset in the exam-room video surveillance.

The content of this paper is arranged as follows: Section 1 summarizes the article; Section 2 introduces the ISHD object detection model proposed in this paper; Section 3 introduces the optimization strategy of the ISHD model; Section 4 conducts experiments and analysis; and Section 5 presents the summary and prospects of this paper.

2 ISHD Object Detection Model

2.1 ISHD Network Model Based on the Scene Features of Exam-Room Surveillance

The SSD is a classical object detection neural network that uses VGG16 as a baseline network and has multiple convolutional layers behind the baseline network to obtain different feature maps for object detection [39]. As a classical convolutional neural network, VGGNet has a traditional network structure, which is mainly improved for AlexNet in terms of depth and convolution kernel size. Its excellent performance benefits from many parameters and relies on numerous sample data. The use of VGG as the baseline network leads to many SSD detection network parameters, numerous computations, a slow training process, and high requirements on samples and hardware [40].

To mitigate the shortcomings of VGGNet and expand lightweight application scenarios, such as mobile terminals, Google has proposed a compact and efficient convolutional neural network model MobileNet [41]. Compared with VGGNet, MobileNet's parameter quantity and calculation amount are greatly reduced. Therefore, in this paper, the improved network changes the baseline of the SSD.

Then, according to the characteristics of test personnel, an a priori box is generated for the specific posture of test personnel to improve the detection accuracy of the network for test personnel.

First, this paper replaced VGGNet with MobileNet and recorded the model as ISHD_v0, the initial version of ISHD. The network structure of model is shown in Fig. 1. As the scale of the input image changes to 300 × 300, the scale of the feature map generated by the convolution layer in the MobileNet network changes to 150 × 150, 75 × 75, 38 × 38, 19 × 19, and 10 × 10. Similar to SSD, ISHD_v0 adds 8 convolutional layers after the last convolutional layer Conv 14. In this paper, 8 new convolutional layers are implemented by 4 depthwise separable convolution units to enhance the network consistency under the condition that the number of newly added convolutional layers remains unchanged (shown in Fig. 1, where Conv refers to the convolutional layer and *Dw* refers to the depthwise separable convolution layer).

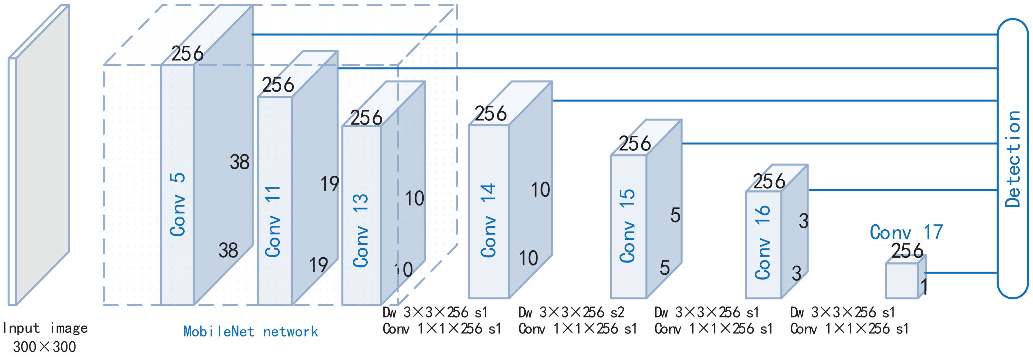


Figure 1: The Network structure of ISHD_v0

To improve the ISHD_v0 network's ability to detect small and medium-sized targets in the examination scene (detection of the standing human position in the exam room mainly focuses on the detection of small and medium-sized objects, as shown in Fig. 2), this paper adjusts the ISHD_v0 structure and marks the improved model structure as ISHD_v1.

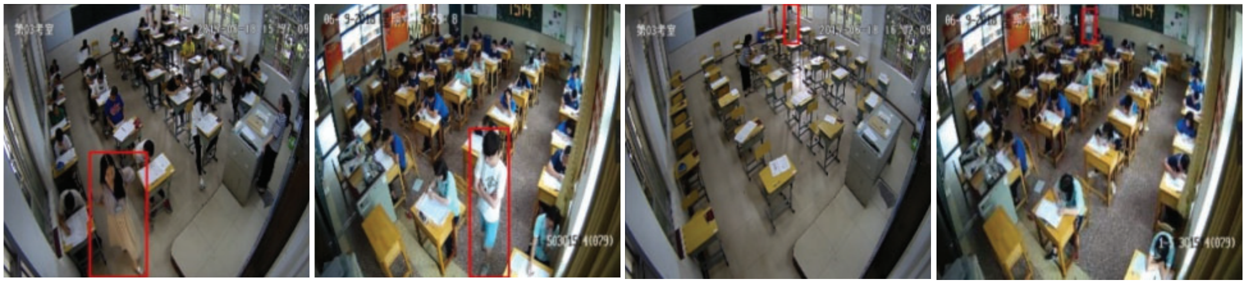


Figure 2: Examples of standing human in an exam room

ISHD_v1 (shown in Fig. 3) retains the output features of Conv3, Conv5, and Conv11, which can achieve feature information about small targets. In addition, since large targets are not the focus of examination room monitoring, ISHD_v1 sets the deep convolution kernel in Conv16 in ISHD_v0 as 5 × 5 and directly reduces the 5 × 5 feature map generated by Conv 15 to a 1 × 1 feature map.

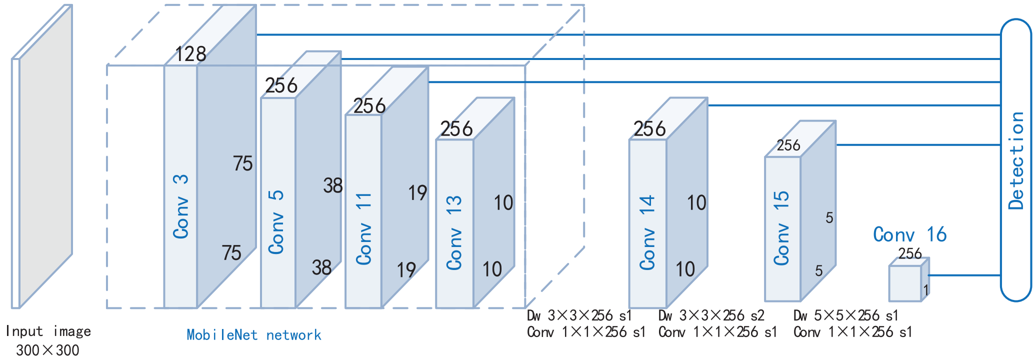


Figure 3: Network structure of ISHD_v1

Since the object size of the video in exam-room surveillance varies within a certain range, to improve the detection ability of all objects in a range from medium to small, our paper fuses the extracted multiscale features to enhance the expression ability of the small to medium-sized object features. The final structure of the ISHD model is shown in Fig. 4. On the basis of the original 6 scale feature maps, ISHD fused the features of the middle three scales in pairs, and the newly generated 2 fusion feature maps were also used for detection.

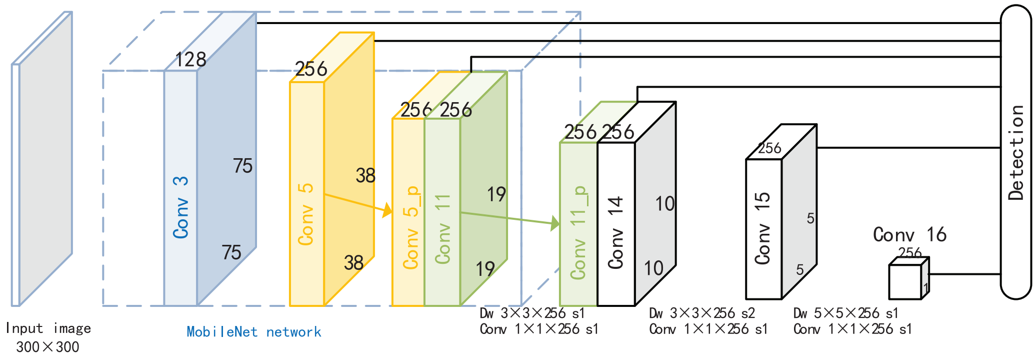


Figure 4: Network structure of the ISHD

ISHD applies pooling operations to fuse feature maps of different scales for detection. Table 1 displays the details of eight feature maps that are selected from different sources and that are divided into six groups according to the size of the feature map. The feature maps f^1_{conv}, f^2_{conv}, f^3_{conv}, f^4_{conv}, f^5_{conv}, f^6_{conv} generated by Conv3, Conv5, Conv11, Conv14, Conv15, respectively. The scales of each map are illustrated in Fig. 4.

Table 1: Feature map of ISHD

Group name	Size of the output feature map	Feature name	Layer name
1	75 × 75	f^1_{conv}	Conv3
2	38 × 38	f^2_{conv}	Conv5
3	19 × 19	f^3_{conv}	Conv11
		$f^{2\_p\_3}_{conv}$	Conv5_p + Conv 11

(Continued)

Table 1 (continued)

Group name	Size of the output feature map	Feature name	Layer name
4	10 × 10	f_{conv}^{4} $f_{conv}^{3\_p\_4}$	Conv14 Conv11_p + Conv14
5	5 × 5	f_{conv}^{5}	Conv15
6	1 × 1	f_{conv}^{6}	Conv16

The feature maps generated by Conv5 and Conv11, Conv11, and Conv14 are merged in ISHD. Taking the feature maps of Conv5 and Conv11 as an example, since the scales of f_{conv}^{2} and f_{conv}^{3} are 38 × 38 and 19 × 19, respectively, the maximum pooling operation is performed on f_{conv}^{2}, which is denoted as Conv5_p. Then, the feature map $f_{conv}^{2\_p}$ generated by Conv5_p is 19 × 19, which is consistent with the size of f_{conv}^{3}. Since the number of feature maps generated by each layer after the Conv 5 layer is 256, $f_{conv}^{2\_p}$ is also composed of 256 19 × 19 feature maps. Through the global average pooling operation, the average values of the feature maps of $f_{conv}^{2\_p}$ and f_{conv}^{3} are sequentially calculated, and the fusion feature map is obtained as shown in Eq. (1):

$$f_{conv}^{2\_p\_3}(i) = glo\_ave\_\_pool(f_{conv}^{2\_p}(i), f_{conv}^{3}(i))\ (i \in [1, 256]) \tag{1}$$

where $f_{conv}^{2\_p}(i)$ and $f_{conv}^{3}(i)$ are the i-th (i ∈ [1,256]) feature maps generated by Conv5_p and Conv11, respectively; *glo_ave__pool* is the average pooling operation; and $f_{conv}^{2\_p\_3}(i)$ is the i-th newly generated feature map. Similarly, the fusion characteristics of the feature maps of Conv11 and Conv14 are shown in Eq. (2).

$$f_{conv}^{3\_p\_4}(i) = glo\_ave\_\_pool\left(f_{conv}^{3\_p}(i), f_{conv}^{4}(i)\right) \quad (i \in [1, 256]) \tag{2}$$

ISHD selects 6 groups of feature maps for detection. A series of a priori boxes are generated at the center of each feature map. The specific size of the map is responsible for monitoring objects of a certain size in the image. Therefore, the scale ratio of a priori boxes on each feature graph is set as shown in Eq. (3).

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{m - 2}(k - 1), \quad k \in [2, m] \tag{3}$$

where m represents the number of feature map groups, k is the group of the feature map, and s_{min} and s_{max} are set to 0.2 and 0.9, respectively. The scale ratio of the a priori boxes of the first layer feature map f_{conv}^{1} is individually reset to $s_{min}/2$. Combining s_k with the actual size of the image, the length of a priori box in each group of feature maps is illustrated in Table 2.

Table 2: Side length of the a priori frame in each feature map

Group name	Size of the output feature map	Feature name	Side length
1	75 × 75	f_{conv}^{1}	30
2	38 × 38	f_{conv}^{2}	60

(Continued)

Table 2 (continued)

Group name	Size of the output feature map	Feature name	Side length
3	19 × 19	f_{conv}^{3} $f_{conv}^{2\_p\_3}$	111
4	10 × 10	f_{conv}^{4} $f_{conv}^{3\_p\_4}$	162
5	5 × 5	f_{conv}^{5}	213
6	1 × 1	f_{conv}^{6}	264

2.2 *Setting for the Loss Function*

The objective function of the ISHD network consists of two parts, as shown in Eq. (4). The first part $L_{conf}(x, c)$ is the calculation of the default box and its target category and the confidence of the target category. The second part L_{loc} calculates the regression results for the corresponding location. The confidence is computed by the Softmax loss function, and the position regression is calculated by the Smooth L1 loss function.

$$L(x, c, l, g) = \frac{1}{N}\left(L_{conf}(x, c)\right) + \alpha L_{loc}(x, l, g) \tag{4}$$

N represents the number of positive samples and the calculation of L_{conf} and L_{loc} is established in Eq. (5), α is the hyperparameter that adjusts the ratio between confidence loss and location loss.

$$\begin{cases} L_{conf}(x, c) = -\sum_{i\in Pos}^{N} x_{ij}^{p}\log\left(\hat{c}_i^p\right) - \sum_{i\in Neg}\log\left(\hat{c}_i^0\right) \\ L_{loc}(x, l, g) = \sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,w,h\}} x_{ij}^{k}\text{smooth}_{\text{L1}}\left(l_i^m - \hat{g}_j^m\right) \end{cases} \tag{5}$$

Pos is the set of positions, Neg is the set of negative samples, and $x_{ij}^{p}\log\left(\hat{c}_i^p\right)$ is the matching probability of the i-th prediction box and j-th real box with respect to category p. The higher the matching probability is, the smaller the loss is. $\hat{c}_i^p$ indicates that there is no target object in the prediction box. The higher the background probability is, the smaller the loss is. By the calculation of Softmax, $x_{ij}^{k} \in \{0, 1\}$ indicates whether the i-th prediction box and j-th true box match on category k, l_i^m is a prediction box, and $\hat{g}_j^m$ is a real box, the four coordinate values of each anchor are $d_i^{cx}, d_i^{cy}, d_i^{w}, d_i^{h}$. The specific calculation is shown in Eq. (6).

$$\begin{cases} \hat{g}_j^{cx} = \left(g_j^{cx} - d_i^{cx}\right)/d_i^{w} \\ \hat{g}_j^{cy} = \left(g_j^{cy} - d_i^{cy}\right)/d_i^{h} \\ \hat{g}_j^{\omega} = \log\left(\dfrac{g_j^{\omega}}{d_i^{\omega}}\right) \\ \hat{g}_j^{h} = log\left(\dfrac{g_j^{h}}{d_i^{h}}\right) \\ \hat{c}_i^p = \dfrac{\exp\left(c_i^p\right)}{\sum_p \exp\left(c_i^p\right)} \end{cases} \tag{6}$$

3 Optimization Strategy the of ISHD Model

3.1 Selection of the a Priori Box in the ISHD Network Model

The examination personnel in the video generally have two postures: sitting and standing. For a human with a standing posture, the outer frame is generally thin and high. Unless the picture is rotated, the wide and flat state will hardly appear. Therefore, this paper sets the a priori frame aspect ratio to a_r ∈ {1,1/2,1/3}. Then, the width and height of the a priori box are determined as shown in Eq. (7):

$$\begin{cases} width_k^a = s_k\sqrt{a_r} \\ height_k^a = s_k/\sqrt{a_r} \end{cases} \tag{7}$$

When $a_r = 1$, the additional scale ratio $s_k' = \sqrt{s_k s_{k+1}}$ is specified. A total of 4 a prior boxes are set on each feature map point. The center position of a priori box is $\left(\frac{i+0.5}{|f_k|}, \frac{i+0.5}{|f_k|}\right)$, where $|f_k|$ represents the size of the kth feature map. In addition, for the 5 × 5 and 1 × 1 feature maps, the corresponding a priori box scale far exceeds the object scale. Therefore, for these two layers, only a priori box is set. Therefore, the total number of prior boxes is: $75 \times 75 \times 4 + 38 \times 38 \times 4 + 19 \times 19 \times 4 \times 2 + 10 \times 10 \times 4 \times 2 + 5 \times 5 \times 1 + 1 \times 1 \times 1 = 22,500 + 5,776 + 2,888 + 800 + 25 + 1 = 31,990$.

3.2 Transfer Training Strategy of the ISHD Network Model

When training data are limited, model performance is often poor [42–44], and some scholars use GAN to alleviate this problem [45]. For the limited video training samples of the examination room, this paper adopts the transfer learning strategy to fine-tune the ISHD network on the relevant dataset after training. This paper adopts two transfer learning strategies, as shown in Fig. 5. In the first strategy, MobileNet adopts a model trained by ImageNet to conduct transfer learning. After MobileNet is connected to ISHD, fine-tuning is carried out through the large dataset to make ISHD learn more complex characteristics of people in street view, and then the micro-adjusted ISHD is transferred to the limited video dataset. In the second transfer learning strategy, the MobileNet framework is directly connected to ISHD, and the pedestrian characteristics are directly learned by using the Caltech Pedestrians Datasets. The trained network is then transferred to the small dataset of video surveillance in the exam room (EMV-2). The experimental part of this paper will compare and analyze the performance of the two transfer modes.

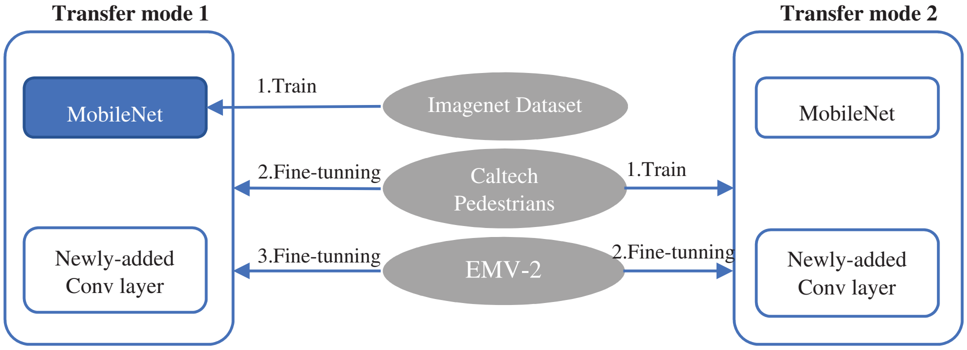


Figure 5: Two transfer modes adopted in ISHD

The ISHD models in the two training modes are recorded as ISHD1 and ISHD2. The intermediate state models in the two training modes, namely, the fine-tuning or training of the Caltech Pedestrians

Datasets, are MISHD1 and MISHD2. These models are fine-tuned by the real exam-room dataset EMV-2 to obtain the final ISHD1 and ISHD2 models.

4 Experimental Results and Analysis

4.1 Dataset for Experiments

Targeting the study of standing posture human detection in the exam room, the dataset is derived from the real standardized exam-room surveillance video. We selected a two-way, 3D convolutional neural network [46] to classify the scenes of the exam-room surveillance video and established the test video dataset EMV-1 according to 6 categories: category 1 (empty examination room), category 2 (exam preparation and closing), category 3 (distribution and withdrawing of papers), category 4 (candidates' admission), category 5 (candidates' leaving) and category 6 (examinations).

The dataset utilized in this paper is collected from the EMV-1 test video dataset. First, the source video frame is scaled to 300 × 300, and then positive and negative samples are collected. The samples in this paper were only collected from categories 2, 3 and 6 of the EMV-1 dataset, including 1666 sample fragments. According to whether there are standing posture personnel and the action range of standing posture personnel in the sample fragments, these three types of sample fragments are divided into the following three types: (1) nonstanding posture personnel; (2) stationary standing posture personnel; and (3) moving standing posture personnel. Fifty video samples of these three types were selected from 1666 sample fragments in this paper. That is, 150 sample fragments were selected as the collection source. For sample segments (1) and (2), only the first frame of their images is retained as the source image, while for sample segments (3), all the frames of images are retained as the source image. A total of 50 frame images of sample fragments of people without standing posture are reserved, and $50 + 50 \times 72$ (frames/sample fragment) of sample fragments of people with standing posture = 3,650. In the third sample segment, not every frame contains moving posture personnel. A total of 854 images containing moving posture personnel were further selected, and 1,006 positive samples were obtained. Through random sampling in these 854 images, the sampling boxes with positive sample IoU values less than 0.5 in the current image are retained as negative samples, and the remaining $3{,}650-854 + 50 = 2{,}846$ sample frames without posture personnel are randomly sampled to generate negative samples. In addition, some images were randomly selected from the $854 + 2{,}846 = 3{,}700$ images to be labeled with artificial negative samples. Students with sitting posture, desks and chairs were labeled negative samples, and 0.5 negative samples with artificial negative samples larger than the IoU value were randomly generated. Ultimately, a total of 10,000 negative samples were obtained. In this paper, these 3,700 300 × 300 images and positive and negative samples are labeled according to the format of VOC2007 and named the EMV-2 dataset (Examination Monitoring Video). In addition, all the experiments in this section are implemented on a common operating platform (an Intel Core i9-10940x CPU @ 3.30 GHz and an NVIDIA RTX 2080ti 11 G GPU), using Python for encoding.

4.2 Analysis of the Transfer Strategy Effect

In this paper, two transfer strategies are adopted to train the ISHD network. Under the two training modes, the Caltech Pedestrian Dataset is fine-tuned to obtain MISHD1 and MISHD2, and then MISHD1 and MISHD2 are fine-tuned to obtain the final ISHD1 and ISHD2, respectively, on the EMV-2 test field dataset. Miss Rate-FPPI curves of MISHD1 and MISHD2 in the Caltech Pedestrian Dataset are shown in Fig. 6, which also compares various traditional methods, such as VJ [47], HOG [48], LatSVM-V1 [49], MultiFtr+CSS [50], and FPDW [51]. This figure shows that the curves of MISHD1 and MISHD2 are much lower than those of other traditional methods, which means that MISHD1 and MISHD2 are greatly improved compared with other traditional methods.

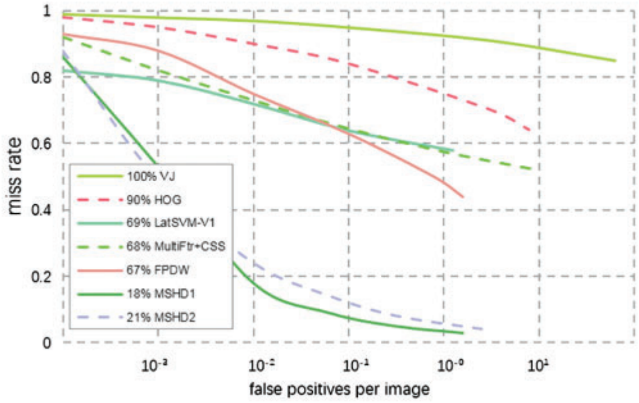


Figure 6: Comparison of miss rate-FPPI in the test set of the caltech pedestrian dataset

MISHD1 and MISHD2 are transferred to the EMV-2 test dataset for fine-tuning to obtain the final network ISHD1 and ISHD2. The loss of the fine-tuning process of both on the EMV-2 test dataset is shown in Fig. 7. ISHD inherits the advantages of good performance of the SSD model. Both converge and stabilize within 10,000 steps due to the reduction in parameter size. Compared with ISHD1, ISHD2 converges to a smaller loss value on the EMV-2 dataset.

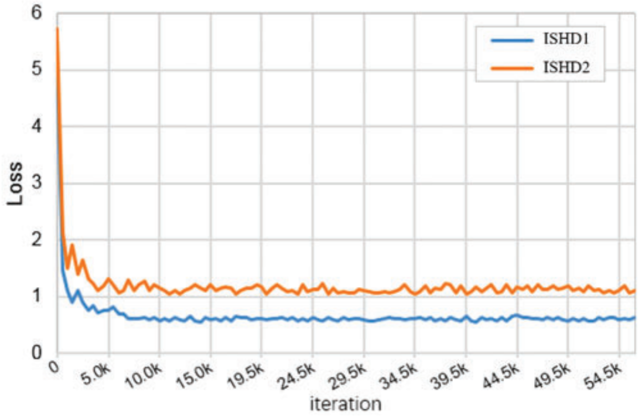


Figure 7: Fine-tuning loss curves of ISHD1 and ISHD2 in EMV-2 datasets

Fig. 8 displays the precision-recall curves of ISHD1 and ISHD2, and neither has reached a large recall rate. The corresponding accuracy is high and stable in the early stage. After the recall rate exceeds 0.6, a sharp drop occurs. The sharp drop in ISHD1 was slightly slower than that in ISHD2. The average accuracy (AP) of ISHD1 on the EMV-2 dataset was also slightly better than that of ISHD2.

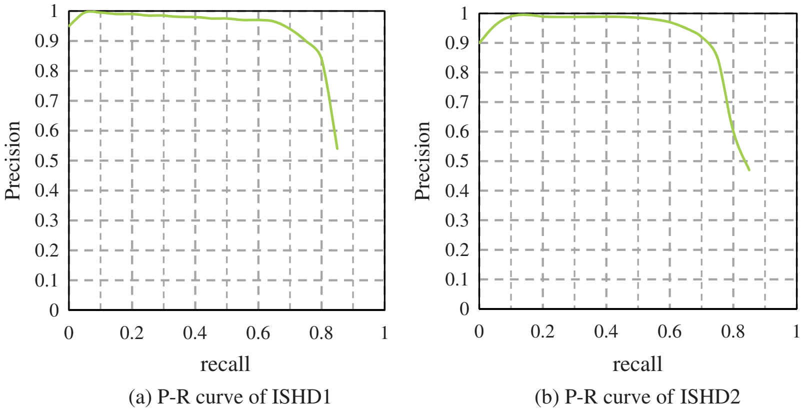


Figure 8: Precision-recall curves of ISHD1 and ISHD2

As shown in Table 3, the final ISHD1 and ISHD2 achieved average precisions of 77.23% and 74.21%, respectively. Table 3 compares the APs of ISHD1 and ISHD2 on the Caltech Pedestrian Dataset. Both achieved an average accuracy greater than 85% and are consistent with the trend on the EMV-2 pedestrian test set. The ISHD1 result (88.30%) was slightly better than the ISHD2 result (84.30%). The training of ISHD1 and ISHD2 consists of two steps. The second step of the two models is fine-tuning on the EMV-2 dataset. The strategy used in the first step is slightly different. The former is fine-tuned on the Caltech Pedestrian Dataset, while the latter is trained. The time difference between the two models is mainly reflected in the first step.

Table 3: Average accuracy of ISHD1 and ISHD2 in different datasets

Data set	ISHD1	ISHD2
Caltech pedestrian dataset	88.30	84.30
EMV-2 dataset	77.23	74.21

The comparison of the single iteration time between the MISHD1 model and the MISHD2 model is shown in Table 4. The difference between the two iterations is large. MISHD2 trains the entire ISHD network structure. Each layer participates in the calculation, while ISHD1 fine-tunes the pretrained MobileNet on ImageNet and only trains the parameters of the full connection layer of the last layer, with a sharp decrease in the calculation amount, increasing the training speed.

Table 4: Video frame data distribution in the examination room

Model	Single iteration time (ms)
MISHD1	2467
MISHD2	437

In terms of the above comparison of the combined time and detection effect, ISHD1 obtains higher detection accuracy with a shorter training time than ISHD2. Therefore, this paper implements the first transfer strategy employed by ISHD1 for subsequent experiments.

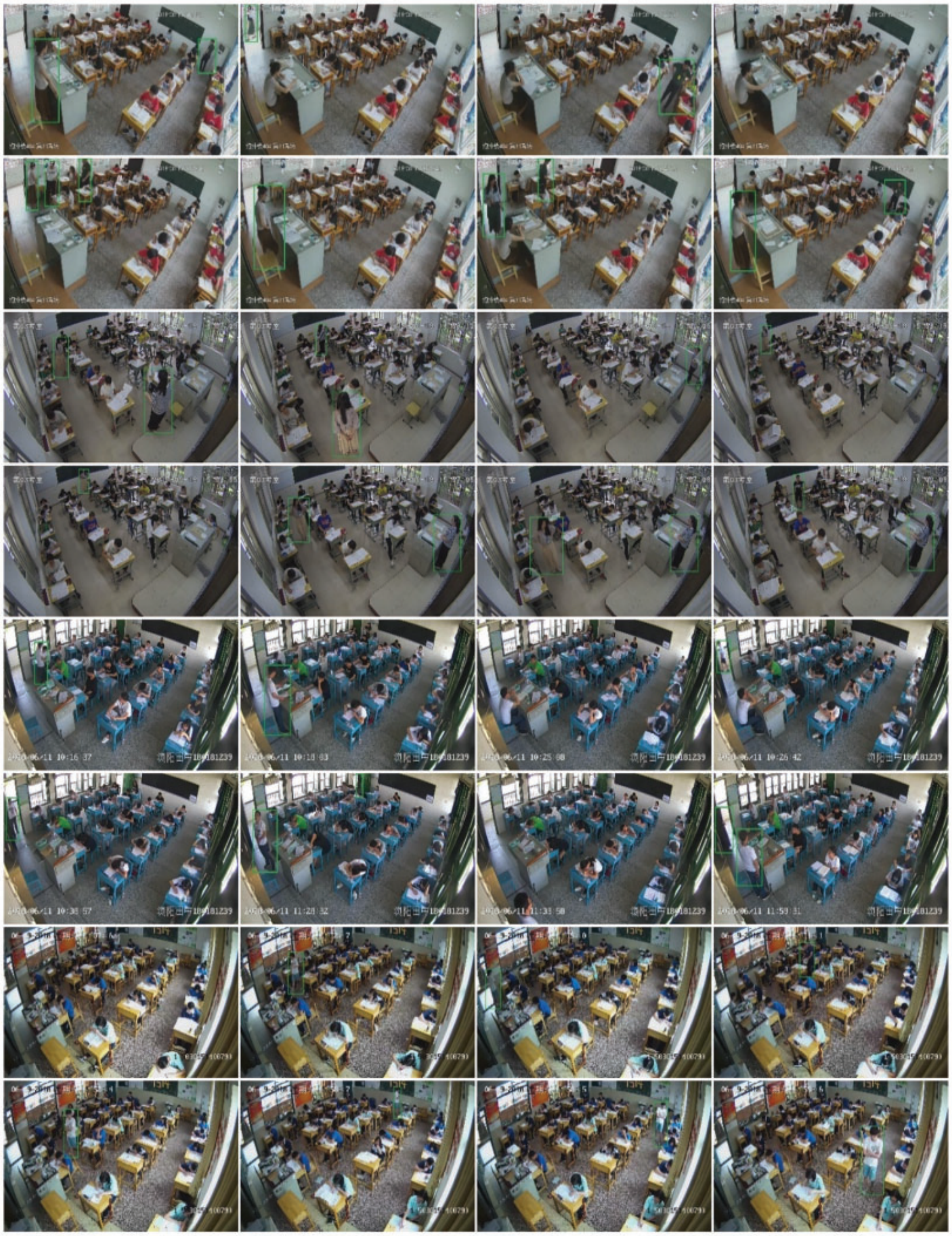
4.3 Comparison with Existing Object Detection Methods

To prove the validity and efficiency of the model, this paper compares ISHD with multiple object detection methods, including HOG-ALBP+LatSVM, UDN, ACF, CrossTalk, HikSVM, faster R-CNN, SSD, ISHD_v0, and ISHD_v1, the above comparison models are all commonly used object detection models nowadays. Table 5 displays the average accuracy of the detection results for each method on the video dataset of the examination room and the average detection time of a single picture. Our proposed ISHD achieves the highest average detection accuracy of 77.23% among these methods, which proves the validity of the test model designed in this paper. In terms of the detection time of a single image, the traditional method generally takes hundreds of milliseconds. The improved ISHD series replaces the baseline network VGG with MobileNet, which has a smaller structure and fewer parameters, thus reducing the detection rate to less than 100 ms. Compared with ISHD_0 and ISHD_1, the number of detection features of ISHD increases, resulting in a more time-consuming detection process. Interestingly, the detection rate of a single image is still only 95 ms, which is much lower than that of other methods. The comprehensive detection of the average accuracy and the time-consuming test results prove that the ISHD method proposed in this paper has a strong detection capacity for standing personnel and can quickly realize accurate retrieval of standing personnel and achieve ideal results.

Table 5: Comparison of the test results of different models on the video dataset in the examination room

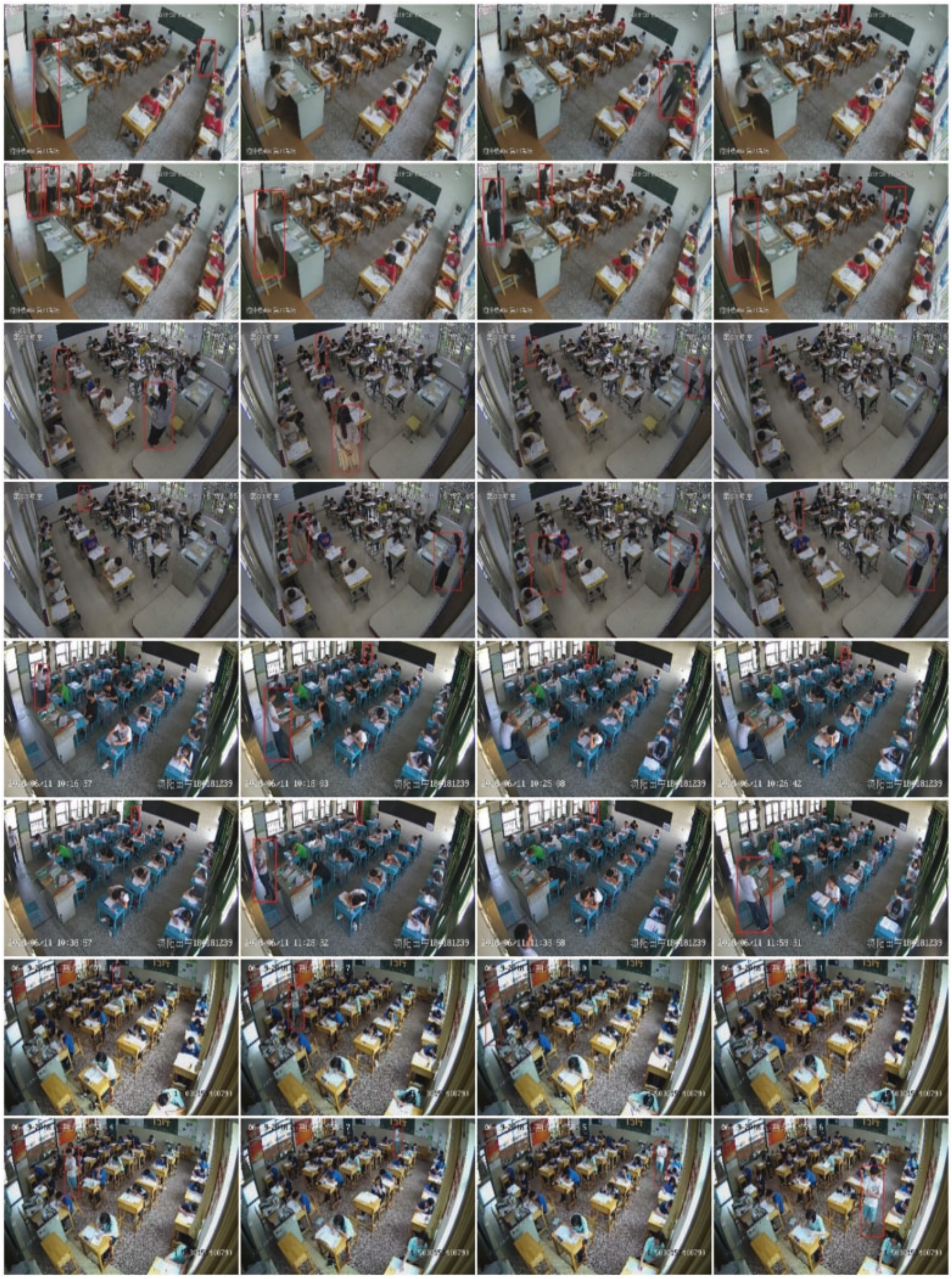
Model	AP (%)	Testing time (ms)
HOG	31.24	457
CrossTalk	45.12	567
HikSVM	34.40	445
YOLO	68.71	45
Faster R-CNN	72.14	185
SSD	72.25	163
ISHD_0	73.56	89
ISHD_1	75.88	87
ISHD	77.23	95

Fig. 9 compares the results of the SSD detection network derived from the VGG network and the ISHD network proposed in this paper on the EMV-2 test field dataset. According to the test results of the SSD network in Fig. 9a, the SSD has a good detection effect on the standing posture personnel at the front end of the classroom but has missed several people standing at the end of the classroom. Compared with the standing staff at the front end of the classroom, the proportion of people standing at the end of the classroom is small in the picture, and the SSD is not sensitive to the detection of small objects. Conversely, in the improved ISHD network, which is shown in Fig. 9b, ISHD can detect the invigilator at the end of the classroom. Regardless of whether the standing personnel stands at the end of the diagonal direction or in front of the camera, ISHD performs perfect detection without missing, completing accurate detection of objects of various scales. However, ISHD still has some defects. For example, the teacher who is near the entrance of the examination room at the edge of the screen cannot be correctly distinguished from the door due to the backlit shooting angle to the teacher, which causes the dark area. However, other normal brightness areas do not exhibit this problem. Overall, ISHD has a good detection effect on the EMV-2 dataset.



(a) SSD test results

Figure 9: (Continued)



(b) ISHD test results

Figure 9: Comparison of SSD and ISHD on EMV-2 test data

5 Conclusion and Prospect

This paper focuses on the detection of standing posture personnel in the video frame of the test. We improve the SSD detection network to obtain the ISHD network, which replaces the baseline network with a lighter MobileNet network and simultaneously improves the detection speed and detection accuracy. This paper reasonably selects multiple feature maps of different sizes and fuses multiple antecedent feature maps to effectively improve the representation and detection of small and medium-sized objects in specific scenes of examination room video surveillance. This paper also optimizes the ISHD network in the prior frame for the specific aspect ratio, aspect ratio, etc., of the standing posture personnel in the examination room so that ISHD has stronger abstract expression and detection ability for the characteristics of small and medium-sized targets. Under the limited test field dataset, the ISHD network uses the transfer training strategy to train, and after learning many pedestrian characteristics, it can strengthen the learning of the characteristics of the standing posture of the examination room to realize the test stand personnel under the limited test video training dataset. Compared with other multimodel methods, our model has better detection ability for small and medium-sized targets, such as standing poses in the examination room video, which is important for the construction of a smart examination environment.

Although the ISHD detection network proposed in this paper has certain effects on the detection of the standing position in the examination room, it still can be improved. The next step is to improve the detection ability of small and medium-sized targets and occluded personnel. By designing appropriate data enhancement methods, the sample data volume of the examination room can be improved, and the network model and parameters can be further tuned to increase the detection rate and realize real-time detection.

Funding Statement: This work was supported by the Natural Science Foundation of China 62102147; National Science Foundation of Hunan Province 2022JJ30424, 2022JJ50253, and 2022JJ30275; Scientific Research Project of Hunan Provincial Department of Education 21B0616 and 21B0738; Hunan University of Arts and Sciences Ph.D. Start-Up Project BSQD02, 20BSQD13; and the Construct Program of Applied Characteristic Discipline in Hunan University of Science and Engineering.

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

References

1. Xu, Y., Liu, Z., Zhang, C., Ren, J., Zhang, Y. et al. (2021). Blockchain-based trustworthy energy dispatching approach for high renewable energy penetrated power systems. *IEEE Internet of Things Journal, 9(12),* 10036–10047.
2. Zhou, X., Xu, X., Liang, W., Zeng, Z., Yan, Z. (2021). Deep-learning-enhanced multitarget detection for end-edge–cloud surveillance in smart IoT. *IEEE Internet of Things Journal, 8(16),* 12588–12596. https://doi.org/10.1109/JIOT.2021.3077449
3. Xu, Y., Yan, X., Wu, Y., Hu, Y., Liang, W. et al. (2021). Hierarchical bidirectional RNN for safety-enhanced B5G heterogeneous networks. *IEEE Transactions on Network Science and Engineering, 8(4),* 2946–2957. https://doi.org/10.1109/TNSE.2021.3055762
4. Ullah, F. U. M., Muhammad, K., Haq, I. U., Khan, N., Heidari, A. A. et al. (2021). AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks. *IEEE Transactions on Industrial Informatics, 18(8),* 5359–5370. https://doi.org/10.1109/TII.2021.3116377

5. Yan, K., Zhou, X., Chen, J. (2022). Collaborative deep learning framework on IoT data with bidirectional NLSTM neural networks for energy consumption forecasting. *Journal of Parallel and Distributed Computing, 163(8),* 248–255. https://doi.org/10.1016/j.jpdc.2022.01.012
6. Zhou, X., Yang, X., Ma, J., Kevin, I., Wang, K. (2021). Energy efficient smart routing based on link correlation mining for wireless edge computing in IoT. *IEEE Internet of Things Journal, 9(16),* 14988–14997.
7. Xu, Y., Zhang, C., Wang, G., Qin, Z., Zeng, Q. (2020). A blockchain-enabled deduplicatable data auditing mechanism for network storage services. *IEEE Transactions on Emerging Topics in Computing, 9(3),* 1421–1432. https://doi.org/10.1109/TETC.2020.3005610
8. Shen, S., Yu, C., Zhang, K., Ci, S. (2021). Adaptive artificial intelligence for resource-constrained connected vehicles in cybertwin-driven 6G network. *IEEE Internet of Things Journal, 8(22),* 16269–16278. https://doi.org/10.1109/JIOT.2021.3101231
9. Liu, Y., Sun, P., Wergeles, N., Shang, Y. (2021). A survey and performance evaluation of deep learning methods for small object detection. *Expert Systems with Applications, 172(4),* 114602. https://doi.org/10.1016/j.eswa.2021.114602
10. Zaidi, S. S. A., Ansari, M. S., Aslam, A., Kanwal, N., Asghar, M. et al. (2022). A survey of modern deep learning based object detection models. *Digital Signal Processing, 126(11),* 103514. https://doi.org/10.1016/j.dsp.2022.103514
11. Huang, Z., Yang, S., Zhou, M., Gong, Z., Abusorrah, A. et al. (2021). Making accurate object detection at the edge: Review and new approach. *Artificial Intelligence Review, 55(3),* 2245–2274.
12. Pal, S. K., Pramanik, A., Maiti, J., Mitra, P. (2021). Deep learning in multi-object detection and tracking: State of the art. *Applied Intelligence, 51(9),* 6400–6429. https://doi.org/10.1007/s10489-021-02293-7
13. Luo, J., Yang, Z., Li, S., Wu, Y. (2021). FPCB surface defect detection: A decoupled two-stage object detection framework. *IEEE Transactions on Instrumentation and Measurement, 70,* 1–11. https://doi.org/10.1109/TIM.2021.3092510
14. Choy, S. K., Carisa, K. W., Lee, T. C., Lam, B. S., Wong, C. Y. (2021). A two-stage variational jump point detection algorithm for real estate analysis. *Land Use Policy, 111(1),* 105687. https://doi.org/10.1016/j.landusepol.2021.105687
15. Li, W., Chen, Z., Li, B., Zhang, D., Yuan, Y. (2021). HTD: Heterogeneous task decoupling for two-stage object detection. *IEEE Transactions on Image Processing, 30,* 9456–9469. https://doi.org/10.1109/TIP.2021.3126423
16. Li, Z., Sun, Y., Tian, G., Xie, L., Liu, Y. et al. (2021). A compression pipeline for one-stage object detection model. *Journal of Real-Time Image Processing, 18(6),* 1949–1962. https://doi.org/10.1007/s11554-020-01053-z
17. Zhang, H., Cloutier, R. S. (2022). Review on one-stage object detection based on deep learning. *EAI Endorsed Transactions on e-Learning, 7(23),* e5. https://doi.org/10.4108/eai.9-6-2022.174181
18. Chu, Y., Guo, J., Shan, W., Wang, Z. (2022). EfficientFCOS: An efficient one-stage object detection model based on FCOS. *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 617–622. IEEE.
19. Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587. IEEE.
20. Cheng, B., Schwing, A., Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems, 34,* 17864–17875.
21. Girshick, R. (2015). Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448. IEEE.

22. Xu, F., Wang, H., Sun, X., Fu, X. (2022). Refined marine object detector with attention-based spatial pyramid pooling networks and bidirectional feature fusion strategy. *Neural Computing and Applications, 34(17),* 1–14. https://doi.org/10.1007/s00521-022-07264-8
23. Ren, S., He, K., Girshick, R., Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems 28*.
24. Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788. IEEE.
25. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. et al. (2016). SSD: Single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds), *Lecture notes in computer science*, vol. 9905.
26. Zhou, X., Xu, X., Liang, W., Zeng, Z., Shimizu, S. et al. (2021). Intelligent small object detection for digital twin in smart manufacturing with industrial cyber-physical systems. *IEEE Transactions on Industrial Informatics, 18(2),* 1377–1386. https://doi.org/10.1109/TII.2021.3061419
27. Magalhães, S. A., Castro, L., Moreira, G., Dos Santos, F. N., Cunha, M. et al. (2021). Evaluating the single-shot multibox detector and YOLO deep learning models for the detection of tomatoes in a greenhouse. *Sensors, 21(10),* 3569. https://doi.org/10.3390/s21103569
28. Bai, D., Sun, Y., Tao, B., Tong, X., Xu, M. et al. (2022). Improved single shot multibox detector target detection method based on deep feature fusion. *Concurrency and Computation: Practice and Experience, 34(4),* e6614. https://doi.org/10.1002/cpe.6614
29. Kim, J., Jang, J., Seo, S., Jeong, J., Na, J. et al. (2022). MUM: Mix image tiles and UnMix feature tiles for semi-supervised object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14512–14521. IEEE.
30. Alairaji, R. A. M., Aljazaery, I. A., ALRikabi, H. T. S. (2022). Abnormal behavior detection of students in the examination hall from surveillance videos. In: *Advanced computational paradigms and hybrid intelligent computing*, pp. 113–125. Singapore: Springer.
31. Xu, Y., Ren, J., Zhang, Y., Zhang, C., Shen, B. et al. (2019). Blockchain empowered arbitrable data auditing scheme for network storage as a service. *IEEE Transactions on Services Computing, 13(2),* 289–300. https://doi.org/10.1109/TSC.2019.2953033
32. Yan, K. (2021). Chiller fault detection and diagnosis with anomaly detective generative adversarial network. *Building and Environment, 201(2),* 107982. https://doi.org/10.1016/j.buildenv.2021.107982
33. Zhou, X., Liang, W., Kevin, I., Wang, K., Yang, L. T. (2020). Deep correlation mining based on hierarchical hybrid networks for heterogeneous big data recommendations. *IEEE Transactions on Computational Social Systems, 8(1),* 171–178. https://doi.org/10.1109/TCSS.2020.2987846
34. Xu, Y., Zeng, Q., Wang, G., Zhang, C., Ren, J. et al. (2020). An efficient privacy-enhanced attribute-based access control mechanism. *Concurrency and Computation: Practice and Experience, 32(5),* e5556. https://doi.org/10.1002/cpe.5556
35. Xiao, B., Kang, S. C. (2021). Development of an image data set of construction machines for deep learning object detection. *Journal of Computing in Civil Engineering, 35(2),* 05020005. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000945
36. Jiao, L., Zhang, R., Liu, F., Yang, S., Hou, B. et al. (2021). New generation deep learning for video object detection: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.
37. Haris, M., Glowacz, A. (2021). Road object detection: A comparative study of deep learning-based algorithms. *Electronics, 10(16),* 1932. https://doi.org/10.3390/electronics10161932 1932.
38. Chang, L., Zhang, S., Du, H., You, Z., Wang, S. (2021). Position-aware lightweight object detectors with depthwise separable convolutions. *Journal of Real-Time Image Processing, 18(3),* 857–871. https://doi.org/10.1007/s11554-020-01027-1

39. Yan, K., Zhou, X. (2022). Chiller faults detection and diagnosis with sensor network and adaptive 1D CNN. *Digital Communications and Networks, 8(4),* 531–539. https://doi.org/10.1016/j.dcan.2022.03.023
40. Zhou, X., Liang, W., Kevin, I., Wang, K., Shimizu, S. (2019). Multi-modality behavioral influence analysis for personalized recommendations in health social media environment. *IEEE Transactions on Computational Social Systems, 6(5),* 888–897. https://doi.org/10.1109/TCSS.2019.2918285
41. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W. et al. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.
42. Yan, K., Chen, X., Zhou, X., Yan, Z., Ma, J. (2022). Physical model informed fault detection and diagnosis of air handling units based on transformer generative adversarial network. *IEEE Transactions on Industrial Informatics, 19(2),* 2192–2199.
43. Xu, Y., Zhang, C., Zeng, Q., Wang, G., Ren, J. et al. (2020). Blockchain-enabled accountability mechanism against information leakage in vertical industry services. *IEEE Transactions on Network Science and Engineering, 8(2),* 1202–1213. https://doi.org/10.1109/TNSE.2020.2976697
44. Zhou, X., Liang, W., Shimizu, S., Ma, J., Jin, Q. (2020). Siamese neural network based few-shot learning for anomaly detection in industrial cyber-physical systems. *IEEE Transactions on Industrial Informatics, 17(8),* 5790–5798. https://doi.org/10.1109/TII.2020.3047675
45. Yan, K., Chong, A., Mo, Y. (2020). Generative adversarial network for fault detection diagnosis of chillers. *Building and Environment, 172(8),* 106698. https://doi.org/10.1016/j.buildenv.2020.106698
46. Chen, Z., Yang, B., Wu, F., Ren, S., Zhong, H. (2020). TCNN: Two-way convolutional neural network for image steganalysis. In: Park, N., Sun, K., Foresti, S., Butler, K., Saxena, N. (eds), *Lecture notes of the institute for computer sciences, social informatics and telecommunications engineering*, vol. 335. Washington DC, USA: Springer, Cham, Springer International Publishing.
47. Viola, P., Jones, M. J., Snow, D. (2005). Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision, 63(2),* 153–161. https://doi.org/10.1007/s11263-005-6644-8
48. Dalal, N., Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893. https://doi.org/10.1109/CVPR.2005.177
49. Felzenszwalb, P., McAllester, D., Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE.
50. Walk, S., Majer, N., Schindler, K., Schiele, B. (2010). New features and insights for pedestrian detection. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1030–1037. IEEE.
51. Dollár, P., Belongie, S., Perona, P. (2010). The fastest pedestrian detector in the west. *British Machine Vision Conference*, Aberystwyth, UK. DBLP.