ITC 1/53
Information Technology and Control
Vol. 53 / No. 1 / 2024
pp.220-236
DOI 10.5755/j01.itc.53.1.33458

YOLOv5s-MEE: A YOLOv5-based Algorithm for Abnormal Behavior Detection in Central Control Room

Received 2023/02/21 | Accepted after revision 2023/09/02

HOW TO CITE: Yuan, P., Fan, C., Zhang, C. (2024). YOLOv5s-MEE: A YOLOv5-based Algorithm for Abnormal Behavior Detection in Central Control Room. *Information Technology and Control*, 53(1), 220-236. https://doi.org/10.5755/j01.itc.53.1.33458

YOLOv5s-MEE: A YOLOv5-based Algorithm for Abnormal Behavior Detection in Central Control Room

Ping Yuan, Chunling Fan, Chuntang Zhang

Department of Automatic and Electronic Engineering, Qingdao University of Science and Technology, Qingdao, 266061, Shandong, China; e-mails: pingyuan@mails.qust.edu.cn; chunlingfan@qust.edu.cn.; 01634@qust.edu.cn

Corresponding author: chunlingfan@qust.edu.cn

Aiming to quickly and accurately detect abnormal behaviors of workers in central control rooms, such as playing mobile phone and sleeping, an abnormal behavior detection algorithm based on improved YOLOv5 is proposed. The technique uses SRGAN to reconstruct the input image to improve the resolution and enhance the detailed information. Then, the MnasNet is introduced to replace the backbone feature extraction network of the original YOLOv5, which could achieve the lightweight of the model. Moreover, the detection accuracy of the whole network is enhanced by adding the ECA-Net attention mechanism into the feature fusion network structure of YOLOv5 and modifying the loss function as EIOU. The experimental results in the custom dataset show that compared with the original YOLOv5 algorithm, the algorithm proposed in this paper improves the detection speed to 75.50 frames/s under the condition of high detection accuracy, which meets the requirements of real-time detection. Meanwhile, compared with other mainstream behavior detection algorithms, this algorithm also shows better detection performance.

KEYWORDS: Abnormal behavior detection, YOLOv5, SRGAN, Attention mechanism, EIOU.

1. Introduction

The central control room is the carrier of information exchange between the production equipment and the central control system in factory area, which plays an essential role in industrial production [18]. For example, in the mixing workshop of the rubber industry, it is of necessity to coordinate the control of various proportions of the recipe and the dosing of carbon black through the central control room. In the process, the personnel on duty in the central control room have to focus on the monitoring situation all the

time and make relevant records. However, some slack behaviors such as playing phones and sleeping of the personnel on duty could lead to untimely control, inaccurate records and even cause serious safety accidents [26]. Thereby, we present a behavior detection algorithm based on deep learning, which can fast and accurately detect the above-mentioned abnormal behaviors in central control room.

Abnormal behavior recognition is one of the basic tasks in the field of computer vision, especially the video surveillance. The commonly used behavior detection methods include human posture recognition [14], detection by smart wearables or smart watches [6, 30], and target detection [36]. However, due to the resolution of camera, the variety of target posture and the complexity of shooting scene, the task of behavior detection is still a huge challenge in the field of computer vision [43, 33].

Lately, with the development of deep learning technology [23, 39], the above-mentioned problems have been addressed. Due to the excellent performance of deep learning technology in field of video surveillance and biometric [1, 13, 28], many studies are using the convolutional neural networks (CNNs) to carry out abnormal behavior detection [8, 9]. Compared with the traditional behavior detection algorithms, the CNNs-based algorithm uses convolutional layer and pooling layer to extract features, which can improve both the detection accuracy and detection speed [19].

In this paper, the target detection technology is applied into abnormal behavior detection, and human is the only target. So far, the target detection algorithms based on CNNs can be divided into two categories: the two-stage object detection and the one-stage object detection.

The two-stage algorithm (such as R-CNN [42], Faster R-CNN [25], etc.) implements the object detection with two processes: extracting the candidate regions and classifying the extracted regions. For example, the Faster R-CNN uses a backbone network to obtain a shared feature map, then uses the RPN (region proposal network) to generate the ROI (region of interest), and resize the ROI feature through ROI Pooling. Finally, a FC (fully connection) layer integrates the feature information for target classification and location. Generally speaking, the two-stage algorithm owns the advantages of higher detection accuracy and lower miss-detection rate. However, the detection speed of two-stage algorithm is lower, which makes it cannot satisfy the real-time detection scenarios, such as video surveillance, medical diagnosis and traffic security.

On the contrary, the one-stage algorithm such as SSD (single-hot multi-Box detector) [17] and YOLO (you only look once) [24, 29], has great performance in detection speed. The core of the one-stage algorithm is that, for the input images, the target size, position and category are regressed directly through the network. Thus, the final detection results can be obtained with a single detection, which makes the detection speed higher. Of course, the detection accuracy of one-stage algorithm is commonly lower than two-stage algorithm. Hence, a variety of one-stage models are proposed to balance the detection accuracy and speed. For example, compared with the YOLOv1, the SSD introduces the pyramid feature hierarchy to enhance the detection ability for small target, but the recall is still rate. While YOLOv2 and YOLOv3 uses the BN (batch normalization) layer and anchor boxes to improve the detection accuracy, especially the small target. Based on this, the YOLOv4 [5] applies Mosaic, CSP, SPP module and FPN+PAN module to balance the detection accuracy and detection speed. In recent years, the YOLOv5 [37] algorithm is proposed. Compared with YOLOv4, it replaces the SPP module with SPPF module, and integrates the CSP module into both backbone and neck, which further improves its comprehensive performance. However, there still exists several problems such as model complexity, false detection and missed detection in actual behavior detection.

Thus, the article presents an indoor abnormal behavior detection algorithm based on improved YOLOv5. The algorithm uses SRGAN (super resolution GAN) [15] to enhance the image resolution for the problems of distortion and blurring in images taken by surveillance devices of conventional cameras. The use of MnasNet [32] to replace the original feature extraction network reduces the number of network parameters and improves the training and detecting speed of the model. By adding the ECA-Net (efficient channel attention neural network) [35] attention mechanism and changing the loss function with EIOU (efficient IOU) [41], the detection accuracy of the overall network is enhanced, especially for the detection of small objects such as mobile phone. The effectiveness of the proposed algorithm is experimented and verified on the custom dataset.

We summarize our main contribution as follows:

1 To enhance the resolution of the video surveillance images, the SRGAN is introduced as image preprocessing.
2 To reduce the computation costs of the YOLOv5 algorithm, we apply the MnasNet to replace the original backbone, aiming to improve the feature extraction speed and achieve lightweight of the proposed model.
3 To reinforce the model's detection ability for small objects such as smart phone, the attention mechanism ECA-Net is introduced into the feature fusion process, which can further simplify the model and accelerate the computing process at the meantime.
4 To further improve the model's robustness, we replace the original loss function of YOLOv5 with EIOU, which can also improve the detecting accuracy of the model.

The remaining section of this paper is as follows: Section 2 presents some related works. Section 3 introduces the proposed methodology, while Section 4 gives the experimental results and analysis. Section 5 concludes our work.

2. Related Work

In this section, we mainly discuss the related works about abnormal behavior detection. As mentioned above, the behavior detection using target detection technology primarily include two categories: two-stage detection and one-stage detection. The two-stage detection algorithm uses candidate regions and CNN classification networks to accomplish target detection. The one-stage detection algorithm directly generates the class probability and position coordinate values of the object.

2.1. Two-stage Object Detection

In 2014, the R-CNN was proposed, which had introduced the deep learning technology into object detection. After that, researchers proposed a variety of two-stage object detection algorithms. Among them, the Faster R-CNN was one of the most classical methods. This method applies RPN to accomplish the ROI generation, which significantly improves the detection speed.

However, the Faster R-CNN does not performance well in small target detection. Thus, scholars improve the structure of original Faster R-CNN to solve the false detection. For example, Mo et al. [20] proposed a research on human behavior detection based on Faster R-CNN. The Faster R-CNN algorithm was combined with OHEM (online-hard-example-mining) and the BN (batch normalization) algorithm to improve the accuracy of behavior detection. Although this method has obtained high detection accuracy on public dataset VOC 2012 Action, the problem of high model complexity still exists. Therefore, Jiang et al. [12] used a multi-time scale dual-stream network to extract features and improved the Faster R-CNN with the principle residual network. The testing results on custom data set have shown that the proposed method can effectively reduce the running speed of the whole network. In addition, Wang et al. [34] designed a variety of network improvement schemes such as difficult case extraction and multilayer feature fusion into Faster R-CNN algorithm. The comprehensive performance of the proposed method is verified on public data set VRU. Overall, to balance the detection accuracy and speed, researchers have improved the original Faster R-CNN with various strategies. However, due to the high computational complexity, two-stage detection models still cannot implement real-time detection for behavior detection in video surveillance. Hence, a variety of one-stage behavior detection algorithms are proposed recently.

2.2. One-stage Object Detection

Compared with two-stage object detection algorithm, one-stage object detection algorithm is simpler in process of feature extraction, which can reduce the computational complexity with expanse of detection accuracy. Therefore, scholars apply various one-stage object detection algorithm into abnormal behavior detection. For instance, Du et al. [4] replaced the feature extraction layer of SSD net with depth-separable convolution, which can reduce the number of model parameters. The improved SSD algorithm is trained and verified on public data set State Farm Distracted Driver Detection. Even though this method can effectively detect the driver's distraction behavior, the detection effect for target occlusion and small target does not perform well.

To increase the detection accuracy for small target, Chen et al. [3] presented an abnormal driving behav-

ior (including smoking and calling) detection algorithm based on YOLOv4, the research introduces the CMBA attention mechanism to enhance the detection of small target. The proposed model shows good performance on self-constructed dataset. Moreover, since the YOLOv5 algorithm was proposed, it had been widely used in various fields such as medical detection [21, 31], car driver detection [22], deep-sea detection [38], and pose recognition [16]. In abnormal behavior detection, Zhang et al. [40] detected unsafe behaviors such as un-wearing safety helmets and smoking in industrial places with an improved YOLOv5 framework. Chen et al. [2] introduced ASAE (adaptive self-attention embedding) module and WFPN (weighted feature pyramid network) module into the original YOLOv5 framework. The improved YOLOv5 model can detect unsafe behavior in industrial field with high accuracy and speed.

The above-mentioned instances have shown that the YOLOv5 algorithm owns better robustness and broad applicability. Compared with two-stage detection algorithm, YOLOv5 uses CNN network for end-to-end processing and directly generates multiple bounding boxes in the image. Thus, the same network outputs the position and classification of the targets, which reduces the consumption of computing resources. Meanwhile, the YOLOv5 model can effectively balance the detection accuracy and speed. Therefore, this paper adopts YOLOv5 as the base framework to detect abnormal behavior in central control room.

3. Methodology

In this section, we mainly describe the detail of our proposed YOLOv5-MEE algorithm. Firstly, the overall network structure of SRGAN is introduced, then, we respectively discuss the framework of MnasNet, ECA-Net and EIOU. Finally, we provide the whole structure of our improvement YOLOv5 algorithm.

3.1. Image Preprocessing with SRGAN

The images with distortion and blurring will cause difficulties in process of feature extraction and reduce the accuracy of behavior detection. Thus, we introduce the SRGAN to improve the resolution of the image taken by surveillance cameras in the central room.

SRGAN (super-resolution generative adver-sarial network) is one of the mainstream algorithms for image super-resolution reconstruction based on deep learning, which aims to convert low-resolution images into high-resolution images.

The whole network mainly contains two network models: the generator network and the discriminator network, the former network reconstructs a low-resolution image into a high-resolution image, and the later one is used to judge the difference between the generated image and the original image. The network structure of SRGAN is shown in Figure 1.

The generator network of SRGAN is an improvement on SR-ResNet (super-resolution residual network) [11], which mainly includes four components: low-level feature extraction, high-level feature extraction, de-convolutional layer and CNN reconstruction layer. The input to the generator network is a low-resolution image, which first passes through a 9×9 convolutional layer and a relu-activated layer. Then, the features are further extracted via N residual modules, and the flow of information between layers is improved by jumping connection. Finally, the size of features has been increased by two up-sampling, and then the output result is obtained through feature reconstruction, that is, a high-resolution image.

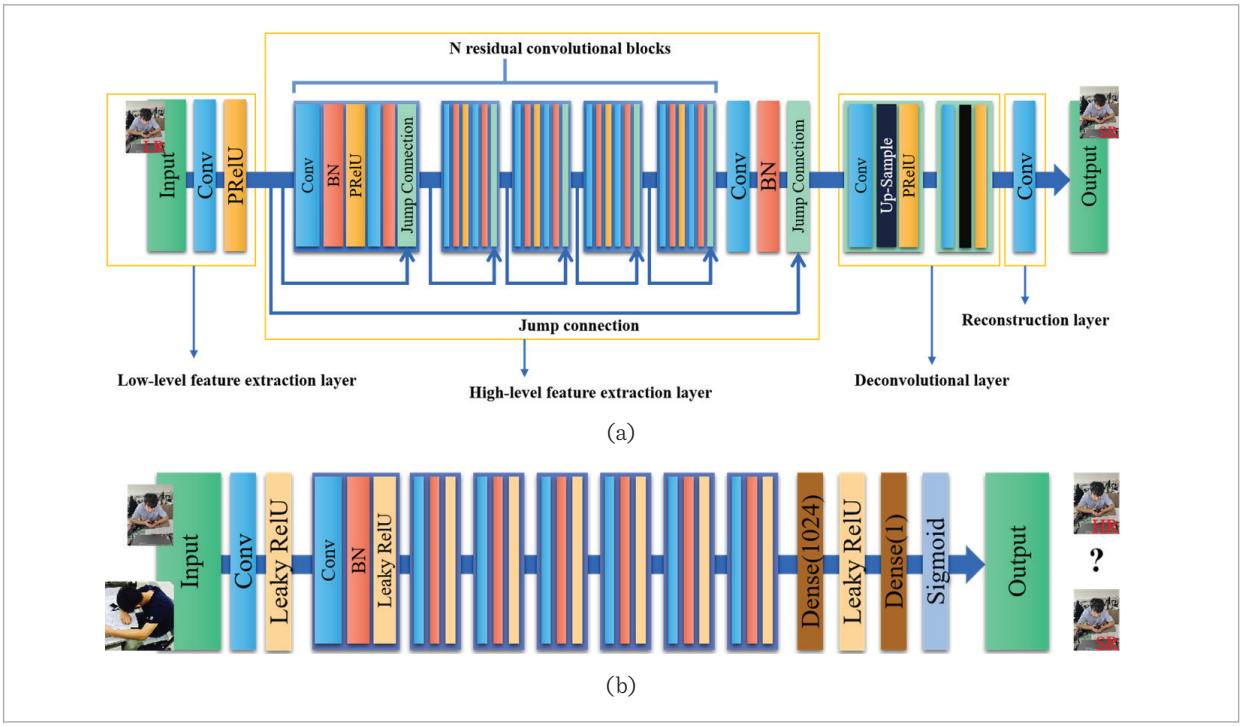
In Figure 1(b), the input of SRGAN's discriminator network is a high-resolution real image or a high-resolution fake image generated by the generator network. The Leaky Relu activation function is used to prevent negative output necrosis, and the calculation of redundant information is reduced by multiple step convolution. At the end of the network, the fully connected layer and the Sigmoid function are used for binary classification, that is, to determine whether the input image data is the data in the real training data, the judgment result is between 0~1, closing to 1 means that it is judged to be a real picture, and closing to 0 means that it is judged to be a false picture.

Different from the loss function of the previous SR model, the loss function of SRGAN consists of two parts, namely G_Loss and D_Loss. where G_Loss is the loss of the generator, which is defined as:

$$l^{SR} = l_X^{SR} + 10^{-3} l_{Gen}^{SR}, \tag{1}$$

where l^{SR} donates the loss of content, which includes the MSE Loss and VGG Loss; l_{Gen}^{SR} is the loss of antag-

Figure 1
Architecture of the SRGAN [15]: (a) generator network structure, (b) discriminator network structure



onism. Three types of these losses are defined as follows:

$$l_{MSE}^{SR} = \frac{1}{r^2 WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}(I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y})^2 \tag{2}$$

$$l_{VGG/i,j}^{SR} = \frac{1}{W_{i,j}H_{i.j}}\sum_{x=1}^{W_{i.j}}\sum_{y=1}^{H_{i,j}}[\varphi_{i,j}(I^{HR})_{x,y} - \varphi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y}]^2 \tag{3}$$

$$l_{Gen}^{SR} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{LR})). \tag{4}$$

D_Loss represents the loss of the SRGAN discriminator network, which is the binary cross-entropy loss function, which is defined as:

$$l_D = -y * \log D(x^{HR}) - (1-y) * \log(1 - D(x^{HR})). \tag{5}$$

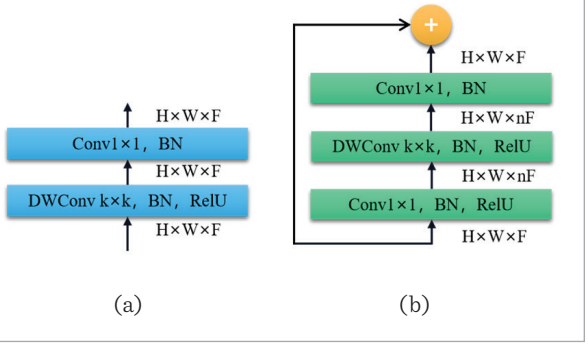
3.2. Feature Extraction Network of MnasNet

MnasNet is a lightweight neural network suitable for mobile terminals proposed by Google in 2019, which is mainly composed of standard convolution, deep separable convolution and inverse residual structure, which can improve the detection speed and reduce the amount of redundant parameters while ensuring that the detection accuracy is not reduced.

From the overall architecture, MnasNet inherits the deep separable convolution module of MobileNetV1 [10] and the inverse residual structure module of MobileNetV2 [27], and uses 5×5 convolution kernels in some convolutional layers to increase the receptive field of the network and improve the detection accuracy of the network. The two basic modules of MnasNet are shown in Figure 2.

The Deep Separable Convolution Module (SepConv) is consisted of two convolutional structures: deep convolution and point-by-point convolution, the former is used to process information in the direction of length and width, and the number of convolution kernels is determined according to the number of chan-

Figure 2
Two main modules of the MnasNet: (a) module of SepConv, (b) module of MBConv-n



nels of the input image, and each channel will generate a corresponding feature map. The latter is used to process information across channel directions, using a 1×1 standard convolution to integrate the output of deep convolution.

In the Inverse Residual Structure Module (MBConv-n), n represents the expansion coefficient of the number of channels of the original image. This module adopts the operation mode of "convolutional dimension –deep convolution - convolution dimensionality reduction", and performs deep convolution in high-dimensional space after dimensiona-lization, which can effectively ensure the richness of network feature extraction and reduce the amount of model parameters and calculation.

The original backbone of YOLOv5 is CSPDarknet-53, which introduces a CSP (cross stage partial) module into the Darknet-53. The CSPDarknet-53 can enhance the learning ability of CNN and reduce the memory cost. To further reduce the training parameters and improve inference speed, this paper replaces the backbone feature extraction network of YOLOv5s with MnasNet, and the specific structure of the replaced network is shown in Table 1. Specifically, the structure of MnasNet consists 132 layers, we extract the layer 52, layer 99 and layer 132 as the effective feature layers. Supposing the size of input image is 640×640×3, the output feature maps of three effective layers are 80×80×240, 40×40×672 and 20×20×1280, respectively. Then we use the three feature layers to construct FPN feature pyramid in the neck of YOLOv5.

Table 1
The backbone feature extraction network structure of the improved YOLOv5s

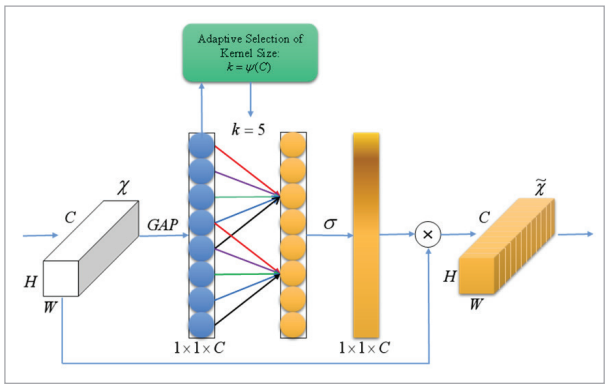
Input Scale	Type	Output Channel	Step
640×640×3	Conv(3×3)	32	2
320×320×32	SepConv(3×3)	16	1
320×320×16	MBConv-6(3×3)	24	2
160×160×24	MBConv-6(3×3)	24	1
160×160×24	MBConv-3(5×5)	40	2
80×80×40	MBConv-3(5×5)	40	1
80×80×40	MBConv-3(5×5)	40	1
80×80×40	MBConv-6(3×3)	80	2
40×40×80	MBConv-6(3×3)	80	1
40×40×80	MBConv-6(3×3)	80	1
40×40×80	MBConv-6(3×3)	80	1
40×40×80	MBConv-6(3×3)	112	2
40×40×112	MBConv-6(3×3)	112	1
40×40×112	MBConv-6(5×5)	160	2
20×20×160	MBConv-6(5×5)	160	1
20×20×160	MBConv-6(5×5)	160	1
20×20×160	MBConv-6(3×3)	320	2
20×20×320	Conv(1×1)	1280	1

3.3. ECA-Net Attention Mechanism

The channel attention mechanism ECA-Net (Efficient Channel Attention-Net) is an improvement on SE-Net [7] in the feature transformation part. Different from SE-Net's fully connected channel information inter-action mode, ECA-Net realizes information interaction between channels through one-dimensional convolution, and uses a method based on adaptive selection of convolution kernel size to realize local interaction of channel information, which reduces the complexity of the model while maintaining the performance of the model. The structure of ECA-Net is shown in Figure 3.

For an input image with size of $H \times W \times C$, the feature map χ is learned through a fast one-dimensional convolution with weight sharing after global average

Figure 3
Module of ECA-Net



pooling(GAP). The kernel size of the fast one-dimensional convolution represents the coverage of local cross-channel interactions, which can be determined by an adaptive function based on the size of input channel *C*. Then the function $\sigma = sigmoid()$ is used to determine the weight matrix for each channel, whose size is $1\times1\times C$. Finally, the original input features and channel weights are multiplied to obtain a feature map $\chi^{\%}$ with channel attention. The relationship between the convolution kernel size k of a fast one-dimensional convolution and the input channel *C* is as follows:

$$k = \psi(C) = \left| \frac{\log_2 C}{\gamma} + \frac{\gamma}{b} \right|_{odd}, \tag{6}$$

where γ and b are two parameters of the mapping function, the values are generally 2 and 1, respectively. odd donates taking an odd number. We can adaptively select the size of convolution kernel for different channel counts through Equation (6).

In this paper, the ECA-Net is introduced into the feature fusion process of YOLOv5. The output feature maps from MnasNet will be further processed by ECA-Net to enhance the useful feature (especially for small target feature) information. By this manner, the network can better extract the key information without increasing the training cost.

3.4. Loss Function of EIOU

The original YOLOv5 algorithm adopts the Complete Intersection over Union Loss (CIOU) as the loss function of the bounding box, which adds the overlapping area, center point distance and aspect ratio of the prediction box and the real box to the calculation at the same time, considering the difference between the predicted box and the real box from more dimensions, making the regression of the bounding box more stable. The CIOU loss function is defined as follows:

$$Loss_{CIOU} = 1 - IOU + \frac{\rho^2(b, b^{gt})}{C^2} + \alpha v \tag{7}$$

$$\alpha = \frac{v}{1 - IOU + v} \tag{8}$$

$$v = \frac{4}{\pi^2}(\arctan(\frac{w^{gt}}{g^{gt}}) - \arctan(\frac{w}{h}))^2. \tag{9}$$

The meaning of *IOU* represents the ratio of the intersection and union of the real box and the predicted box and $\rho^2(b, b^{gt})$ is the Euclidean distance of the prediction box from the center point of the real box. α donates a weight parameter, v is a parameter that measuring the similarity of the aspect ratio of the prediction box and the real one, which are defined as Equations (8)-(9), respectively. Moreover, w^{gt} and h^{gt} represent the width and height of the prediction box, while w and h donate the width and height of the real box.

As can be seen in Equation (9), the CIOU uses the relative proportions of width and height, rather than their true difference from their confidence, respectively. When the width and height of the prediction box and the true box satisfy the proportional relationship at any multiple, the usage of αv as a penalty is out of meaning, which prevents the model from effectively optimizing for similarity. Hence, we replace the bounding box loss function of the original YOLOv5 with EIOU loss.

The definition of EIOU loss function can be seen in Equation (10), where C_w and C_h represent the width and height of the smallest outer frame that covers the prediction box and the real box, respectively.

$$\begin{aligned} Loss_{EIOU} &= 1 - IOU + \rho^2(b, b^{gt}) / C^2 \\ &+ \rho^2(w, w^{gt}) / C_w^2 + \rho^2(h, h^{gt}) / C_h^2 \end{aligned}. \tag{10}$$

The EIOU is consisted of three parts: overlapping loss, width and height loss, and center distance loss. The first two parts are the same as the calculation method of CIOU. The width and height loss calculates the length and width of the prediction box and the real box, respectively, so that the difference between

the width and height of the two boxes is minimized, thereby accelerating the conver-gence speed and improving the detection accuracy.

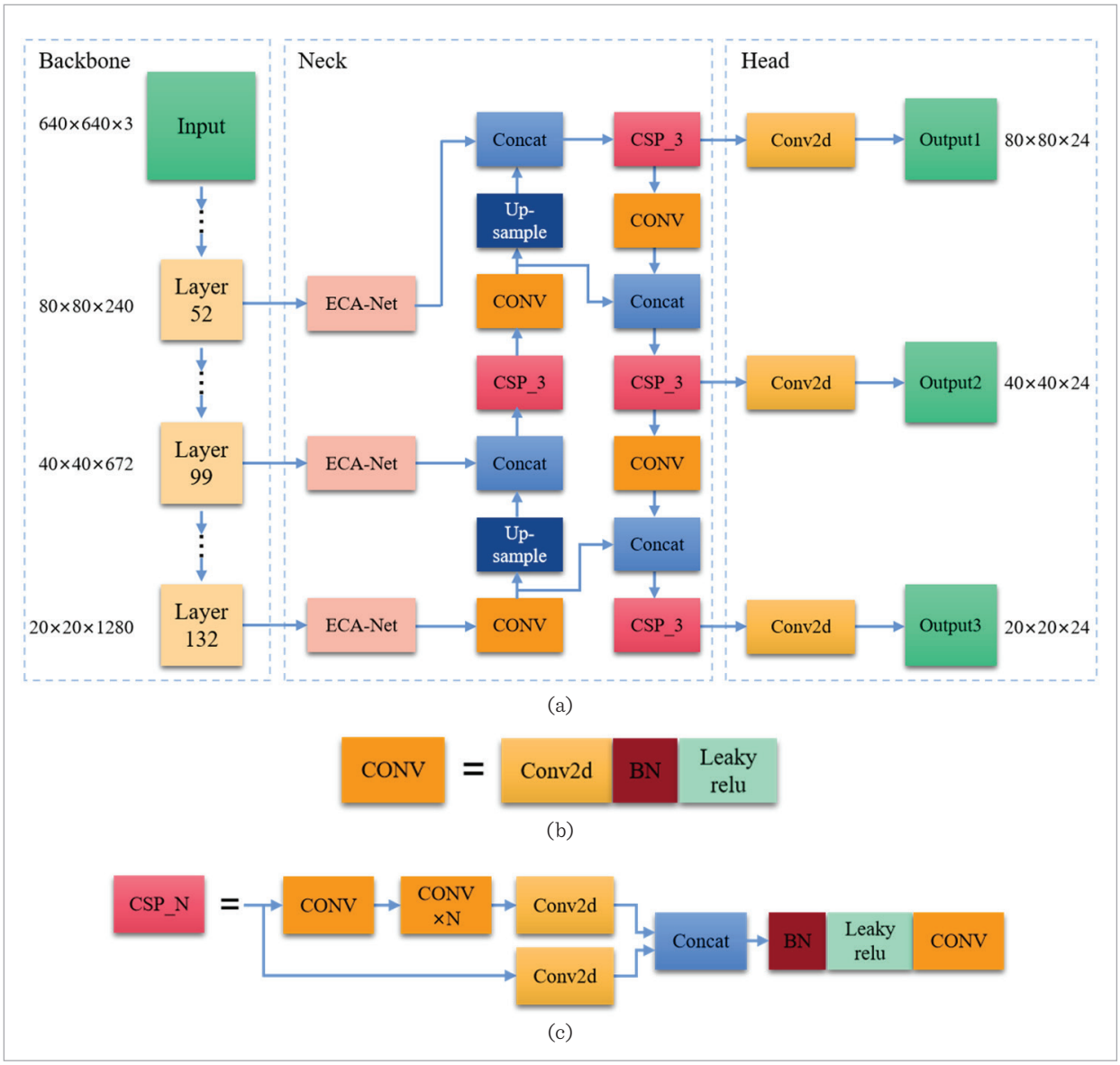
3.5. Overall Structure of Improved YOLOv5

The overall network architecture of the improved YOLOV5 algorithm (hereinafter referred to as YOLOv5s-MEE) proposed in this paper is shown in Figure 4. In Figure 4(a), the input is a high-resolution image pre-processed by SRGAN, with a size of 640×640×3; the backbone network MnasNet performs feature extraction, and the feature layers of 80×80×240, 40×40×672, and 20×20×1280 are sent to the Neck part for feature fusion.

In the Neck part, firstly, the effective feature layers with channel attention information are obtained by

Figure 4

The network structure of YOLOv5s-MEE: (a) the overall structure of the algorithm we proposed, (b) the module of CONV in (a), (c) the module of CSP-N in (a)



the introduced ECA-Net. Then, the features are further extracted through the FPN feature pyramid and PAN (path aggregation network structure). In process of feature fusion, the CONV module contains a standard 2D convolutional layer, a batch normalization layer and a leaky relu activation layer, which can be seen in Figure 4(b). The structure of CSP module is shown in Figure 4(c). In addition, the up-sample module is used to magnify the high lever features. As for down-sampling, we adopt a CONV module (kernel size is 3×3, stride is 2) to reduce the dimensionality of feature maps and avoid overfitting. With the FPN feature pyramid, three strengthened feature layers with rich semantic information are obtained. Finally, these strengthened feature layers are used as the input of the Head part to obtain the prediction results.

In YOLO Head, a convolutional layer is applied to adjust channels for each feature layer from Neck part. The number of channels can be calculated as follows:

$$channels = N \times (4 + 1 + classes). \quad (11)$$

The N=3 donates the number of priori boxes, 4 represents four coordinate values of prediction box, 1 represents one confidence parameter, and classes is the number of categories. In this paper, the class value is 3, thus the output channel is 24 (i.e. 3×(4+1+3)=24). Therefore, the YOLO head outputs prediction results for each effective feature layer, and their shape are 20×20×24, 40×40×24 and 80×80×24, respectively.

4. Experiment

In this section, we introduce the generated dataset used for model training and testing firstly, then we describe the configuration detail of our work. Finally, the results of ablation and comparative experiments are analyzed res-pectively.

4.1. Dataset Generation

According to the relevant safety regulations of central control room, more than two staff members should be guaranteed to operate in the central control room during working, and some slack behaviors such as playing mobile phones and sleeping are forbidden. Based on this, we define lacking people, playing phone and sleeping as abnormal behaviors. Moreover, several normal behaviors such as recoding, monitoring and drinking water are considered into our dataset.

Aiming to ensure the integrity of our dataset, the article simulates the environment of central control room in a laboratory. Two volunteers play as the staff members to emulate the five behaviors above. We use an industrial camera with two million pixels to take videos, the resolution of these video is 1920×1080, and the frame rate is 30 FPS. Then, we obtain images with 1920×1080 pixels by extracting key frame from these videos. Moreover, a variety of images with normal behaviors and abnormal behaviors are also obtained by internet crawling. There-fore, our dataset contains various images with different size, different resolution and different angle.

At last, we obtain 1200 images in total. Each image contains at least one human, each abnormal behavior (playing phone, sleeping) contains 450 images and each normal behavior (recoding, monitoring, drinking water) contains 100 images. These images are annotated in PASCAL VOC format by the LabelImg tool. Then, we divide the dataset by a ratio of 8:2, and 960 images are selected as the training set, 240 images as the test set. A partial sample of the dataset used in this article is shown in Figure 5.

4.2. Experimental Environment

The environment for model training in this paper is configured as an Ubuntu 16.04 64-bit operating system, with an NVIDIA GeForce RTX 2080Ti GPU and 11GB of video memory. Meanwhile, the environment for model testing is configured as a Windows 10 64-bit operating system, with an NVIDIA GeForce RTX 1050 GPU and 4GB of video memory. All experiments have used the Pytorch deep learn-ing framework and been programmed using the Python 3.8.

4.3. Model Training

As for SRGAN, we use the trained model in literature [15]; For the YOLOv5 model proposed in this paper, the training strategy is adopted as follows: the number of iterations is 100, the initial learning rate is 0.01, the learning rate momentum is 0.937, the weight decay coefficient is 0.0005, the batch-size is 8, the cosine annealing learning rate decay strategy is adopted, the optimizer is stochastic gradient descent (GSD), and the pre-training weight selection is yolov5s.pt. The loss curve during training is shown in Figure 6.

Figure 5
Sample image examples of the generated dataset. The first three images are frontal images and the last three images are side images counting from left to right

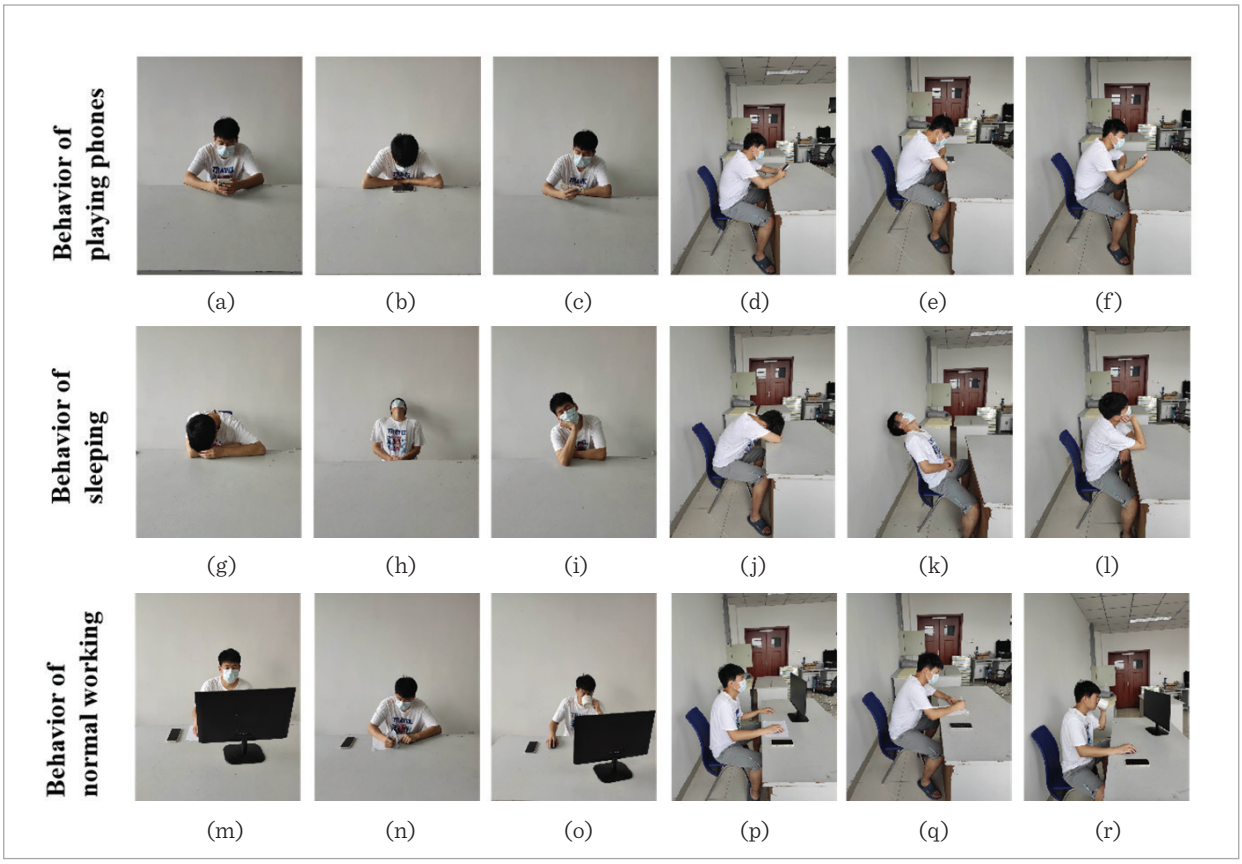
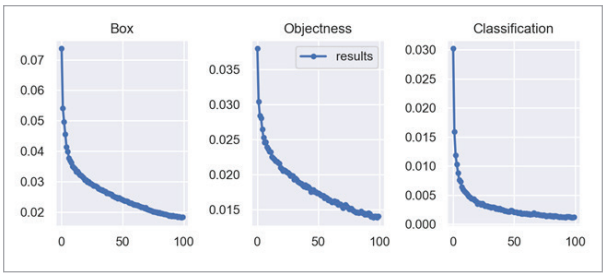


Figure 6
The loss variation of YOLOv5s-MEE



The three curves in Figure 6 correspond to the loss of box, the loss of confidence and the loss of classification. As it can be seen, the model's training loss gradually converges after 100 rounds of training, which means we have obtained a model with good generalization ability.

4.4. Evaluation Indicators

We mainly use the model complexity and model comprehensive as the evaluation indicators. Among them, the model complexity includes the model volume, parameter quantity and operation amount. The comprehensive performance of the model uses the average precision (AP), mean average precision (mAP) and frame per second (FPS), which are commonly used in the detection field to measure the performance and reliability of the algorithm. AP is related to the precision (P) and recall (R) of the model, which refers to the average accuracy at different recall rates. mAP is the average of AP across all categories and is used to measure detection accuracy. P, R, AP, and mAP are defined respectively as follows:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{12}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{13}$$

$$AP = \int_0^1 P(R)dR \tag{14}$$

$$mAP = \frac{\sum_{n \in N} AP(n)}{N}. \tag{15}$$

The meaning of *TP* is the sample size of the abnormal behavior predicted to be abnormal behavior.*FP* represents the sample size of what is predicted to be normal behavior but is actually abnormal behavior. *FN* donates the sample size of what is predicted to be normal behavior but is actually abnormal behavior; *N* is the number of target categories.

4.5. Ablation Experiments and Results Analysis

In order to verify the advancement and effectiveness of YOLOv5s-MEE proposed in this paper, ablation experiments are carried out on the improvement points one by one.

The first is the image preprocessing comparison experiment, in which the images in our dataset are SRGAN processed to obtain high-resolution images as input to the YOLOv5s algorithm. The results of some of the preprocessed images are shown in Figure 7, and the experimental results are shown in Table 2.

In Figure 7, the original image has the problems of blurring and distortion due to the low resolution, and lots of detail information have been lost. After super-resolution reconstruction by SRGAN, blur can be effectively removed and a clearer image is obtained. It is observed that the SRGAN can maximize the preservation of image details, and increase the image solution.

Figure 7
Example of SRGAN preprocessing results

In Table 2, S-YOLOv5s is the YOLOv5 model under the application of SRGAN image pre-processing module, and the input is RGB high-resolution images after processing with the size of 640×640. Two models use the same para-meters during training. Compared with the original YOLOv5 algorithm, by adding the SRGAN module, except the FPS, the mAP is increased by 0.96%, and the detection AP values of the three abnormal behaviors are increased by 0.52%, 1.83% and 0.34%, respectively, which shows that the detection accuracy of the YOLOv5s algorithm can be effectively improved after using SRGAN, especially for the detection of small targets.

The contour in-formation of the target in the reconstructed image is more obvious, which strengthens the feature extraction ability of network for small targets. For other improve-ments, several similar ablation experiments are designed, and the dataset used in these experiments has been processed by SRGAN. The experimental results are shown in Table 3. As can be seen in Table 3, four methods with different improvements are compared.

Table 2
Analysis of the effect of SRGAN image preprocessing. S-YOLOv5s: SRGAN for image preprocessing and original YOLOv5s for behavior detection

Algorithms	AP/%			mAP/%	FPS/frame×s-1
	Lacking people	Playing phone	Sleeping		
YOLOv5s	95.32	84.25	93.81	91.12	46.08
S-YOLOv5s	95.84	86.08	94.15	92.02	47.62

Table 3
Analysis of algorithm improvements ablation experiments. Method1: SRGAN + YOLOv5s; Method2: SRGAN + YOLOv5s + MnasNet; Method3: SRGAN + YOLOv5s + MnasNet + ECA-Net; Method4: SRGAN + YOLOv5s + MnasNet + ECA-Net + EIOU

Algorithms	Improvements			AP/%			mAP/%	FPS/frame×s-1
	MnasNet	ECA-Net	EIOU	Lacking people	Playing phone	Sleeping		
Method1	×	×	×	**95.81**	86.08	94.15	92.09	47.62
Method2	√	×	×	94.73	85.45	92.67	90.95	**79.50**
Method3	√	√	×	95.41	88.62	94.37	92.80	77.07
Method4	√	√	√	95.53	**88.98**	**95.03**	**93.18**	75.50

Among them, method 1 means S-YOLOv5s in Table 2. Method 2 means replacing the backbone network of method 1 with MnasNet. Method 3 means incorporating the ECA-Net attention mechanism on the basis of Method 2. Method 4 indicates the use of EIOU as the loss function of method 3, which is the algorithm proposed in this article.

Compared with Method 2 and Method 1, although mAP decreased by 1.14%, the FPS increased by 33.42 frames/s, which indicates that the introduction of lightweight network MnasNet can sacrifice a small amount of detection accuracy in exchange for a substantial increase in detection speed, which can meet the requirements of real-time detection.

For the comparison with method 2, method 3 improves the detection accuracy by adding the ECA-Net attention mechanism, the detection accuracy of "playing mobile phone" behavior is increased by 3.17%, indicating that the attention mechanism can effectively improve the feature extraction ability of the model

The results of Method 4 show that the introduction of EIOU as the bounding box loss function can further improve the model detection performance: the AP of the three behaviors is increased by 0.12%, 0.36%, and 0.66%, and the mAP is increased by 0.38%, which proves the effectiveness of EIOU.

In comprehensiveness of Tables 2-3, compared with the original YOLOv5s algorithm, our proposed algorithm in this paper has improved the accuracy of abnormal behavior detection to 93.18%, and the detection speed has been increased to 75.50 frames/s, which meets the requirements of accuracy and real-time performance.

Table 4 shows the comparison results of the model complexity between the proposed algorithm and the original YOLOv5s algorithm. It can be seen from the comparison that the proposed algorithm reduces the model volume by 53.43%, the number of parameters by 12.87%, and the amount of operation by 4.74%. It shows that the algorithm in this paper has a lighter weight model and is more suitable for devices with low hardware performance.

Table 4
Comparison of the complexity of the improvement algorithm with YOLOv5s

Algorithms	Model Size/MB	Number of Parameters/M	Computation/G
YOLOv5s	28.8	7.296	17.06
YOLOv5s-MEE	13.7	6.357	16.25

4.6. Comparative Experiments and Results Analysis

In this paper, several classical detection algorithms including SSD, YOLOv3, YOLOv4 and Faster R-CNN, are introduced for comparative experiments to further verify the effectiveness of the proposed algorithms. For the reliability of the results, all comparison experiments are performed in the same environment (software and hardware configurations, dataset). Moreover, we have set the same training and testing protocols for each methods: the momentum is 0.937, the learning rate is 0.01, the weight decay is 0.0005, the epoch is 100, and the batch size is 8. The experimental results are shown in Tables 5-6.

Table 5
The performance of our algorithm with other mainstream algorithms

Algorithms	AP/%			mAP/%	FPS/frame×s-1
	Lack of people	Playing phone	Sleeping		
SSD	91.22	86.92	92.52	90.45	**75.97**
Faster R-CNN	93.60	78.62	93.71	88.65	11.17
YOLOv3	55.90	78.92	85.34	73.39	21.39
YOLOv4	41.36	69.73	80.83	63.98	35.91
YOLOv5s-MEE	**95.53**	**88.98**	**95.03**	**93.18**	75.50

Table 6
The complexity comparison of our algorithm with other mainstream algorithms

Algorithms	Model Size/MB	Number of Parameters/M	Computation/G
SSD	91.625	26.285	62.747
Faster R-CNN	108.24	137.099	370.21
YOLOv3	235	58.7	65.312
YOLOv4	244	63.363	60.527
YOLOv5s-MEE	13.7	6.357	16.25

It is observed from Table 5 that the detection accuracy of the proposed algorithm is higher than that of these three mainstream algorithms, and the detection speed is slightly lower than that of SSD, indicating that the proposed algorithm can better balance detection accuracy and detection speed. As can be seen from Table 6, the model complexity of the proposed algorithm is much lower than other algorithms, and it has better device applicability. Combined with the two sets of comparative experiments, it has been proved that the comprehensive performance of the algorithm in this paper is the best among other mainstream algorithms.

Aiming to more intuitively show the difference in detection effect between the proposed algorithm and other algorithms, we extract three sets of images from the test set as verification, corresponding to mobile phone behavior, sleeping behavior and mixed behavior (including normal behaviors and abnormal behaviors), and some of the detection results are shown in Figure 8.

It can be seen from the figure that the other three algorithms exist the phenomenon of missed detection and false detection. However, the algorithm in this paper can effectively detect the abnormal behavior of playing mobile phones and sleeping in various postures, and the detection confidence is high. For the detection of multiple targets, far targets and small targets, the algorithm in this paper also has no missed detection and false detection.

Moreover, the predefine behaviors shown in Figure 5 are also detected by the proposed model, and the detection results can be seen in Figure 9.

It is observed that most of the behavior detection results are correct, for action of playing phone, the YOLOv5s-MEE can precisely recognize the behavior of holding phones. However, if the mobile phone places on the table, our algorithm cannot well predict whether the staff is playing phone or not. If there exists a behavior of playing phone with head down, the detection results will be the sleeping behavior. Of course, if the phone has been put aside while normal working, the YOLOv5s-MEE can also detect accurately. For behaviors of sleeping, our algorithm can recognize them with high accuracy, both from the front and the side. As for normal working, recording with head down and drinking water, the proposed algorithm can detect most of the correctly.

Figure 8

Detection results of different algorithms. Phone-sleep: one person is playing phone and the other one is sleeping. Phone-phone: both of them are playing phone. Phone-normal: one person is playing phone and the other one is working normally

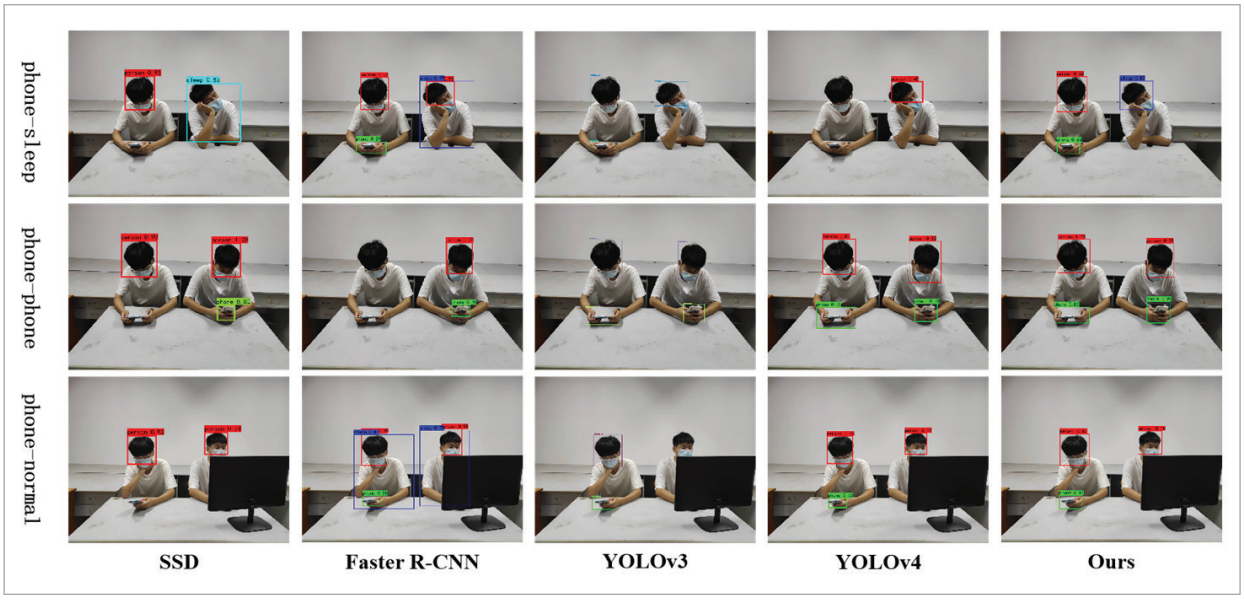
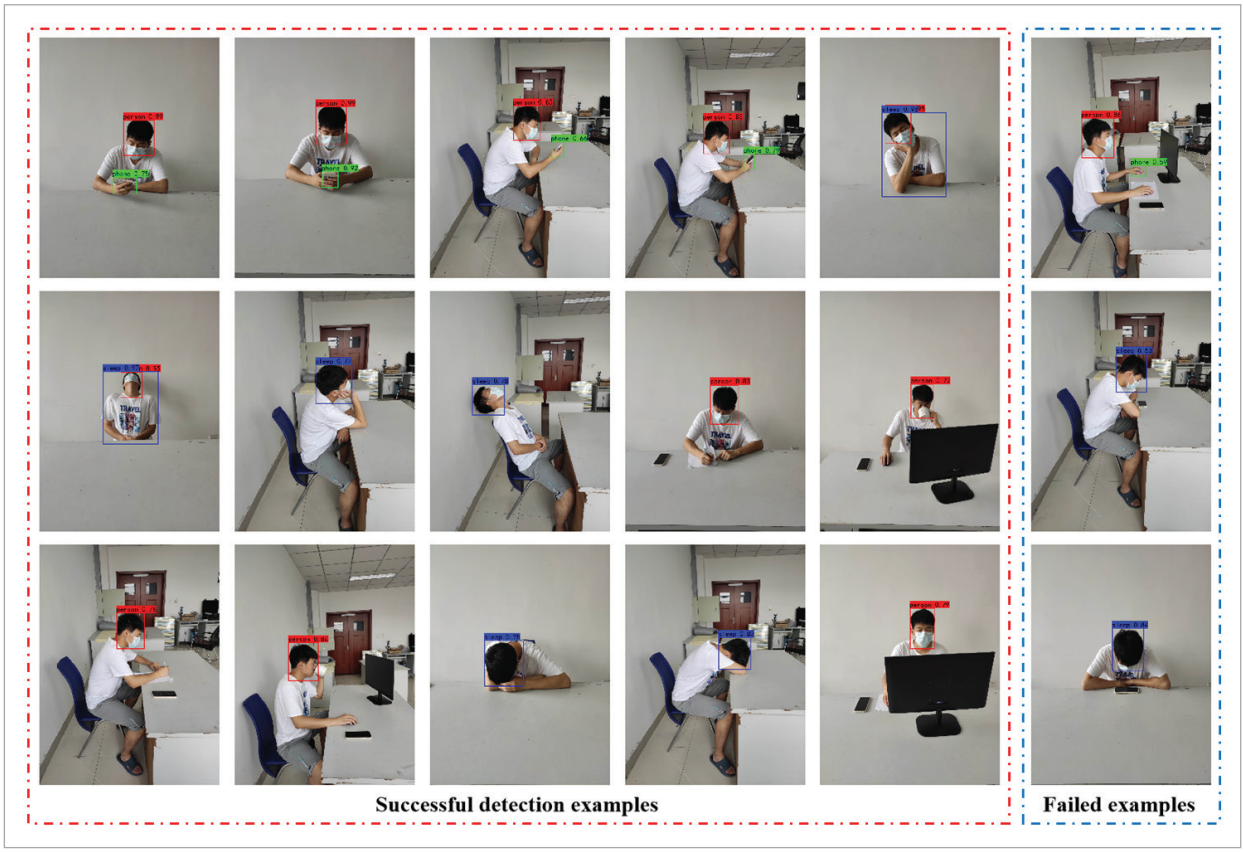


Figure 9

The detection results with our proposed YOLOv5s-MEE. The original images can be seen in Figure 5. The images in orange box are several successful detection examples, and the images in blue box are some failed results



However, our YOLOv5s-MEE has several false detection results at the same time, which can be seen in the blue box in Figure 9. On the one hand, although the SRGAN can increase the images' resolution, some details are not process well, which will cause local distortion for those far targets. On the other hand, in the process of dataset generation, the amount of images about playing phone with putting phone on the table are not enough, thus the type of this behavior cannot detect well.

5. Conclusion

In this paper, we propose a lightweight ab-normal behavior detection algorithm based on the improved YOLOv5, which is used to detect abnormal behaviors such as playing mobile phones, sleeping and missing people in the central control room. We use the SRGAN to preprocess the images with resolution im-provement, and solve the problem of blurry and distorted image caused by low hardware device configuration. Secondly, the MnasNet is applied to reduce the amount of model parameters. The ECA-Net is integrated into the neck part to improve the accuracy of model detection. Finally, we introduce the EIOU loss function to enhance the model's detection performance. The experimental results on self-constructed dataset show that our algorithm has reached 93.18% in detection accuracy and 75.5 frames/s in detection speed. Compared with SSD, YOLOv3 and YOLOv4, our proposed algorithm has great advantage in terms of detection comprehensive performance and model com-plexity. Meanwhile, our proposed technique is more dependable in categorization of activities with fewer instances and homogenous classes with overlapping attributes.

However, the algorithm still has certain defects: due to the limitations of the dataset, the placement of the shooting equipment needs to be considered in the detection process, which is inconsistent with the actual situation. In future, we will consider further improving the gen-eralization ability of the model by enriching the dataset.

Acknowledgement

This research was funded by the National Natural Science Foundation of Shandong (NO. ZR2019MEE071) and the Taishan Scholar Project Fund of Shandong Province.

References

1. Ahan, T., Khalid, S., Najam, S., Khan, M. A., Kit, Y. J., Chang, B. Hrneto: Human Action Recognition Using Unified Deep Features Optimization Framework. Computers, Materials & Continua, 2023, 75(1), 1089-1105. https://doi.org/10.32604/cmc.2023.034563
2. Chen, B.B, Wang, X. H, Bao, Q. F, Jia, B., Li, X. S, Wang, Y. R. An Unsafe Behavior Detection Method Based on Improved YOLO Framework. Electronics, 2022, 11(12), 1912. https://doi.org/10.3390/electronics11121912
3. Chen, N. T., Man, Y. Z., Sun, Y. C. Abnormal Cockpit Pilot Driving Behavior Detection Using YOLOv4 Fused Attention Mechanism. Electronics, 2022, 11, 2538-2549. https://doi.org/10.3390/electronics11162538
4. Du, X. L., Yu, H. P. Detecting Driver's Distracted Behavior Based on Improved Mobile Net-SSD Network. Journal of Highway and Trans Research and Development, 2022, 39, 160-166.
5. Feng, D., Liang, M., Wang, G. Improved YOLOv4 Based on Dilated Convolution and Focal Loss. In Proceedings of the 2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA), 2021, 966-971.https://doi.org/10.1109/AEECA52519.2021.9574147
6. Gupta, R., Gupta, S. H., Agarwal, A., Choudhary, P., Bansal, N., Sen, S. A Wearable Multisensor Posture Detection System. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). 2020, 818-822. https://doi.org/10.1109/ICICCS48265.2020.9121082
7. Hu, J., Shen, L., Sun, G. Squeeze-And-Excitation Networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 42(8), 7132-7141. https://doi.org/10.1109/TPAMI.2019.2913372
8. Ibrar, K., Fayyaz, A. M., Khan, A. M., Alhaisoni, M., Tariq, U., Jeon, S., Nam, Y. Human Personality Assessment Based on Gait Pattern Recognition Using Smartphone Sensors. Computer Systems Science and Engineering, 2023, 46(2), 2351-2368. https://doi.org/10.32604/csse.2023.036185
9. Jahangir, F., Khan, M. A., Alhaisoni, M., Alqahtani, A., Alsubai, S., Sha, M., Al Hejaili, A., Cha, J.-H. A Fusion-As-

sisted Multi-Stream Deep Learning and ESO-Controlled Newton-Raphson-Based Feature Selection Approach for Human Gait Recognition. Sensors, 2023, 23(5), 2754. https://doi.org/10.3390/s23052754

10. Jia, X. L., Peng, Y. L., Ge, B., Xin, Y. H., Liu, S. G. Dual-Complementary Convolution Network for Remote-sensing Image Denoising. IEEE Geoscience and Remote Sensing Letters, 2021, 19, 1-5. https://doi.org/10.1109/LGRS.2021.3101851
11. Jiang, X. H., Xu, Y. F., Wei, P. P., Zhou, Z. M. CT Image Super Resolution Based on Improved SRGAN. 2020 5th International Conference on Computer and Communication Systems (ICCS), 2020, 363-367. https://doi.org/10.1109/ICCCS49078.2020.9118497
12. Jiang, Z. G., Shi, X. T. Application Research of Key Frames Extraction Technology Combined with Optimized Faster R-CNN Algorithm in Traffic Video Analysis. Complexity, 2021. https://doi.org/10.1155/2021/6620425
13. Khan, M. A., Arshad, H., Khan, W. Z., Alhaisoni, M., Tariq, U., Hussein, H. S., Alshazly, H., Osman, L., Elashry, A. HGRBOL2: Human Gait Recognition for Biometric Application Using Bayesian Optimization and Extreme Learning Machine. Future Generation Computer Systems, 2023, 143, 337-348. https://doi.org/10.1016/j.future.2023.02.005
14. Kulikajevas, A., Maskeliunas, R., Damaševičius, R. Detection of Sitting Posture Using Hierarchical Image Composition and Deep Learning. Peer Journal of Computer Science, 2021, 7, e442. https://doi.org/10.7717/peerj-cs.442
15. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. H., Shi, W.Z. Photo-realistic Single Image Super-Resolution Using a Generative Adversarial Network. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, 105-114. https://doi.org/10.1109/CVPR.2017.19
16. Ling, L., Tao, J., Wu, G. Research on Gesture Recognition Based on YOLOv5. 2021 33rd Chinese Control and Decision Conference (CCDC), 2021, 801-806. https://doi.org/10.1109/CCDC52312.2021.9602731
17. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., Berg, A. C. SSD: Single Shot Multibox Detector. Computer Vision- ECCV 2016, 2016, 9905, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2
18. Mentler, T., Rasim, T., Müßiggang, M., Michael, H. Ensuring Usability of Future Smart Energy Control Room Systems. Energy Informatics, 2018, 1(26). https://doi.org/10.1186/s42162-018-0029-z
19. Mneymneh, B. E., Abbas, M., Khoury, H. Vision-based Framework for Intelligent Monitoring of Hardhat Wearing on Construction Sites. Journal of Computing in Civil Engineering, 2019, 33(2). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000813
20. Mo, H. W., Wang, H. B. Research on Human Behavior Detection Based on Faster R-CNN. CAAI Transactions on Intelligence Systems, 2018, 13, 976-973.
21. Mushtaq, M., Akram, M. U., Alghamdi, N. S., Fatima, J., Masood, R. F. Localization and Edge-Based Segmentation of Lumbar Spline Vertebrae to Identify the Deformities Using Deep Learning Models. Sensors, 2022, 22, 1547-1569. https://doi.org/10.3390/s22041547
22. Nemcova, A., Svozilova, V., Bucsuhazy, K., Smisek, R., Mezl, M., Hesko, B., Belak, M., Bilik, M., Maxera, R., Seitl, M. Multimodal Features for Detection of Driver Stress and Fatigue. IEEE Transactions on Intelligence Transactions Systems, 2021, 22, 3214-3233. https://doi.org/10.1109/TITS.2020.2977762
23. Qi, W., Ovur, S. E., Li, Z. J., Marzullo, A., Song, R. Multi-sensor Guided Hand Gesture Recognition for a Teleoperated Robot Using a Recurrent Neural Network. IEEE Robotics and Automation Letters, 2021, 6(3), 6039-6045. https://doi.org/10.1109/LRA.2021.3089999
24. Redmon, J., Divvala, S., Girshick, R, Farhadi, A. You Only Look Once: Unified, Real-time Object Detection. Computer Vision and Pattern Recognition, 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91
25. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031
26. Roe, E., Schulman, P. R. A Reliability & Risk Framework for the Assessment and Management of System Risks in Critical Infrastructures with Central Control Rooms. 2018, 110, 80-88. https://doi.org/10.1016/j.ssci.2017.09.003
27. Sandler, M., Howard, A., Zhu, M. L., Zhmoginov, A., Chen, L. C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 4510-4520. https://doi.org/10.1109/CVPR.2018.00474
28. Shehzad, F., Khan, M. A., Yar, M. A. E., Sharif, M., Alhaisoni, M., Tariq, U., Majumdar, A., Thinnukool, O. Two-stream Deep Learning Architecture-Based Human Action Recognition. Computers, Materials & Continua, 2023, 74(3), 5931-5949. https://doi.org/10.32604/cmc.2023.028743

29. Shu, Z. K., Yan, Z. Y., Xu, X. H. Pavement Crack Detection Method of Street View Images Based Deep Learning. Journal of Physics: Conference Series, 2021, 1952(2), 22-29. https://doi.org/10.1088/1742-6596/1952/2/022043

30. Skach, S., Stewart, R., Healey, P. G. T. Smart Arse: Posture Classification with Textile Sensors in Trousers. Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI'18). Association for Computing Machinery. 2018, 116-124. https://doi.org/10.1145/3242969.3242977

31. Su, H., Qi, W., Chen, J. H., Zhang, D. D. Fuzzy Approximation-based Task-space Control of Robot Manipulators with Remote Center of Motion Constraint. IEEE Transactions on Fuzzy Systems, 2020, 30, 1564-1573. https://doi.org/10.1109/TFUZZ.2022.3157075

32. Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Haward, A., Le, Q. V. Mnasnet: Platform-aware Neural Architecture Search for Mobile. 2019 IEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 2815-2823. https://doi.org/10.1109/CVPR.2019.00293

33. Uzair, M., Brinkworth, R. S., Finn, A. Bio-Inspired Video Enhancement for Small Moving Target Detection. IEEE Transaction Image Processing, 2020, 30, 1232-1244.https://doi.org/10.1109/TIP.2020.3043113

34. Wang, K. L., Zhou, W. Pedestrian and Cyclist Detection Based on Deep Neural Network Fast R-CNN. International Journal of Advanced Robotic Systems, 2019, 16(2). https://doi.org/10.1177/1729881419829651

35. Wang, Q. L., Wu, B. G., Zhu, P. F., Li, P. H., Zuo, W. M., Hu, Q. H. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 11531-11539. https://doi.org/10.1109/CVPR42600.2020.01155

36. Wang, Y., Zhou, L. L., Peng, X. Y. Detection of Abnormal Behavior in the Elderly Indoor Environment. Journal of Civil Engineering and Management, 2022, 39, 145-152.

37. Wang, Z., Wu, L., Li, T, Shi, P. B. A Smoke Detection Model Based on Improved YOLOv5. Mathematics, 2022, 10, 1190-1203. https://doi.org/10.3390/math10071190

38. Wu, R., Bi, X. J. A Coral Benthic Recognition Method Based on the Improved YOLOv5 Algorithm. Journal of Harbin Engineering University, 2022, 43, 580-586.

39. Yao, J., Fan, X., Li, B., Qin, W. Adverse Weather Target Detection Algorithm Based on Adaptive Color Levels and Improved YOLOv5. Sensors, 2022, 22, 8577-8598. https://doi.org/10.3390/s22218577

40. Zhang, J., Qu, P. Q., Sun, C., Luo, M. Safety Helmet Wearing Detection Algorithm Based on Improved YOLOv5. Journal of Computer Application, 2022, 42, 1292-1300.

41. Zhang, Y. F., Ren, W. Q., Zhang, Z., Wang, L., Tan, T. N. Focal and Efficient IOU Loss for Accurate Bounding Box Regression. Neurocomputing, 2022, 506, 146-157. https://doi.org/10.1016/j.neucom.2022.07.042

42. Zhou, C. H., Zhoy, J. Y., Yu, C., Zhao, W., Pan, R. L. Multi-channel Sliced Deep RCNN with Residual Network for Text Classification. Chinese Journal of Electronics, 2020, 29(5), 880-886. https://doi.org/10.1049/cje.2020.08.003

43. Zhou, K., Hui, B., Wang, J., Wang, C., Wu, T. A Study on Attention-based LSTM for Abnormal Behavior Recognition with Variable Pooling. Image and Vision Computing, 2021, 108, 104-120. https://doi.org/10.1016/j.imavis.2021.104120