Hindawi
Journal of Electrical and Computer Engineering
Volume 2022, Article ID 4260543, 12 pages
https://doi.org/10.1155/2022/4260543

Research Article

A New Multiface Target Detection Algorithm for Students in Class Based on Bayesian Optimized YOLOv3 Model

Dongmei Shi[1] **and Hongyu Tang**[2]

[1]*Department of Computer Science and Technology, Suzhou College of Information Technology, Suzhou, China*
[2]*School of Electrical and Information, Zhenjiang College, Zhenjiang, China*

Correspondence should be addressed to Hongyu Tang; t_redrain@126.com

Received 2 November 2021; Revised 27 November 2021; Accepted 30 November 2021; Published 4 January 2022

Academic Editor: Yang Li

Deep learning theory is widely used in face recognition. Combined with the needs of classroom attendance and students' learning status monitoring, this article analyzes the YOLO (You Only Look Once) face recognition algorithms based on regression method. Aiming at the problem of small target missing detection in the YOLOv3 network structure, an improved YOLOv3 algorithm based on Bayesian optimization is proposed. The algorithm uses deep separable convolution instead of conventional convolution to improve the Darknet-53 basic network, and it reduces the amount of calculation and parameters of the network. A multiscale feature pyramid is built, and an attention guidance module is designed to strengthen multiscale fusion, detecting different sizes of targets. The loss function is improved to solve the imbalance of positive and negative sample distribution and the imbalance between simple samples and difficult samples. The Bayesian function is adopted to optimize the classifier and improve the classification efficiency and accuracy, ensuring the accuracy of small target detection. Five groups of comparative experiments are carried out on public COCO and VOC2012 datasets and self-built datasets. The experimental results show that the proposed improved YOLOv3 model can effectively improve the detection accuracy of multiple faces and small targets. Compared with the traditional YOLOv3 model, the mean mAP of the target is improved by more than 1.2%.

1. Introduction

In recent years, biometric authentication has been widely used in all walks of life. There are mature technologies, such as fingerprint recognition, face recognition, and iris recognition [1], which have been applied in university classroom. Due to the loose discipline control in university classroom, with the popularity of mobile devices, students like to use mobile phones in class, which seriously affects the quality of learning. In order to improve the effect of classroom learning, we need to use face recognition technology. On the one hand, students' attendance in class can be realized. On the other hand, through the face images in the classroom, we can analyze the students' learning status and improve the students' participation in classroom learning. Therefore, it is very necessary to use face recognition technology to realize class attendance and head-up rate recognition, monitoring teaching process, analyze learning and teaching situation, so as to improve methods and improve teaching quality.

In face recognition and behavior recognition, scholars have carried out a lot of research studies. At present, face detection methods are mainly divided into two categories: knowledge-based methods and statistics-based methods [2]. These two methods extract the features of the region to judge by calculating the similarity of the features or the response value of the classifier. The knowledge-based method has the features of skin color, texture, structure, edge, shape, etc. Affected by different environments, the recognition accuracy of this method is quite different. However, statistics-based methods have been deeply studied, such as artificial neural network (ANN), AdaBoost method, support vector machine (SVM) [3], feature space method, long-term recurrent convolutional networks (LRCN), and convolutional neural

networks (CNN) [4, 5]. The network structures of these methods include AlexNet, VGG (visual graphics generator), and Inception Net model [6].

The methods based on feature space include principal component analysis (PCA), linear discrimination analysis (LDA), and local binary pattern (LBP) [7]. One of the common characteristics of these methods is to use the mapping of space vector in one feature space to distinguish face from nonface. Gabor, histogram of oriented gradients (HOG) [8], and scale-invariant feature transform (SIFT) are also used to combine global features with local effective features to form the final features of face recognition. Face recognition is performed on learning video by the Fisher weighting criterion [9], and Gabor features and cooperative representation are combined. The face recognition algorithm (Gabor CRC) and speed are proposed [10]. In video image behavior recognition, face feature extraction and appearance expression are the important basis of behavior recognition algorithm. The common methods of appearance expression include contour template, light flow, and feature point. In the process of candidate region video image processing and feature extraction, color histogram, Haar feature or Haar-like feature, HOG feature operator [11], and wavelet algorithm [12] are usually used. Then, machine learning classifiers, such as Softmax and SVM, and boosting or random forest classification algorithms [13–15], are used.

The second is single-stage target detection algorithm, represented by single shot multi-box detector (SSD) [16–19] and YOLO [20, 21], which is based on regression and classification of target detection, learning from the Faster R-CNN, sacrificing a little speed to further improve the accuracy. At present, YOLO has developed to YOLOv3. Compared with the RetinaNet [22] algorithm with the best accuracy before YOLOv3, under the same detection accuracy, the detection speed of YOLOv3 is 3.8 times that of the RetinaNet algorithm. Under the 320×320 resolution, compared with the SSD algorithm, YOLOv3 can detect map better than SSD, which can reach 28.2%. The processing time of each picture is 3 times faster than SSD, which only takes 22 ms. In [23], the SSD algorithm combines the advantages of YOLO and accurate positioning of the region proposal network (RPN), but the disadvantage is that the speed is slower than YOLO.

In view of the misdetection rate of occluded targets in pedestrian detection by the YOLOv3 algorithm, the YOLOv3 network structure is improved [24], which can enhance the ability of multi-scale feature fusion. Based on the fusion of GIoU and Focal loss, the YOLOv3 target detection algorithm is proposed [25]. An improved YOLOv3 network structure is designed [26], and through saliency mapping, the most significant part of the mesh is selected to detect the object. A new multi-sensor multi-level enhanced YOLO convolution network model is proposed for robust vehicle detection in traffic monitoring [27]. The micro-YOLO network model convolution layer is optimized by the deep separable convolution [28]. It decomposes a complete convolution operation into deep convolution and point convolution, which reduces the parameter of CNN and improves the speed of operation. A multi-scale parallel network structure from dense to sparse is proposed for the face detection of different sizes in document [29]. But the recognition accuracy of multi-face targets in classroom and other places needs to be improved, because there are many faces, different sizes, and many priori bounding boxes for target detection in these occasions. Although most face targets can be detected, there are some missed detection rates, which makes some errors in the recognition accuracy of the class students' head-up rate.

In view of the above problems, this article proposes an improved YOLOv3 network structure based on Bayesian optimization for face recognition and face state analysis in class. The main contributions of this article are as follows:

(1) On the basis of the Darknet-53 network structure, the deep separable convolution is used to improve the network structure, which can lighten the model and reduce the network computation.

(2) Because of the students' different seats in the classroom, the size of the face target is different, so the feature pyramid is used to extract different scale features. In view of the diversity of the size of the face in the classroom, the attention guidance module is designed for multi-scale fusion, and it was combined with the method of DeepID network, so can recognize different sizes of the face targets.

(3) In order to solve the problem of unbalanced distribution of positive and negative samples, simple samples, and difficult samples, the Bayesian model is used to optimize the classifier to improve the classification accuracy of face recognition.

(4) The research team carried out experiments and result analysis in COCO, VOC, and self-built classroom face datasets, which further verified the face recognition efficiency of the network designed in this article and the effectiveness of the method.

2. YOLO Principle

YOLO is an end-to-end convolutional neural network for target detection. The YOLO grid cell has three prediction bounding boxes, which take the largest bounding box among intersection over union (IoU) of the current target box as the current target of prediction. The first two dimensions of the predicted output map are extracted feature dimensions. The third dimension is $B * (5 + C)$, where B is the number of bounding boxes predicted by each grid unit and 5 is one confidence level plus 4 coordinates (x, y, w, h). The bounding box with confidence less than threshold value set to 0 C is the number of bounding boxes. The output characteristic diagram contains parameters to be optimized for loss function. Finally, the nonmaximum suppression (NMS) is used to remove the repeated bounding box to detect various targets, as shown in Figure 1.

In 2015, Joseph Redmon et al. proposed the YOLOv1 network structure [30], which drew on GoogleNet thought, including 24 convolution layers and 2 full connection layers. In the YOLOv1 network, the last output layer uses linear activation function, and Leaky ReLU activation function is used after each convolution layer and full connection layer. Because the network structure contains the full connection layer, the problem of small target leakage in YOLOv1 is caused.

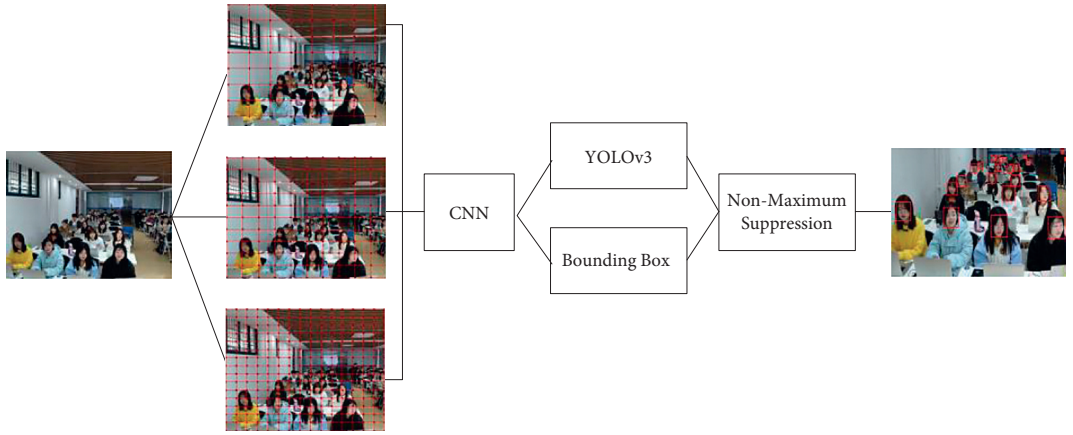


Figure 1: Process of YOLO face recognition.

In order to focus on solving the errors in the recall rate and positioning accuracy of YOLOv1, an improved YOLOv2 detection algorithm was proposed in 2016, which uses darknet-19 as the feature extraction network [31]. A variety of strategies, such as using anchor box, multi-scale training, and batch standardization processing, were proposed to improve the mean average precision (mAP) detected by the algorithm. Finally, the global average pooling layer is used to replace the full connection layer for prediction, and the mAP of the algorithm is increased by 3.7%.

In 2018, the YOLOv3 detection algorithm was further improved [32]. Based on the Darknet-53 and feature pyramid networks (FPN), multi-scale detection is carried out on three-scale feature graphs, which improves the detection ability and accuracy of small targets. YOLOv3 has no full connection layer, and it is a full convolution network. Softmax is replaced by multi-label classification. The detection speed of the algorithm YOLOv3 is 78.6% when the detection speed is 40 Fps (frames per second) on the VOC207 dataset. On the COCO dataset, the detection speed of 20 Fps can be maintained when the mAP reaches 57.9% (IoU = 0.5). In this article, the HD camera used for data collection is 1280 * 960, which is 1.3 million pixels.

YOLOv3 uses clustering to obtain anchor frames and initializes three anchor frames on three scales. During prediction, each grid will predict 3 bounding boxes, each of which contains 5 parameters and probability of each category. The YOLOv3 network combines multi-scale features, so the detection accuracy and capability of small targets are improved. The prediction method of the bounding box coordinate of YOLOv3 is as follows:

$$\begin{cases} b_x = \sigma(t_x) + c_x, \\ b_y = \sigma(t_y) + c_y, \\ b_w = p_w e^{t_w}, \\ b_h = p_h e^{t_h}, \\ \Pr(\text{object}) * \text{IoU}(b, \text{object}) = \sigma(t_o), \end{cases} \tag{1}$$

where σ is the sigmoid activation function; t_x, t_y, t_w, and t_h are the prediction output of the model; and c_x and c_y are the grid cell coordinates. p_w and p_h are the size of the bounding box before prediction, and t_o is the confidence level of YOLOv3. b_x, b_y, b_w, and b_h are the center coordinates and dimensions of the bounding box obtained from prediction.

Pr (object) indicates whether there is a target in the bounding box of the current grid. If it exists, the value is 1; otherwise, 0 is taken. IoU (b, object) is IoU loss function, which represents the distance between the target and the center of the anchor frame. The K-means clustering method is used.

The YOLOv3 loss function consists of four parts: the center error of the bounding box, the width and height error of the bounding box, the classification error, and the confidence error. Among them, the center error of the bounding box and the width and height error of the bounding box are calculated by sum variance. Suppose there are a set of samples y_i and its estimated value, $\hat{y}_i$, $i = 1, 2, \ldots, n$, then the sum variance is as follows:

$$\text{loss} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{2}$$

Classification error and confidence error are calculated by using the calculation method of binary cross-entropy loss, as shown in the following formula:

$$\text{loss} = \sum_{i=1}^{n} y_i \log\ y_i + (1 - \hat{y}_i) \log\ (1 - \hat{y}_i)^2. \tag{3}$$

The error of the width and height of the bounding box is increased by a scale factor ε. The calculation formula is shown as follows where w and h are width and height, respectively. Because the scale factor can adjust the regression loss, it can detect small targets better.

$$\varepsilon = 2 - w \times h. \tag{4}$$

3. Improved YOLOv3 Algorithm

3.1. Improving the Network Structure of YOLOv3. In the classroom video, the number of students is large, and the face sizes are different in different scenes. Especially for the students sitting in the back row, the face is small, and

sometimes, it will be missed. Therefore, it is necessary to further improve the YOLOv3 algorithm, on the one hand, to build the amount of computation and, on the other hand, to improve the accuracy of face recognition. The improved YOLOv3 network structure introduces feature pyramid networks (FPN) for multi-scale detection. Deep features are used to detect large targets in the network, and shallow features are used to detect small targets, which effectively improves the detection ability of small targets. The improved network structure of YOLOv3 is shown in Figures 2 and 3.

From Figure 3, the YOLOv3 algorithm has been improved as follows:

(1) The dependence of gradient on parameters and its initial value scale is reduced. The network training can be avoided by using a large learning rate and batch normalization (BN). BN improves the generalization ability of network and reduces dropout and optimizes network structure. Spatial pyramid pooling (SPP) is used instead of average pooling at last, which reduces the adverse effect of average pooling on network performance.

(2) In this article, the lightweight network model MobileNet is used for reference, and the depth separated convolution (DSC) is used to reduce the parameters and computation. The width multiplier and resolution multiplier are used to achieve an effective tradeoff between classification accuracy and speed. In view of the different sizes of different targets, combined with the DeepID face recognition algorithm, the attention guidance module is designed to further carry out multi-scale fusion and strengthen the relationship between different size eigenvalues.

(3) In order to solve the problem of unbalanced distribution of positive and negative samples, simple samples, and difficult samples, the super parameter of scaling factor is added and the loss function is improved.

(4) Bayesian is used to optimize the classifier, so as to improve the efficiency and accuracy of classification.

3.2. Deep Separable Convolution. In order to simplify the network model, width multiplier α is introduced to act on the number of channels to reduce the amount of parameters and calculation. The number of channels M in the input layer becomes αM, and the number of channels N in the output layer becomes αN. The total calculation amount of deep separable convolution $\mathrm{N}_{\mathrm{DWS}\text{-}\alpha}$ and the parameter $P_{\mathrm{DWS}-\alpha}$ is shown in the following formulas, respectively:

$$N_{\mathrm{DWS}-\alpha} = H \times W \times \alpha M \times D_K \times D_K + H \times W \times \alpha M \times \alpha N, \quad (5)$$

$$P_{\mathrm{DWS}-\alpha} = \alpha M \times D_K \times D_K + \alpha M \times \alpha N, \quad (6)$$

where the value range of α is (0, 1], and the parameters and calculation amount of the model are reduced by the order of α^2. The resolution multiplier β is introduced to act on the input features, thus reducing the amount of calculation. The input feature changes from $H \times W$ to $\beta H \times \beta W$. After introducing β, the total amount of computation of deep separable convolution is $N_{\mathrm{DWS}\text{-}\beta}$, as shown in the following formula:

$$N_{\mathrm{DWS}-\beta} = \beta H \times \beta W \times \alpha M \times D_K \times D_K + \beta H \times \beta W \times \alpha M \times \alpha N, \quad (7)$$

where the value range of β is (0, 1], which is generally used implicitly by setting the input resolution. It can be seen from formula (7) that the calculation amount is reduced by β^2, and the parameter quantity is independent of the super parameter.

3.3. Improved Loss Function. In single-stage target detection, because the distribution of positive and negative samples is extremely unbalanced, the loss of target detection is easily submerged by a large number of negative samples. The key information provided by a small number of positive samples cannot play a normal role in the loss function, so it is impossible to obtain a loss function that can provide correct guidance for model training. Increasing the weight of cross-entropy loss can effectively solve the problem of sample imbalance. The typical cross-entropy loss is widely used in image classification, $p \in [0, 1]$, which is given as

$$\mathrm{CE}(p, y) = \begin{cases} -\log(p), & y = 1, \\ -\log(1 - p), & \text{otherwise.} \end{cases} \quad (8)$$

where p represents the output class probability of the model and y is the class label.

Formula (8) is improved, supposing

$$p_t = \begin{cases} p, & y = 1, \\ 1 - p, & \text{otherwise.} \end{cases} \quad (9)$$

Considering the imbalance of positive and negative samples in the dataset, the weight is increased by using the coefficient, which is inversely proportional to the probability of the target. In order to reduce the loss of simple samples automatically, a dynamic scaling factor super parameter λ is added to correct the cross-entropy loss by using the weight coefficient γ, which is inversely proportional to the probability of target existence, which is given as

$$\mathrm{FL}(p_t) = -\gamma(1 - p_t)^{\lambda}\log(p_t). \quad (10)$$

3.4. Bayesian Optimization Classifier. According to the video image sequence feature map, we need to establish a classifier to predict the unknown image types. Bayesian classification is based on the Bayesian theory, and it can estimate the posterior probability after training a large number of samples. Suppose that the attribute set of face features to be classified is $X = \{x_1, x_2, \ldots x_m\}$, $C = \{c_1, c_2, \ldots c_n\}$, and $P(c_i)$ is a priori probability. Face recognition is essentially a binary classification problem, $C = \{c_1, c_2\}$, where c_1 is face and c_2 is nonface, and according to the Bayes theorem, the probability is calculated as follows:

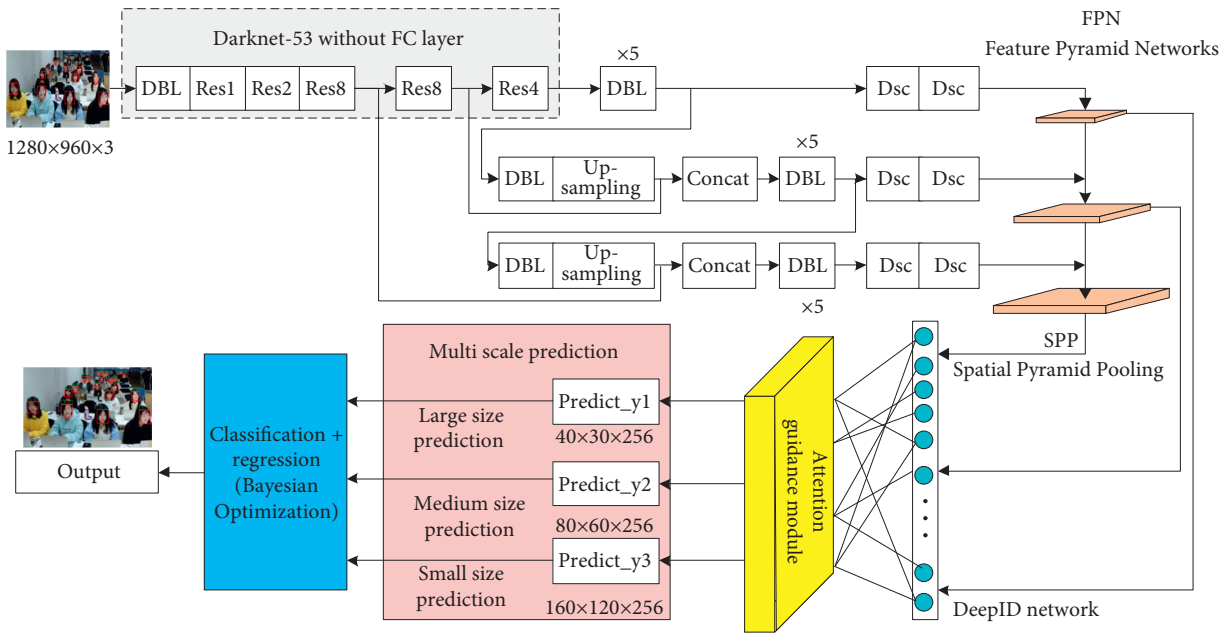


Figure 2: Improved YOLOv3 network structure.

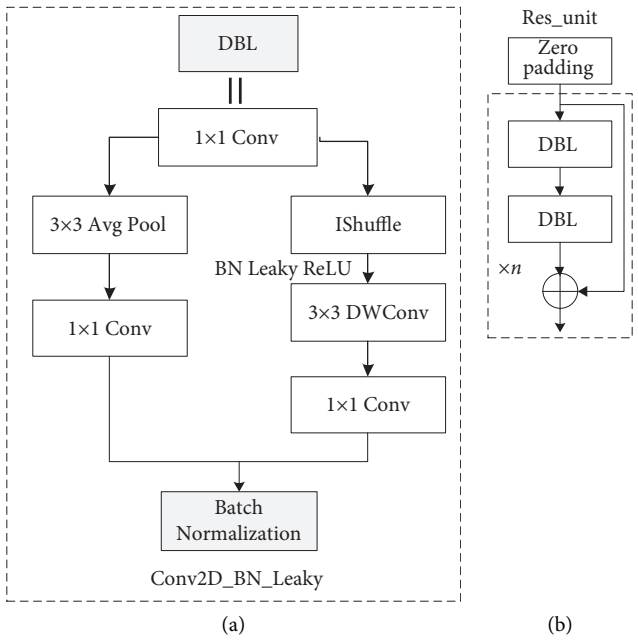


Figure 3: Basic components. (a) Basic components of DBL. (b) Resn components.

$$P(c_1|x),$$
$$P(c_2|x), \quad (11)$$
$$P(c_i|x) = \max\{P(c_1|x), P(c_2|x)\}.$$

The key is to calculate the conditional probabilities and establish the training sample set firstly. The guiding principle of Bayesian classification is that if the probability of feature X belonging to pattern class c_1 is greater than that of feature X belonging to pattern class c_2, then the decision pattern belongs to pattern class c_1, and on the contrary, the decision pattern belongs to pattern class c_2.

If $P(X|c_1) > P(X|c_2)$, then $x \in c_1$, and if $P(X|c_1) < P(X|c_2)$, then $x \in c_2$. When there is an unknown

data sample vector X, the Bayesian method calculates the maximum category of posterior probability. The Bayesian formula is as follows:

$$P(c_i|X) = \frac{P(X|c_i)P(c_i)}{P(X)},$$
$$P(X|c_i) = P(x_1, x_2, \ldots, x_n|c_i) = \prod_{j=1}^{n} P(x_j|c_i). \quad (12)$$

Since $P(X)$ is a constant, only $P(X|c_i)P(c_i)$ needs to be calculated when calculating the posterior probability. $P(c_i) = N/N_c$, where N is the number of training samples, and N_c is the number of training sample categories. The conditional probability estimates of each feature attribute in each category are calculated and recorded. Then, the conditional probability estimates of each feature component under the two categories are calculated. The formula for calculating $P(X|c_i)$ is as follows.

For the unknown sample X category, $P(X|c_i)P(c_i)$ of each category is calculated. The category with the highest probability is the prediction category of sample X, that is,

$$C = \arg\max P(c_i) \prod_{j=1}^{n} P(x_j|c_i). \quad (13)$$

4. Analysis of Experimental Results

In order to verify the feasibility of the algorithm, an experimental system is built. The system version is Ubuntu 16.04LTS (64 bit), the CPU is Intel i7-9750, the graphics card is GeForce GTX1060, and the memory is 8 g. Using the Darknet learning framework, the running environment of the program is *Python* 2.7. This article designs five groups of comparative experiments in VOC2012, COCO datasets, and self-built classroom datasets.

4.1. Comparative Experiments of Different Basic Networks. For the algorithm basic network, the evaluation indexes are Top1 and Top5 error rates. The lower the index value is, the better the classification accuracy of the model is; that is, the better the model is.

$$E_{\text{Top1}} = \frac{\text{the number of samples with real markers different from Top1}}{\text{total samples}},$$
$$E_{\text{Top5}} = \frac{\text{the number of samples with real markers different from Top5}}{\text{total samples}}. \quad (14)$$

The VOC2012 dataset is used as the test dataset, and the comparison of calculation results of different basic networks is shown in Table 1.

It can be seen from Table 1 that the error rate of the improved YOLOv3 operation is reduced, and the deep separation convolution operation can greatly reduce the flops and complexity without losing the model capacity. Compared with GoogleNet and Darknet-53, the Top1 error rate of the model is reduced by 1.1% and 0.13%, respectively, and the Top5 error rate of the model is increased by 1.8% and 0.45%, respectively, and the parameters and FLOPs (floating point operations per second) is 90.4% and 89.1% of GoogleNet.

Taking COCO dataset as an example, the comparison of calculation results of different methods is shown in Table 2.

It can be seen from Table 2 that compared with other methods, the method can increase the depth of convolution and attention guidance module, which makes the detection speed and parameter amount slightly decrease. However, under the same IoU, the leakage rate of this method is the smallest and the average accuracy is higher than 1.2% of other methods, which indicates that the method has certain advantages.

In the VOC2012 dataset, FPN comparative experiments are carried out to verify the improvement of detection effect by adding the FPN algorithm in YOLOv3. The mAP and Fps of the target detected by the algorithm without FPN are 32.6 and 57.2, respectively, whereas the mAP and Fps of the target detected by the algorithm with FPN are 34.8 and 55.4, respectively. Using the FPN-improved YOLOv3 method for multi-scale detection, due to the addition of two scales and attention guidance module to assist in feature enhancement, the detection accuracy of the algorithm is improved to a certain extent. Compared with the method without FPN, the mAP index is improved by 2.2%, and the Fps is reduced by 1.8 due to the use of larger features, but the detection speed is still high.

4.2. Improving the Comparison Experiment of Loss Function. The loss functions commonly used in SSD, YOLOv1, and YOLOv2 models are used in the experiment. The improved local loss function in this article is proposed and tested under the IoU index of 0.5 and confidence of 0.5 in the VOC2012 dataset. The average accuracy of the test is shown in Table 3.

It can be seen from Table 3 that the detection accuracy index mAP of the improved local loss function is enhanced, compared with the conventional loss function and IoU loss function used in YOLOv3. Compared with the original loss function, $\text{mAP}_{0.5}$ and mAP of the improved algorithm are improved by 2.3% and 2.7%, respectively.

It can be seen from Figure 4 that the improved loss function tends to zero after 500 iterations, and the accuracy is higher than loss and IoU.

Table 1: Comparison of calculation results of different basic networks.

Model	Top1 error rate (%)	Top5 error rate (%)	Parameter (MB)	FLOPs
MobileNet	31.22	9.08	14.83	22.63
VGG19	32.37	9.05	15.75	26.23
GoogleNet	32.15	9.02	16.78	26.15
Resnet	32.52	9.03	18.08	25.06
Darknet-19	31.29	8.96	13.93	22.06
Darknet-53	31.18	8.89	14.63	22.28
Ours	31.05	8.85	15.17	23.31

Table 2: Comparison of calculation results of different methods.

Algorithm	IoU = 0.5 missed detection rate	mAP (%) IoU = 0.5	AP (%)			Detection speed (Fps)	Parameter (MB)
			IoU = 0.6	IoU = 0.7	IoU = 0.8		
Fast R-CNN	15.8	84.1	81.2	79.9	76.8	45.5	15.32
FaceNet	11.3	88.6	86.3	83.8	81.2	36.6	18.61
DeepID	10.8	89.2	86.5	83.9	82.1	38.3	17.85
SSD	12.2	87.8	85.3	82.9	80.4	40.5	14.83
YOLOv1	9.8	90.1	88.2	86.5	83.9	41.3	15.65
YOLOv2	9.5	90.5	89.3	87.5	85.2	43.7	16.21
YOLOv3	8.9	91.0	90.1	89.5	88.7	46.2	17.52
Ours	7.8	92.2	91.3	90.9	90.1	45.1	18.35

Table 3: Comparison test results of three loss functions.

Method	$mAP_{0.5}$	mAP
Loss	52.2	32.8
IoU	51.6	33.5
Improved local loss	54.5	35.5

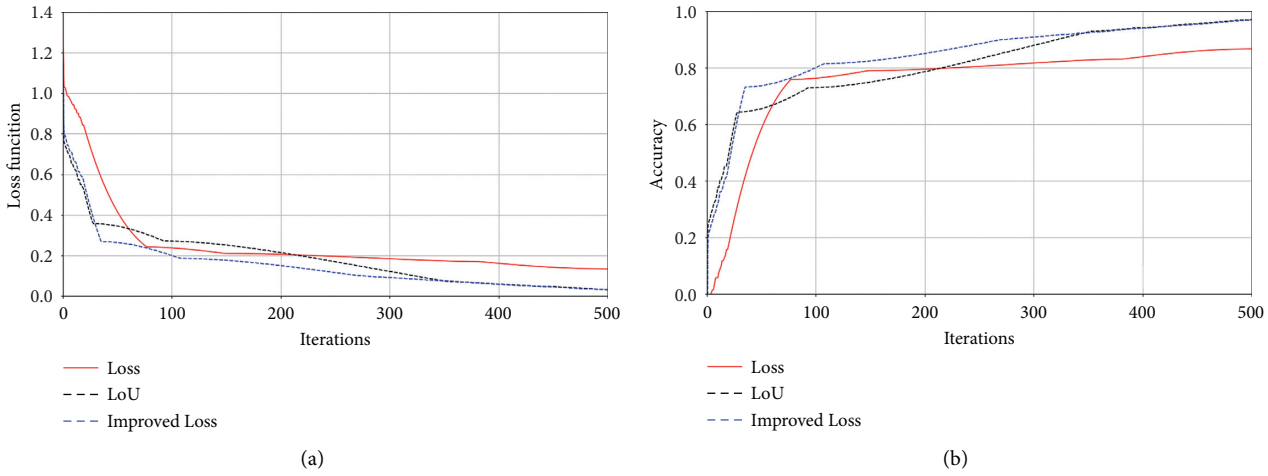


Figure 4: Curve of loss function and accuracy.

4.3. Performance Comparison Experiment of Different Test Methods. Each target detection algorithm is tested on VOC2012 dataset and COCO dataset, and the ROC curve and recall curve are used for comparative analysis.

It can be seen from Figure 5 that the recognition accuracy and success rate of this algorithm are higher than those of other methods, due to using attention guidance module and Bayesian Optimization classifier. The ROC curve area of this algorithm is 0.845, which is 0.015 more than the SSD algorithm and 0.01 more than the YOLOv3 algorithm. The recognition success rate is 83.7%, which is 1.5% and 0.6% higher than the SSD algorithm and YOLOv3 algorithm, respectively.

The recall rate is analyzed in VOC2012 dataset, and the curve is shown in Figure 6. As can be seen from Figure 6, this method adopts the improved YOLOv3 algorithm to realize

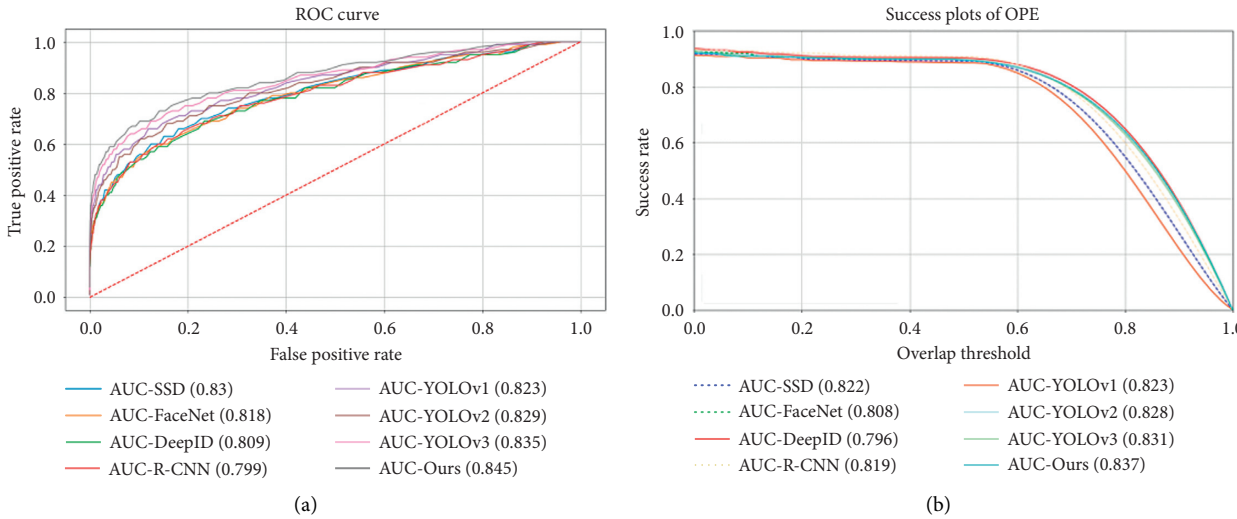


Figure 5: ROC and success rate curve of VOC2012 dataset.

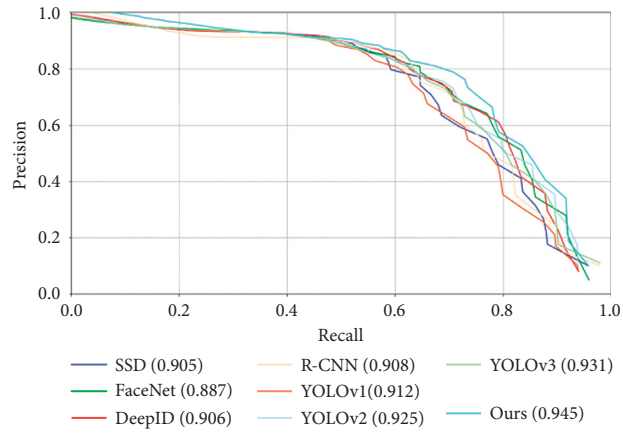


Figure 6: Recall curve of VOC2012 data set.

the recognition of multi-scale face target and uses Bayesian to optimize the classification method. So the recall rate is higher than other methods, which is about 4% and 1.4% higher than that of SSD and YOLOv3, respectively.

The test results in the COCO dataset are shown in Table 4.

From the experimental results in Table 5, it can be seen that the above algorithms can better complete the task of student detection. Fast R-CNN algorithm is not accurate in positioning and will miss detection. The accuracy of the YOLO algorithm is very high, but there will still be some missed detection. Compared with the traditional YOLO method, YOLOv3-based Bayesian optimization can increase the depth separation convolution and feature pyramid and improve the recall rate of 1.8% and the accuracy of 2.21%, but the average detection speed is reduced by 0.4. This algorithm uses the DeepID face detection model for reference, which reduces the average detection time, but the accuracy and recall rate are improved, and the comprehensive performance of the improved YOLOv3 algorithm is enhanced.

4.4. Image Scaling Experiment. Firstly, the original image (1280 * 960) is compressed into a low-pixel image (320 * 240) from the video sequence, and then, the low-pixel image is reconstructed into (640 * 480), (960 * 720) size images by using the algorithm in this article. By comparing and predicting the number of face frames, the face detection recall rate under different image scales is calculated.

As can be seen from Figure 7 and Table 6, the original image pixel is 1280 * 960, and the face detection recall rate is 98.5%. The compressed low-pixel image pixel is 320 ∗ 240, and the face detection recall rate is only 25.6%. The experimental results show that with the increase of the image reconstruction scale, the face detection recall rate increases from 25.6% to 98.5%.

4.5. Experimental Analysis of Self-Built Data Set. The test effect of the self-built classroom dataset is shown in Figure 8.

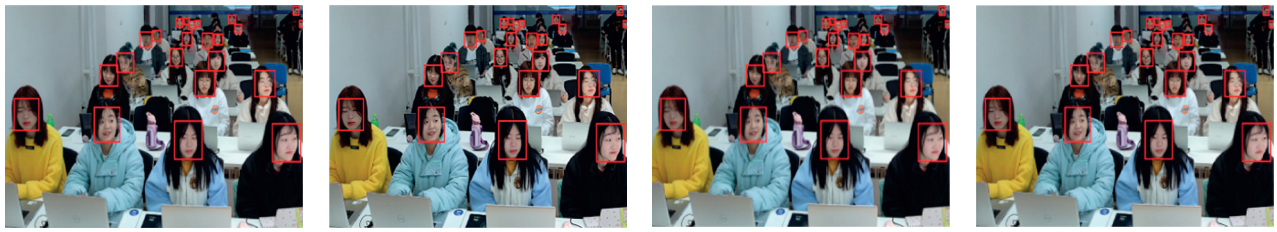
The experiment adopts a medium-sized classroom scene, the length and width of the classroom is 18 * 9 meters, which can accommodate up to 120 students. The

Table 4: Test results in COCO dataset.

Algorithm	Total number	Accurate detection number	Total detection number	Accuracy rate (%)	Recall (%)	Detection speed (Fps)	Parameter (MB)
Fast R-CNN	500	445	590	75.42	89.00	38.5	15.32
FaceNet	500	437	589	74.19	87.40	35.4	18.61
DeepID	500	456	601	75.87	91.20	35.6	17.85
SSD	500	453	586	77.30	90.60	37.8	14.83
YOLOv1	500	455	597	76.21	91.00	36.2	15.65
YOLOv2	500	463	585	79.15	92.60	37.5	16.21
YOLOv3	500	470	592	79.39	94.00	38.3	17.52
Ours	500	479	587	81.60	95.80	37.9	16.55

Table 5: Influence of different scales on the recall rate of face detection.

Scale	Recall (%)
320 * 240	25.6
640 * 480	53.8
960 * 720	86.2
1280 * 960	98.5



(a) (b) (c) (d)

Figure 7: Face detection results of this algorithm. (a) Original image $(1280 * 960)$. (b) A low-pixel image $(320 * 240)$. (c) Image reconstructed by $(640 * 480)$. (d) Image reconstructed by $(960 * 720)$.

Table 6: Test results of self-built classroom dataset.

Algorithm	45 students				100 students			
	Accurate number	Accuracy rate (%)	Detection speed (Fps)	Parameter (MB)	Accurate number	Accuracy rate (%)	Detection speed (Fps)	Parameter (MB)
Fast R-CNN	32	86.49	4.92	57.3	66	88.00	9.85	93.5
FaceNet	33	89.19	7.35	55.9	70	93.33	12.21	95.1
DeepID	34	91.89	6.56	56.1	72	96.00	11.52	91.7
SSD	34	91.89	5.82	53.2	69	92.00	11.28	87.5
YOLOv1	32	86.49	6.23	56.6	71	94.67	11.32	92.3
YOLOv2	33	89.19	5.37	54.5	70	93.33	11.15	89.4
YOLOv3	34	94.59	5.12	55.3	72	96.00	10.59	90.6
Ours	36	97.30	5.35	54.2	73	97.33	11.26	90.1

shooting position is located at the top left of the platform, facing the students. Considering the angle, the distance between the students and the camera is about 3 m~18 m. During face recognition, we need to consider the environment, light, distance factors, the farthest, the face size of students is small. However, the size of face has less impact on face detection, but more impact on face recognition and head-up rate. Because face detection is a two-classification problem, face recognition is a multi-classification problem. Affected by the environment, the algorithm is more complex, and the lack of information will have a great impact on face feature extraction. Firstly, the face state analysis experiments of multiple groups of classrooms were carried out, with 110 people as the benchmark. The statistical results are shown in Figure 9.

It can be seen from Figure 9 that only 50.91% of the total number has good detection and recognition conditions in a large classroom. Too much illumination will lead to face

(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 8: Test effect of self-built classroom dataset.

features covered by strong light, which is more common in outdoor scenes; occlusion will lead to incomplete face features, and only part of the face features can be obtained. Bow head and small face also account for 26%.

In five self-built classroom datasets, the number of students is 45 and 100. The actual number of students in 45-student classroom is 37 and that in 100-student classroom is 75. The test results are shown in Table 6.

It can be seen from Table 5 that the accuracy of this method is higher than that of other methods. In the 45-student classroom, this method is 5.4% higher than the SSD method and 2.71% higher than the YOLOv3 method. In the 100-student classroom, this method is 5.33% higher than the SSD method and 1.33% higher than the YOLOv3 method. Although the detection speed of this method is slightly lower than that of fast R-CNN, SSD, and YOLOv3, it is still faster than other methods.

4.6. Comparison with Other Neural Network Depth Quantization Methods. Furthermore, in order to verify the training ability of the model, the image prediction method proposed in this article needs to be compared with other model prediction methods. In automatic reinforcement learning, quantization is widely used as an important means of image compression, and the contradiction between bit width and accuracy always exists. At present, the Bayesian optimizer proposed in this article determines the bit width. Another problem is the selection of quantization value, which is obtained by alternating training in LQ-Net. Deep reinforcement learning (RL) is good at mapping original sensory input to action [33], while AutoRL is an evolutionary automation layer around deep RL, which uses large-scale hyperparametric optimization to search reward and neural network architecture. The AutoRL method has been used in long-distance robot navigation, multi-stage prediction of microgrid, and so on [34].

In order to verify the effectiveness of this method, the quantization bit width of automatic reinforcement learning is determined by using ReLeQ, AutoRL framework, and this

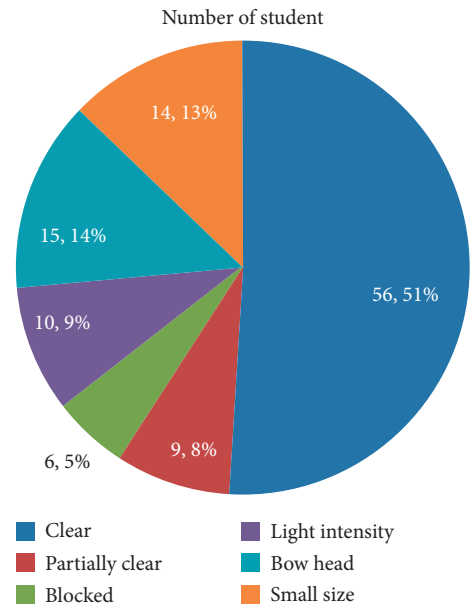


Figure 9: Statistical chart of face state classification of.

method separately. The experiments in three experimental data sets are shown in Table 7.

It can be seen from Table 7 that the Bayesian optimized image-type prediction method proposed in this article can determine the quantization bit width, and the network recognition accuracy is slightly higher than that of the AutoRL method, indicating the superiority of this method in image classification prediction.

Furthermore, extreme learning machine (ELM) is based on feed forward neural network (FNN) [35], or an improvement of FNN and its back propagation algorithm. Its characteristic is that the weight of hidden layer nodes is random or artificially given, and does not need to be updated. The learning process only calculates the output weight. ELM has strong generalization ability and high accuracy in approximating datasets. It

Table 7: Comparison of quantization for different networks.

Dataset	Average quantized bit width weights	Network accuracy		
		Ours (%)	ReLeQ (%)	Auto RL (%)
VOC2012	4	94.3	92.9	94.2
COCO	3	92.2	90.5	92.0
MINIST	2.75	96.9	95.8	96.9

has been successfully used in the transformation from low-resolution to high-resolution images. In the COCO dataset, the model prediction accuracy of this method, and methods described in literature [35] and literature [36] has reached 96.5%, 95.9%, and 96.3%, respectively, indicating that the image prediction method optimized by Bayesian can achieve high accuracy.

5. Conclusion

In order to solve the problem of low accuracy of the traditional object detection model, a new YOLOv3 model based on Bayesian optimization is proposed. The depth integral separation convolution is used to replace the standard convolution for information fusion, which reduces the amount of network structure parameters and calculation. The feature pyramid is used to replace the network full connection layer, and the attention guidance module is designed to enhance the multi-scale feature fusion ability and reduce the overfitting weight coefficient. A dynamic scaling factor super parameter is added to the loss function to improve the imbalance between positive and negative samples, and the Bayesian method is used to optimize the classification. Compared with the traditional YOLOv3 algorithm, the recall rate of the optimized YOLOv3 model is improved from 94% to 95.8%, and the average mAP value is increased by more than 1.2%. It shows that the optimization method of the training process in this article enhances the training effect of the model. The algorithm in this article realizes the synchronous improvement of the detection accuracy and improves the real-time detection of multi-face small target in the classroom.

5.1. Future Prospects. Recently, YOLOv4 and YOLOv5 are released, and these algorithms will be the next research goal of our team. We will combine with more efficient network structure and model to research deeply, and the research results will be applied to classroom face recognition, driving safety behavior recognition and other aspects.[36]

Data Availability

At present, the data are still in the experimental stage and cannot be disclosed temporarily.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Jiangsu Natural Science Foundation of China (Project no. BK20191225) and the second batch of production-university-research cooperation bases in Suzhou Higher Vocational College in 2020 (Project no. 2020-5).

References

[1] R. Hartanto and N. Marcus, "Face recognition for attendance system detection," in *Proceedings of the 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pp. 376–381, Bali, Indonesia, July 2018.

[2] D. Shi and H. Tang, "Research on safe driving evaluation method based on machine vision and long short-term memory network," *Journal of Electrical and Computer Engineering*, vol. 2021, pp. 1–13, Article ID 9955079, 2021.

[3] Y. Gao, C. H. Zhou, and F. Z. Su, "Study on SVM classifications with multi-features of OLI images," *Engineering of Surveying & Mapping*, vol. 47, no. 11, pp. 3084–3086, 2014.

[4] W. H. Tian, K. M. Zeng, and Z. Q. Mo, "Driver unsafe behavior recognition based on convolutional neural network," *Journal of University of Electronic Science and technology*, vol. 48, no. 3, pp. 381–387, 2019.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 8, pp. 84–90, 2017.

[6] Y. Zhang, Z. Y. Gong, and W. W. Wei, "Traffic sign detection based on improved faster R-CNN model," *Laser & Optoelectronics Progress*, vol. 57, no. 18, pp. 181015–181021, 2020.

[7] S. F. Liang, Y. H. Liu, and L. C. Li, "Face recognition under unconstrained based on LBP and deep learning," *Journal on Communications*, vol. 35, no. 6, pp. 154–160, 2014.

[8] Y. Wang, X. J. Shen, and H. P. Chen, "Multi instance learning video face recognition algorithm based on improved Fisher criterion," *Acta Automatica Sinica*, vol. 44, no. 12, pp. 69–77, 2018.

[9] H. X. Zhang, G. Zou, and J. Zhao, "Face recognition algorithm based on Gabor feature and collaborative representation," *Computer engineering and design*, vol. 35, no. 2, pp. 666–670, 2014.

[10] H. Tan, B. Yang, and Z. Ma, "Face recognition based on the fusion of global and local HOG features of face images," *IET Computer Vision*, vol. 8, no. 3, pp. 224–234, 2014.

[11] W. J. Li, J. Wang, Z. H. Huang, T. Zhang, and D. K. Du, "LBP-like feature based on Gabor wavelets for face recognition," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 15, no. 5, 2017.

[12] Y. Zhu and J. K. Zhao, "A review of human action recognition based on deep learning," *Acta Automatica Sinica*, vol. 42, no. 6, pp. 848–857, 2016.

[13] S. Srinivasan, V. Ravi, V. Sowmya et al., "Deep convolutional neural network based image spam classification," in *Proceedings of the 2020 6th Conference on Data Science and*

Machine Learning Applications (CDMA), Riyadh, Saudi Arabia, March 2020.

[14] H. W. Mo and H. B. Wang, "Research on human behavior detection based on Faster R-CNN," *CAAI Transactions on Intelligent Systems*, vol. 13, no. 6, pp. 107–113, 2018.

[15] S. Parvathi and S. T. Selvi, "Detection of maturity stages of coconuts in complex background using Faster R-CNN model," *Biosystems Engineering*, vol. 202, pp. 119–132, 2021.

[16] W. Pei and Y. M. Xu, "The target detection method of aerial photography images with improved SSD," *Journal of Software*, vol. 30, no. 3, pp. 738–758, 2019.

[17] D. W. Liu, S. Gao, W. D. Chi, and D. Fan, "Pedestrian detection algorithm based on improved SSD," *International Journal of Computer Applications in Technology*, vol. 65, no. 1, 2021.

[18] F. Xia and H. Li, "Fast detection of airports on remote sensing images with single shot MultiBox detector," *Journal of Physics: Conference Series*, vol. 960, Article ID 012024, 2018.

[19] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multibox detector," in *Proceedings of the European Conference on Computer Vision-ECCV 2016*, pp. 21–37, Springer, Amsterdam, The Netherlands, October 2016.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[21] H. L. Luo and H. L. Chen, "Survey of object detection based on deep learning," *Acta Electronica Sinica*, vol. 48, no. 6, pp. 1230–1239, 2020.

[22] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, Venice, Italy, October 2017.

[23] J. Yi, P. Wu, and D. N. Metaxas, "ASSD: attentive single shot multibox detector," *Computer Vision and Image Understanding*, vol. 189, pp. 102827–102835, 2019.

[24] J. Zhou, Y. Tian, C. Yuan, K. Yin, G. Yang, and M. Wen, "Improved UAV opium poppy detection using an updated YOLOv3 model," *Sensors*, vol. 19, no. 22, pp. 4851–4810, 2019.

[25] Li Liu, Y. Zheng, and D. M. Fu, "Occluded pedestrian detection algorithm based on improved network structure of YOLOv3," *Pattern Recognition and Artificial Intelligence*, vol. 33, no. 6, pp. 568–574, 2020.

[26] L. Zhao and S. Li, "Object detection algorithm based on improved YOLOv3," *Electronics*, vol. 9, no. 3, pp. 537–545, 2020.

[27] J. Y. Hu, C. J. R. Shi, and J. S. Zhang, "Saliency-based YOLO for single target detection," *Knowledge and Information Systems*, vol. 2021, no. 1, pp. 1–16, 2021.

[28] J. X. Zhu, X. Li, P. Jin, Q. Xu, Z. L. Sun, and X. Song, "MME-YOLO: multi-sensor multi-level enhanced YOLO for robust vehicle detection in traffic surveillance," *Sensors*, vol. 21, no. 1, 2020.

[29] S. Zhang, Y. Wu, C. Men, and X. Li, "Tiny YOLO optimization oriented bus passenger object detection," *Chinese Journal of Electronics*, vol. 29, no. 1, pp. 132–138, 2020.

[30] X. Zhang, X. Dong, Q. Wei, and K. Zhou, "Real-time object detection algorithm based on improved YOLOv3," *Journal of Electronic Imaging*, vol. 28, no. 5, pp. 1–10, 2019.

[31] J. Liu and D. Zhang, "Research on vehicle object detection algorithm based on improved YOLOv3 algorithm," *Journal of Physics: Conference Series*, vol. 1575, no. 1, Article ID 012150, 2020.

[32] A. T. Elthakeb, P. Pilligundla, F. Mireshghallah, A. Yazdanbakhsh, and H. Esmaeilzadeh, "ReLeQ: a reinforcement learning approach for automatic deep quantization of neural networks," *IEEE Micro*, vol. 99, p. 1, 2020.

[33] Y. Li, R. Wang, and Z. Yang, "Optimal scheduling of isolated microgrids using automated reinforcement learning-based multi-period forecasting," *IEEE Transactions on Sustainable Energy*, p. 1, 2021.

[34] Y. Li and Z. Yang, "Application of EOS-ELM with binary jaya-based feature selection to real-time transient stability assessment using PMU data," *IEEE Access*, vol. 5, pp. 23092–23101, 2017.

[35] Y. Zhang, T. Li, G. Na, G. Li, and Y. Li, "Optimized extreme learning machine for power system transient stability prediction using synchrophasors," *Mathematical Problems in Engineering*, vol. 2015, no. 1, 8 pages, Article ID 529724, 2015.

[36] H. Liu, T. Xu, X. Wang, and Y. Qian, "Related HOG features for human detection using cascaded adaboost and SVM classifiers," in *Proceedings of the the International MultiMedia Modeling Conference*, pp. 345–355, Springer, Huangshan, China, January 2013.