Hindawi
Advances in Multimedia
Volume 2022, Article ID 5785108, 11 pages
https://doi.org/10.1155/2022/5785108

Research Article

Face Detection Method based on Lightweight Network and Weak Semantic Segmentation Attention Mechanism

Xiaoyan Wu

Sichuan University of Arts and Science, Dazhou 635000, China

Correspondence should be addressed to Xiaoyan Wu; 20040026@sasu.edu.cn

Received 14 January 2022; Revised 24 February 2022; Accepted 28 February 2022; Published 23 May 2022

Academic Editor: Qiangyi Li

A face detection method based on lightweight network and weak semantic segmentation attention mechanism is proposed in this paper, aiming at the problems of low detection accuracy and slow detection speed in face detection in complex scenes. K-means++ algorithm is employed to perform clustering analysis on YOLOv4 model prior frames in this paper, and smaller size prior frames are set to capture small face information to solve the missing detection problem of small face targets in scenes. The backbone network structure is improved by introducing Mobile Net lightweight network model, to reduce the number of parameters and calculation of the model and improve the detection speed. The convolutional block attention module model with dual attention mechanism is embedded to improve the sensitivity of the model to target features, which can suppress interference information and improve the accuracy of target detection. A dynamic enhancement attachment based on weak semantic segmentation is added in front of the detector head, whose output is used as the spatial weight distribution to correct the activation area, to suppress the false detection and missed detection caused by the decrease of extraction ability evoked by the pursuit of lightweight. The experimental results on WIDEFACE dataset indicate that this method not only can detect face in real time and with high accuracy, but also has better performance than other existing methods.

1. Introduction

Face detection technology has important theoretical research significance and wide application value in the field of computer vision. Since the small human face target has less pixels and less obvious features, and its recall rate is low compared with the large target, how to improve the detection accuracy and model robustness has become a crucial problem [1].

The target detection methods based on convolution neural network are mainly divided into two categories [2]. The first is two stage target detection algorithms, which is divided into two parts: regional selection and positioning regression, mainly represented by Region Convolutional Neural Network (R-CNN) [3], such as Fast R-CNN [4] and Faster R-CNN [5], SPPNET [6], R-FCN [7], and mask R-CNN [8]. However, the delay of the two-stage method is higher. The second is the single step detection algorithm. The probability and position coordinates of the target can be obtained directly by regression. Classic algorithms include RetiaNet, Single Shot MultiBox Detector (SSD) [9] and You Only Look Once (YOLO) series [10], such as YOLOv2, YOLOv3, and YOLOv4. In order to deal with multi-scale or small objects, YOLO series proposes a new anchor frame matching strategy, and reweights the width and height of the object. SSD uses multi-level feature map combination structure, while FPN introduces the characteristic pyramid structure.

In addition to the new methods proposed by general object detection, the development of other fields also promotes face detection. Nonmaximum Suppression (NMS) is a common postprocessing method for target detection [11], which is used to solve the problem of redundant prediction frame for the same target in detection. The detection frame with high confidence is extracted, and the detection frame with low confidence is suppressed, so that the repeated frame is removed, and the correct detection frame is obtained. The common NMS method will lead to false

suppression when multiple prediction frames overlap. The combination of Diou and NMS can further consider the location information of the center point of the two frames, which makes the prediction box more realistic. Activation function is the key to the introduction of nonlinearity into deep neural networks. Currently, the most used activation function in neural networks is Rectified Linear Unit (RELU). However, the RELU function has the problems of hard zero boundary and too simple nonlinear processing, while the Mish [12] activation function is smoother, which can effectively alleviate the hard zero boundary problem. Attention mechanism can identify the key features in image data in the field of target detection. The deep neural network can learn the areas that need attention in each image through learning and training, to select the information that is more critical to the current mission objectives from a large number of information. Convolutional Block Attention Module (CBAM) [13] is a simple and effective attention module for feedforward neural networks. Given intermediate feature mappings, attention mappings are derived sequentially along both channel and space dimensions, and then the attention mappings are multiplied onto the input feature mappings for feature adaptive learning.

Face is a special target. In order to obtain better detection results, researchers have improved and optimized the target detection algorithm, and put forward many excellent face detection algorithms. CMS-RCN [14] is improved based on fast R-CNN and integrates human context information to improve face detection performance. Multitask Cooperative Neural Networks (MTCNN) [15] adopt the cascade mode of three separate network modules. The detection of key points is added to the network, which is conducive to face detection. Fang et al. [16] combined Fast R-CNN with syntax guided network (SG-Net) to fuse the generated image with the original convolution features, enhancing the focus on the face area in the feature map. The algorithm can effectively realize face detection. Zhang et al. proposed a multiscale face detection method [17] which improved the SSD framework and had good performance for faces of different scales, especially for small scale faces. These methods have achieved good detection results under controllable conditions. However, when the face information is insufficient, and the size is small in complex scenes, the accuracy of these face detection methods is relatively low.

YOLOv4 is improved in this paper to improve the accuracy of small face detection. The algorithm introduces the mobile net lightweight network model, improves the backbone network structure of YOLOv4 model, and reduces the number of parameters and calculation of yolov4 model. Simultaneously, it embeds convolutional block attention module (CBAM) model with dual attention mechanism to improve the sensitivity of YOLOv4 model to target features. The proposed algorithm ensures the detection speed, improving the detection accuracy and the detection ability of small targets. Section 2 is a description of related work. Section 3 describes the algorithm in detail. Section 4 is the analysis of experimental data. Section 5 is the conclusion.

2. Related Work

2.1. YOLOv4 Target Detection Algorithm. YOLOv4 [18] algorithm is a target detection algorithm based on the YOLO target detection architecture, which adopts the excellent optimization strategy in the field of convolutional neural network (CNN). The algorithm is optimized in data processing, backbone network, network training, activation function, and loss function, so that anyone can use a 1080Ti or 2080 Ti GPU (graphs processing unit) to train a super-fast and accurate target detector. YOLOv4 verified the influence of a series of mainstream target detector training methods and modified these mainstream methods to make them more effective and adaptive when training with a single GPU, including cross iteration batch normalization, CBN, path aggregation network (PANet), and spatial attention module (SAM). There are two kinds of training methods in YOLOv4: (1) bag of freebies (BOF), which only changes the training strategy or only increases the training cost, such as data enhancement. (2) bag of specials (BOS), including plug-in modules and post-processing methods, which only increases a small amount of reasoning cost, but can greatly improve the accuracy of target detection.

2.2. MobileNetv3. Convolution layer is the most time-consuming structure for data flowing through neural network. In 2017, Andrew g. Howard et al. proposed a lightweight model for mobile devices, MobileNetv1. This algorithm redesigns the convolution strategy, which is called depthwise separable convolution (DW). The model parameters and computation are reduced via using DW convolution to replace the traditional convolution. The specific quantitative formula is as follows:

$$\frac{\mathrm{Cal}_{dw}}{\mathrm{Cal}_{\mathrm{conv}}} = \frac{1}{\mathrm{Chal}_{\mathrm{out}}} + \frac{1}{K}, \tag{1}$$

where Cal_{dw} and $\mathrm{Cal}_{\mathrm{conv}}$ represent the computational complexity of separable convolution and traditional convolution respectively. $\mathrm{Chal}_{\mathrm{out}}$ and K represent the number of channels of the output feature and the size of the convolution kernel. In the next two iterations, inverted residuals and attention module based on squeeze and exception (SE) mechanism were integrated into the two iterations to form MobileNetv3.

2.3. Semantic Segmentations. The semantic segmentation task classifies the whole image pixels one by one by using semantic labels to obtain the segmented image with semantic information. It is not until the emergence of fully convolutional network (FCN) that semantic segmentation has a milestone breakthrough. It abandons the sliding window-based method and uses a full convolution network instead, and establishes the model architecture of encode-decode, which overcomes the low precision and low efficiency of the region annotation method. Later, Mask R-CNN is based on the classical target detection method Faster R-CNN, and an

additional branch is added in the detection header as the output of semantic segmentation. By sharing deep features with classification and regression branches, the proposed method modifies the network modularized and extends the target detection algorithm to jump in semantic segmentation tasks. The excellent performance of Mask R-CNN on MS COCO common data set indicates that different tasks in a network can complete joint learning by sharing deep features and using different training tags and loss functions.

In addition to classical semantic segmentation methods, there are also semantic segmentation methods based on attention mechanisms, such as CCNet, DANet, A2-Nets, PGNet, SAGNN, and CMN. CCNet uses two serial cross-attention modules. DANet uses spatial and channel attention modules. A2-Nets uses self-attention mechanisms. ACFNet is a coarse-fine segmentation network based on the attention class feature module. The model PGNet (Pyramid Graph Networks) represents features as graph structures and uses attention weighting to establish relationships between graph nodes. SAGNN designs a multiscale graph neural network structure. In the network design of the model Cyclic Memory Network (CMN), the same structure of self-attentive mechanism like the IEM module is used to aggregate the relationships between multi-scale features.

In a recent work, Classifier Weight Transformer (CWT) proposes a simple and novel transformer structure. This structure dynamically migrates the classifier weights trained in the support set to the query set images for prediction, effectively reducing the intraclass differences between the support set images and the query set images. In addition, other semantic segmentation methods based on the transformer structure include SETR, Trans4trans, and SegFormer.

3. Improved YOLOv4 Model

The improved YOLOv4 model integrates lightweight network and dual attention mechanism. And its overall structure is shown in Figure 1. K-means++ clustering algorithm is used to recluster the prior frame. According to the given data set samples, the anchor frames of similar samples are classified into one class through distance calculation, and the anchor frames suitable for the data sets are obtained, to improve the model learning ability. In addition, this paper introduces the Mobile Net lightweight network model to replace the backbone network CSPDarknet53 of YOLOv4 model. This method can effectively reduce the model parameters and improve the model detection speed without loss of accuracy. The CBAM model with dual attention mechanism is introduced to expand the range of network perception and make the network more sensitive to the detection target by paying attention to the dual information of channel and space, to improve the problem of missing detection caused by poor image quality and false detection evoked by the lack of obvious edge features between faces.

3.1. Improved MobileNetv3 Lightweight Network. To solve the problem of slow model detection due to redundancy caused by image clipping, MobileNetv3-large is further optimized in this paper. Specific approach: retain the first fourteen layers of the original model and discard the rest of the structure. In order to avoid introducing more convolution to align the number of input channels of feature fusion, the number of output channels of the eighth layer was changed from 80 to 40, and the number of channels of the last layer was changed from 160 to 112. The specific network structure is shown in Table 1. In order to suppress the drastic effect of the reduction of feature channels on the model learning ability without increasing the amount of computation, the important module SENet in MobileNetv3 was improved and enhanced, and MobileNetv3-lite was obtained.

The original SENet module transforms the original feature channel into a one-dimensional vector, which can represent the global receptive field by compressing the input features in the spatial dimension. The calculation principle is as follows:

$$T_c = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} f(x, y), \tag{2}$$

where $f(x, y)$ represents the value of each pixel on the feature map. H and W represent the length and width of the characteristic graph respectively. T_c is the real number representing the channel? However, SENet uses global average pooling (GAP) as the compression mechanism. For tasks with balanced target share, the average value of channel characteristics can better represent the response of the channel, thereby obtaining the global context relationship through this method. Average pooling of features in the spatial dimension will cause the background and interference information to drown the foreground, resulting in distorted response of the network to the foreground target. Since the target is small, global max pooling (GMP), which is generally used as texture extraction, will screen out the area with the strongest signal, better reflecting the response of the channel to the foreground target. The calculation principle is as follows:

$$T_c' = \max_{x,y \in H,W} f(x, y). \tag{3}$$

Based on this, the compression mechanism of SENet is redesigned in this paper, and the EN-SENet module is obtained by using two real numbers to characterize the global sensory field. Specifically, GAP and GMP are used to compress the features of each channel respectively, and the pooled results are spliced in the channel direction to obtain a vector with the dimension of $1 \times 1 \times 2C$. Then, the vector is sent to the fully connected network to obtain the channel attention weight Ac.

3.2. K-Means++ Reclustering. In the process of target detection using YOLOv4 model, it is difficult for the pregenerated anchor frame size of the object to be detected to be fully suitable for different detection objects, which will affect the generalization ability of network training and learning.

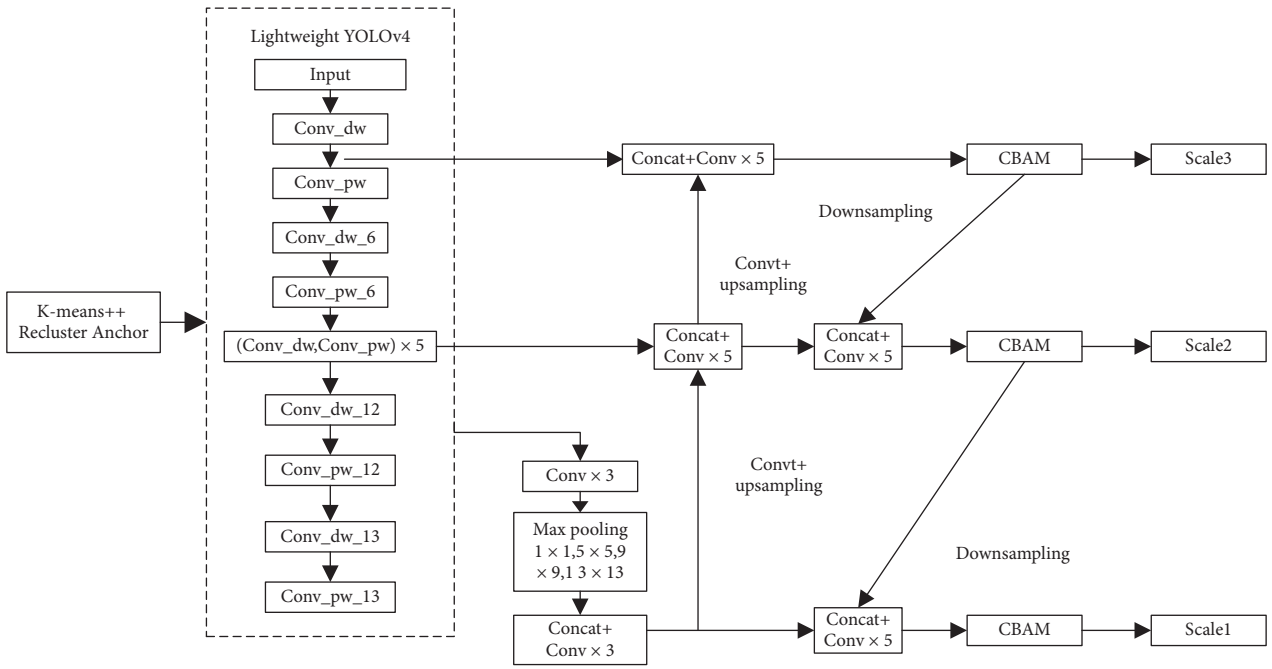

Figure 1: The overall structure of the improved YOLOv4 model integrating lightweight network and dual attention mechanism.

In this paper, K-means++ is used to cluster the width and height of the real face frame in the data set WIDERFACE, and nine anchor boxes with width and height combinations suitable for the data set are generated. K-means++ adopts Euclidean distance. The larger the candidate box, the larger the error. Therefore, YOLOv4 uses the intersection and union ratio IoU of the candidate frame and the real frame to eliminate the error caused by the candidate frame. Here, the average IoU is used to analyze the clustering results, and the average IoU objective function g of clustering can be expressed as:

$$g = \arg\max \frac{\sum_{k}^{x=1} \sum_{y=1}^{n_k} X_{\mathrm{XoU}}(O, C)}{n}, \tag{4}$$

where O represents the sample. n represents the total number of samples. C represents the cluster center. k represents the number of clusters and the number of samples in the kth cluster. x represents the serial number of the sample. y represents the serial number of the cluster center. $X_{\mathrm{XoU}}(O, C)$ indicates the intersection and union ratio of the area of the bounding box and the cluster center box.

The clustering anchor box can make the network converge faster and ensure the detection accuracy of the network. The specific feature map and its prior box size allocation are shown in Table 1.

3.3. YOLOv4 Model Backbone Network Improvement. The backbone network of YOLOv4 model adopts CSPDarknet53, and Spatial Pyramid Pooling (SPP) and PANet module are used for feature fusion, which can effectively extract the feature information of video images. However, the accuracy and speed of model detection are not ideal in that the network has many parameters. In order to meet the speed requirement of face detection, MobileNet lightweight network model is introduced to replace the backbone network of YOLOv4 model. The backbone network of the improved YOLOv4 model adopts the deep separable convolution layer (Figure 2(a)), which divides the traditional standard convolution layer (Figure 2(a)) into two convolution modes: deep convolution and point convolution. As shown in Table 2, Features are extracted by depth convolution and point convolution, and feature maps of three sizes are output to detect faces of different sizes.

Table 1: Characteristic graph and its prior box size allocation.

Feature map	Feature graph size	Anchor box size
Predict one	13×13	(44, 56) (76, 100) (178, 236)
Predict two	26×26	(15, 19) (21, 27) (30, 38)
Predict three	52×52	(5, 6) (8, 9) (10, 13)

Assume that the convolution kernel size is $S \times S$ and the number is R. The input characteristic map size is $H_R \times H_R$, and the number of characteristic map channels for input and output is P, Q respectively. And the input and output feature maps have the same size. Then, the calculation amount after traditional standard convolution is

$$U_1 = S \times S \times H_R \times H_R \times P \times Q. \tag{5}$$

When the depth separable convolution layer is used, the calculated amount of point convolution is

$$U_2 = H_R \times H_R \times 1 \times 1 \times P \times Q. \tag{6}$$

The calculation amount of depth convolution is

$$U_3 = S \times S \times H_R \times H_R \times P. \tag{7}$$

Therefore, the calculation amount of depth separable convolution is

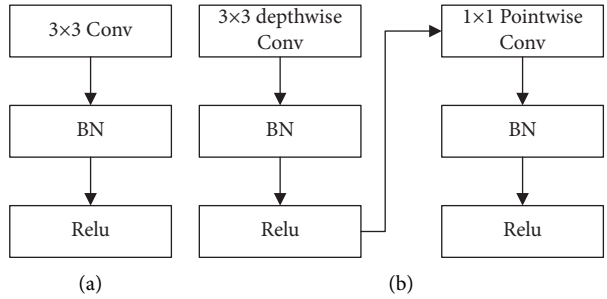

Figure 2: Traditional standard convolution and depth separable convolution. (a) Traditional standard convolution. (b) Depth separable convolution.

Table 2: Improved YOLOv4 model backbone network structure.

Layer name	Filter size	Parameter	Output size
Input_1		0	(416,416,3)
Convl	$3\times3\times3\times32$	864	(208,208,32)
Conv_dw_1	$3\times3\times32$dw	288	(208,208,32)
Conv_pw_1	$1\times1\times32\times64$	2048	(208,208,64)
Conv_dw_2	$3\times3\times64$dw	576	(104,104,64)
Conv_pw_2	$1\times1\times64\times128$	8192	(104,104,64×2)
Conv_dw_3	$3\times3\times128$dw	1152	(104,104,64×2)
Conv_pw_3	$1\times1\times128\times128$	1 6384	(104,104,64×2)
Conv_dw_4	$3\times3\times128$dw	1152	(52,52,128)
Conv_pw_4	$1\times1\times128\times256$	32768	(52,52,256)
Conv_dw_5	$3\times3\times256$dw	2304	(52,52,256)
Conv_pw_5	$1\times1\times256\times256$	65536	(52,52,256)
Conv dw_6	$3\times3\times256$dw	2304	(26,26,256)
Conv_pw_6	$1\times1\times256\times512$	131072	(26,26,512)
(Conv_dw, conv_pw)×5	$3\times3\times512$dw $1\times1\times512\times512$	262144	(26,26,512)
Conv_dw_12	$3\times3\times512$dw	4608	(13,13,512)
Conv_pw_12	$1\times1\times512\times1024$	524288	(13,13,1024)
Convdw_13	$3\times3\times1024$dw	9216	(13,13,1024)
Conv_pw_13	$1\times1\times1024\times1024$	1048576	(13,13,1024)

$$U_4 = S\times S\times H_R\times H_R\times P + H_R\times H_R\times 1\times 1\times P\times Q. \quad (8)$$

The ratio of depth separable convolution computation to traditional standard convolution computation is as follows:

$$\frac{U_4}{U_1} = \frac{1}{Q} + \frac{1}{S^2}. \quad (9)$$

When the size of the input image is 416×416, the output $13\times13\times1024$ feature map has a larger receptive field, which can be used to detect larger faces. The output feature map of $26\times26\times512$ size has moderate receptive field and can be used to detect medium-sized faces. The output $52\times52\times256$ feature map has a small receptive field and can be used to detect smaller faces. Without losing the accuracy, the convolution calculation in this paper has less computation and fewer parameters. And the detection speed is improved.

3.4. CBAM of Dual Attention Mechanism. CBAM model combines the channel and space information of feature map, which makes the model more perceptive of target information and suppresses the interference caused by invalid information without changing the size of feature map. In order to extract richer high-level semantic features of the target to be detected, the CBAM model with dual attention mechanism is embedded, as shown in Figure 3. First, the channel number of pooling as the channel number of input feature map is set, and two spatial information of average pooling and maximum pooling of feature map are calculated. Multilayer Perceptron (MLP) is used to compress the spatial dimension of average pooling and maximum pooling spatial information, and two feature vectors are obtained and summed to further improve the spatial feature representation of the target to be detected. Then, the channel attention weight coefficient WC is obtained by activation function $\sigma(.)$, and the new feature graph G_D' is obtained by multiplying the original input feature graph G_D. The average pooling and maximum pooling information of the original input feature graph G_D were calculated and spliced. Then, the spliced feature vectors are input into the convolution layer, and the spatial attention weight coefficient W_s is obtained through activation function $\sigma(.)$, so that network feature learning is more focused on face features. Finally, W_s is multiplied by the new feature graph $G_D'$to obtain the final feature graph G_o.

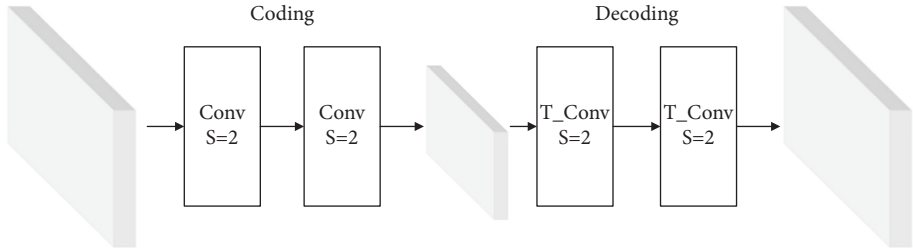

Figure 3: Semantic segmentation module SSA.

$$\begin{aligned} W_c(G_D) &= \sigma\left(g\left(P_{\text{mean}}(G_D)\right) + g\left(P_{\max}(G_D)\right)\right), \\ G'_D &= W_c(G_D) \otimes G_D, \\ W_s(G'_D) &= \sigma\left(f^{d\times d}\left(P_{\text{mean}}(G'_D); P_{\max}(G'_D)\right)\right), \\ G_o &= W_s(G'_D) \otimes G'_D, \end{aligned} \tag{10}$$

where $g(\cdot)$ is the function form of MLP network model? $P_{\text{mean}}(\cdot)$ is the average pooling function. $P_{\max}(\cdot)$ is the maximum pooling function. $f^{d\times d}$ is a convolution layer operation of size $d \times d$. As shown in Figure 4.

3.5. Spatial Attention Mechanism based on Weak Semantic Segmentation

3.5.1. Weak Semantic Segmentation Algorithm. Research on the output of the neural network shows that there is mapping overlap between the active area of the feature map and the detection area in the original map, indicating that the neural network will gradually converge the attention range and focus the visual threshold on the real detection area in the process of data flow. However, in the detection task of this paper, some faces account for a small pixel area of the image, and the redundant part of the image is not a pure background. These noises distort the spatial mapping of the detected objects characterized by deep features. If the learned features are directly input into the detection head, the identification and detection of targets will be seriously affected by the network. Some researchers have proposed a solution: adopting an unsupervised way. By stacking the residual attention modules, the effective part in the feature map is enhanced. But the disadvantage is also obvious: the overall cost increases.

A weak semantic segmentation module is proposed in this paper to design a lightweight detection model. Abandoning the idea of building the network basic module, it is used as a dynamic enhancement attachment only once before the detection head of YOLOv4 lightweight network, and the output of this module is used as the spatial weight distribution learned by the model to correct the features, thus improving the detection ability of the model.

In the spatial attention algorithm based on weak semantic segmentation, the output of the input image through MobileNet v3-Lite and feature fusion submodule is used as the input feature map of the weak semantic segmentation module. The input feature map is sent to the semantic segmentation module to predict the foreground and background, to obtain the spatial attention weight. Weights are assigned to the corresponding input feature maps to strengthen the target features.

3.5.2. Semantic Segmentation Modules. Semantic segmentation module is a typical coding-decoding model. The coding part generally uses convolution or pooling to reduce the size of feature map, while the decoding part uses bilinear interpolation to restore feature map step by step. This paper designs two semantic segmentation modules: basic semantic segmentation module SSA and enhanced semantic segmentation module SSB, as shown in Figures 3 and 5.

SSA module uses the simplest framework in semantic segmentation task. Only two convolutions with step 2 are used as encoders and two transposed convolutions with step 2 are used as decoders to verify whether semantic segmentation is effective. In order to expand the receptive field of modules and combine the multiscale context, the Atrous Spatial Pyramid Pooling (ASPP) of DeepLabv3 was improved and introduced into SSA to form SSB.

ASPP module adopts the form of multi-branch parallel connection and obtains larger receptive field through the cavity convolution (DC, dilated convolution) with different expansion rates, and then fuses the features of each branch to obtain accurate context information. In this paper, the ASPP module is improved: in order to reduce the computation of the module, 1×1 standard convolution is used to reduce the dimension of each branch. The expansion rate of DC layer is reduced to 1, 3 and 5. 3×3 standard convolution is added to branches with expansion rates of 3 and 5 to obtain basic features. The features of 4 branches are spliced. Finally, the feature fusion and channel dimension reduction are carried out by 1×1 standard convolution.

3.5.3. Supervise the Generation of Information and Attention Weight. Since the data set provided by this task lacks semantic labels, this paper can only use annotation information to generate semantic masks, which are called weak semantic masks to distinguish them from traditional methods. Specific approach: if the pixel point on the output of the semantic segmentation module falls into the coordinate box, the value of the pixel at this point is set to 1; otherwise, it is set to 0. To avoid ambiguity, the pixel value of the points on the annotation box was set to 255, which was ignored during training. This pixel numeric map is then mapped to a size matching the detection head of the YOLOv4 lightweight network. In this paper, binary values

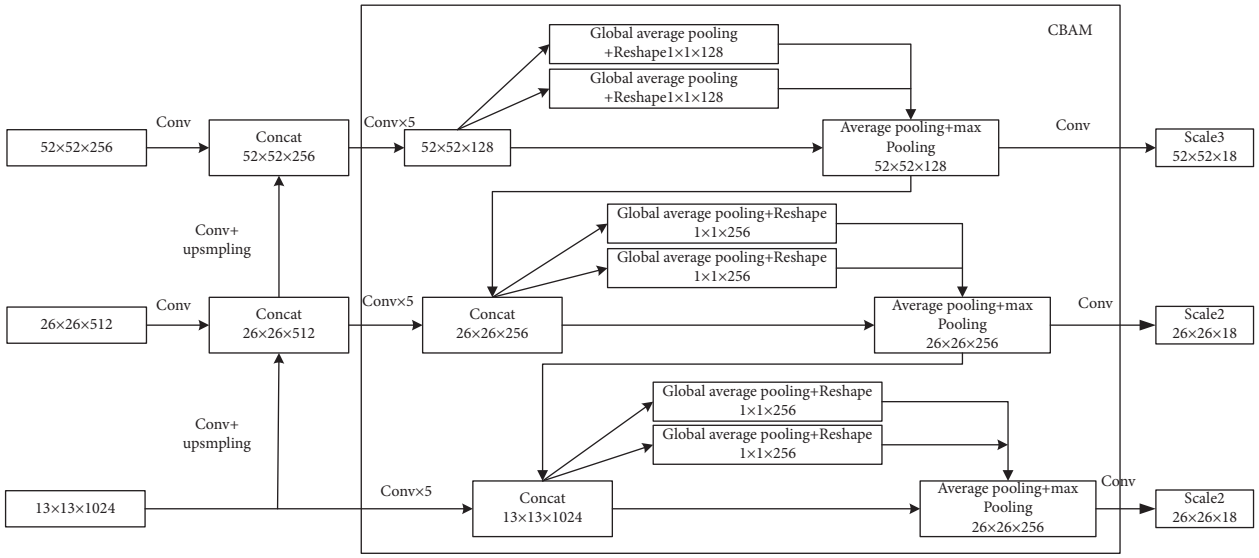

Figure 4: CBAM model structure with dual attention mechanism.

are selected as the distinction to obtain the final weak semantic mask.

The weak semantic mask is used for supervision training, and the output B_x of the module SSA or SSB is the spatial attention weight. In order to prevent the degradation of network performance, residual structure is introduced as identity mapping, that is,

$$F_3 = (1 + B_x)F_2. \tag{11}$$

Since this paper only needs this module to predict the foreground and background, it is a binary classification task. Therefore, only the cross-entropy loss needs to be calculated for the segmentation result B and the weak semantic mask B^*, that is,

$$L_{\text{mask}} = -\sum_{x}^{h}\sum_{y}^{w} B^*_{xy}\log\left(B_{xy}\right) + \alpha\left(1 - B^*_{xy}\right)\log\left(1 - B_{xy}\right), \tag{12}$$

where h and w are the length and width of the weak semantic mask. B_{xy} is the pixel value of the point with coordinates (x, y) on the feature map.

3.6. Loss Function. The loss function of the improved YOLOv4 model, which combines lightweight network and dual attention mechanism, consists of three parts: loss of position of bounding box L_{ClOU}, loss of confidence L_{conf} and loss of classification L_{cls}.

$$L = L_{\text{CIOU}} + L_{\text{conf}} + L_{\text{cls}},$$

$$L_{\text{CIOU}} = 1 - I(C, D) + \frac{\rho^2\left(C_{\text{ctr}}, D_{\text{ctr}}\right)}{m^2} + \alpha v,$$

$$\alpha = \frac{v}{[1 - I(C, D)] + v},$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2,$$

$$L_{\text{conf}} = \sum_{x=0}^{V^2}\sum_{y=0}^{G} T^{\text{oby}}_{xy}\left[E^y_x\ln\left(E^y_x\right) + \left(1 - E^y_x\right)\ln\left(1 - E^y_x\right)\right] + \lambda_{\text{nooby}}\sum_{x=0}^{V^2}\sum_{y=0}^{G} T^{\text{nooby}}_{xy}\left[\overline{E^{y^-}_x}\ln\left(E^y_x\right) + \left(1 - E^y_x\right)\ln\left(1 - E^y_x\right)\right],$$

$$L_{\text{cls}} = \sum_{x=0}^{V^2} T^{\text{oby}}_{xy}, \sum_{c\in k}\{\overline{p}^y_x(c)\ln p^y_x(c) + [1 - \overline{p}^y_x(c)]\ln[1 - p^y_x(c)]\}. \tag{13}$$

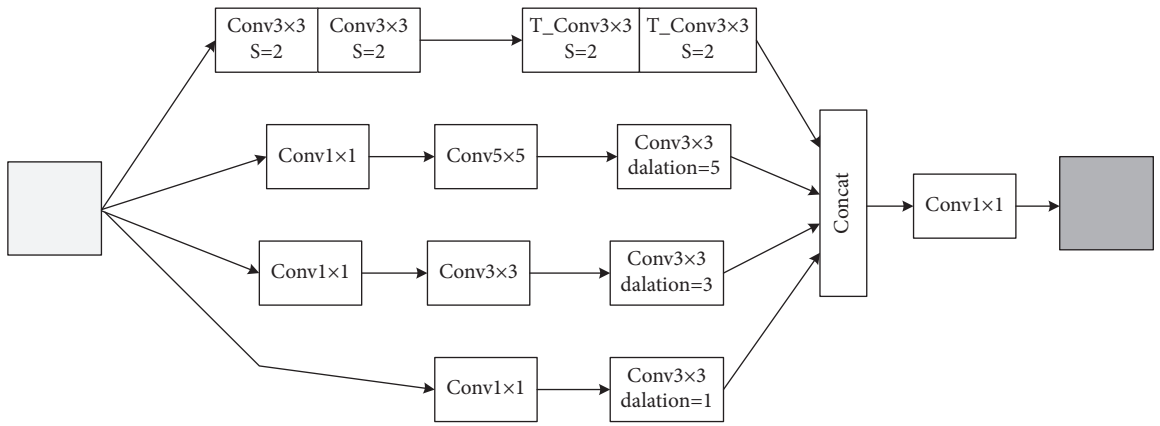

Figure 5: Semantic segmentation module SSB.

where $I(C, D)$ is the intersection and union ratio of prediction box C and real box D. $\rho^2(C_{\text{ctr}}, D_{\text{ctr}})$ is the Euclidean distance between the predicted box center point D_{ctr} and the real box center point C_{ctr}. α is a weight function. v is the aspect ratio similarity measurement coefficient. w^{gt} and h^{gt} are the width and height of the real frame respectively. w and h are the width and height of the prediction frame, respectively. V^2 is the number of grids. G is the number of prior frames in each grid. T_{xy}^{oby} is the target contained in the y-th prediction frame generated on the x-th grid, $x \in [0, V^2]$, $y \in [0, G]$。 T_{xy}^{nooby} is the y-th prediction box generated on the x-th grid that does not contain the target. $\overline{E_x^y}$ is the true confidence. E_x^y is the prediction confidence. λ_{nooby} is a self-set calculation coefficient. k is the target classification number. $\overline{p}_x^y(c)$ is the true probability that the object in the frame belongs to a certain category. $p_x^y(c)$ is the prediction probability that the object in the frame belongs to a certain category c.

4. Experimental Data Analysis

The experiment uses a benchmark data set WIDERFACE for face detection. The face in this dataset picture has great changes in scale, posture, and occlusion. Through random sampling from 61 scene categories, the data set is divided into training set, verification set, and test set according to the ratio of 4 : 1 : 5. The experimental environment configuration of this paper is as follows Table 3:

4.1. Model Training and Testing

4.1.1. Training Model. The training parameters of the network have been set before the model training. The target category is 1, and the anchor box is (5, 6), (8, 9), (10, 13), (15, 19), (21, 27), (30, 38), (44, 56), (76, 100), (178, 236). The training batch size is 64. The momentum is 0.9. The attenuation rate is 0.0005. The maximum number of iterations is 60000. At the beginning of training, set the learning rate to 0.001 to stabilize the whole network. After 10000 iterations, it is adjusted to 0.01. After 30000 iterations, it is adjusted to 0.001. After 40000 iterations, it is adjusted to 0.0001, which makes the loss function converge further.

Table 3: Configuration of experimental environment.

Experimental environment	Environment configuration
The operating system	Linux 64
GPU	TITAN Xp
CPU	Intel(R)Core i7-3770 CPU@
Memory	12G
Deep learning framework	Darknet + TensorFlow

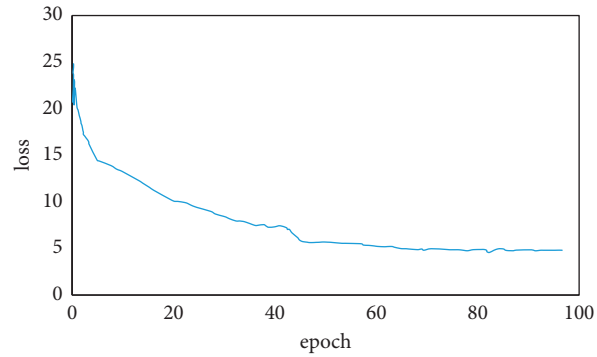

Figure 6: Graph of average loss function.

The loss function curve of the improved YOLOv4 model is shown in Figure 6. During the iteration process, the loss values of the training set show an obvious downward trend, and the convergence process is stable after 70 iterations, showing that the improved YOLOv4 model, which combines lightweight network and dual attention mechanism, has not experienced the fitting phenomenon.

4.1.2. Test Result. Figure 7 shows the results of face detection in complex conventional scenes. Figure 7(a) can detect faces in dim light. Even in high-density population, many small faces can be well detected in Figure 7(b). Figure 7(c) itself has low resolution and poor visual effect, but it can be seen from the detection results that there are still many faces detected. It can be seen from Figure 7(d) that this method also has a good detection effect on shielding human face. The proposed model can effectively perform face detection in complex scenes.

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

Figure 7: Detection results on the WIDERFACE set. (a) Light change. (b) Dense small faces. (c) Fuzzy scene. (d) Blocking faces.

4.2. Analysis of Experimental Results. In this paper, the mean average precision, mAP is used to evaluate the performance of the network model for face detection, and the calculation formula is

$$\text{mAP} = \frac{\sum_{x=0}^{n} \text{AP}(x)}{n}. \tag{14}$$

AP represents the average of the accuracy rate under different recall rates. mAP represents the average of the detection accuracy of all categories. n is the number of sample categories in the dataset.

The algorithm is compared with other existing ones to illustrate its superiority. The comparison results are shown in Table 4. Existing methods are RetinaNet50, YOLOv3, cascade-CNN, Adaptive Training Sample Selection (ATSS), Probabilistic Anchor Assignment (PAA), YOLOv4, and YOLOv4-Tiny. These comparison methods include the classic frameworks and the latest improvements that have emerged in recent years. The mean Average Precision (mAP) of the proposed algorithms was 96.9%, 94.3%, and 81.7% respectively. Compared with RetinaNet50 algorithm, it is improved by 11.7%, 13.2%, and 15.8% respectively. Compared with YOLOv3 algorithm, it is improved by 3.5%, 5.5%, and 6.2% respectively. And compared with CascadeR-CNN algorithm, it is improved by 2%, 4.4%, and 8%, respectively. Compared with ATSS algorithm, it is improved by 4%, 7.1%, and 7.6%, respectively. Compared with PAA algorithm, it is improved by 3.7%, 6%, and 6.1%, respectively. Compared with YOLOv4 algorithm, it is improved by 0.2%, 0.8%, and 4.1%, respectively. Compared with YOLOv4-Tiny algorithm, it is improved by 7.9%, 10.6%, and 10.7%, respectively.

Table 4: Comparative results (mAP%).

Algorithm	mAP/%		
	Easy	Medium	Hard
RetinaNet50	85.2	81.1	65.9
Cascade R-CNN	94.9	89.9	73.7
YOLOv3	93.4	88.8	75.5
ATSS	92.9	87.2	74.1
PAA	93.2	88.3	75.6
YOLOv4	96.7	93.5	77.6
YOLOv4-tiny	89	83.7	71.0
Proposed	96.9	94.3	81.7

In order to compare the advantages of the algorithms in this paper more comprehensively and comprehensively, two indexes of model size and detection speed are added to compare these algorithms based on accuracy, as shown in Table 5.

It is found that the detection speed of this model is second only to YOLOv4-Tiny, but the map of this model is much higher than YOLOv4-Tiny, and the size of this model is much smaller than YOLOv4-Tiny.

In order to verify the effectiveness of the algorithm in this paper, the algorithm is evaluated on the FDDB dataset, which consists of 2845 images with a total of 5171 faces. This is a large-scale face detection dataset using normalized

Table 5: Comprehensive comparison between this method and other mainstream methods.

Method	mAP/%	FPS	Model Size/MB
RetinaNet50	85.2	3.1	296.8
Cascade R-CNN	94.9	1.6	553.6
YOLOv3	93.4	3.3	255.8
ATSS	92.9	3.3	257.1
PAA	93.2	2.1	253.6
YOLOv4	96.7	3.6	242.9
YOLOv4-tiny	89	16.2	23.7
Proposed	96.9	12.0	8.4

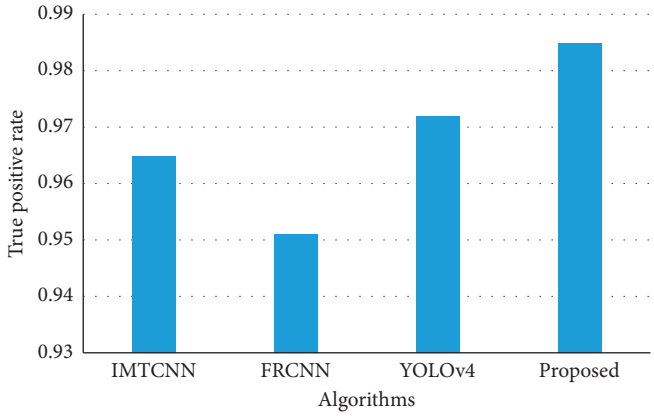

Figure 8: Performance comparison on FDDB dataset.

operation steps. FDDB uses elliptical face boundaries, and the operation defines discontinuous and continuous values. In the discontinuous value operation, the number of detected faces is calculated relative to the number of false positives. The detection bounding box (or ellipse) can be recognized as a real face only if it contains an IOU ratio greater than 0.5. In the continuous value operation, the extent to which a face is localized can be calculated by treating the IOU ratio as a matching metric of the detection bounding box.

When the false certificate reaches 1000, the performance comparison on FDDB dataset is given in Figure 8. It can be intuitively seen from Figure 8 that the algorithm proposed in this paper outperforms other algorithms in face detection, and the true positive rate in this paper can reach 0.985.

5. Conclusion

This paper proposes a face detection method based on lightweight network and weak semantic segmentation attention mechanism aiming at the detection problem caused by the changeable face scale in practical application. The main idea of this paper is to replace the backbone network of YOLOv4 model with Mobile Net lightweight network model, to improve the speed of model detection without losing accuracy. K-means ++ clustering algorithm was used to recluster prior frames, which improved the detection accuracy of the model. Then, the CBAM model with channel and space dual attention mechanism is embedded, and the weak semantic segmentation module is designed. Under the premise of not affecting the detection speed, the perceptual ability of different scales of faces is improved, and the missed and false detection situation is reduced. Compared with other face detection algorithms in the same environment, the results confirm the feasibility and superiority of the proposed method via using the same data set. In the following work, the detection network should be further optimized. In addition, the generalization of the model for complex illumination can be improved by adding face images under complex illumination into the dataset.

Data Availability

The labeled datasets used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares no competing interests.

Acknowledgments

This study was sponsored by Sichuan University of Arts and Science.

References

[1] Z. Liu, X. Qi, and P. H. S. Torr, "Global Texture Enhancement for Fake Face Detection in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8060–8069, IEEE, Seattle, WA, USA, June, 2020.

[2] S. Chen, Y. Yao, Y. Zhang, and G. Fan, "CRISPR system: d," *Cellular Signalling*, vol. 70, Article ID 109577, 2020.

[3] P. Voigtlaender, J. Luiten, and P. H. S. Torr, "Siam r-cnn: visual tracking by re-detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6578–6588, IEEE, Seattle, WA, USA, June, 2020.

[4] Z. Y. Chen, I. Y. Liao, and I. Y. Liao, "Improved fast R-CNN with fusion of optical and 3D data for robust palm tree detection in high resolution UAV images," *International Journal of Machine Learning and Computing*, vol. 10, no. 1, pp. 122–127, 2020.

[5] S. Wan and S. Goudos, "Faster R-CNN for multi-class fruit detection using a robotic vision system," *Computer Networks*, vol. 168, Article ID 107036, 2020.

[6] L. Li, Z. Yang, L. Jiao, and F. X. Liu, "High-resolution SAR change detection based on ROI and SPP net," *IEEE Access*, vol. 7, Article ID 177009, 2019.

[7] J. Wang, J. Luo, B. Liu, and R. L. H. Feng, "Automated diabetic retinopathy grading and lesion detection based on the modified R-FCN object-detection algorithm," *IET Computer Vision*, vol. 14, no. 1, pp. 1–8, 2020.

[8] M. Wu, H. Yue, J. Wang, and Y. M. Y. C. C. Huang, "Object detection based on RGC mask RCNN," *IET Image Processing*, vol. 14, no. 8, pp. 1502–1508, 2020.

[9] S. Zhai, D. Shang, S. Wang, and S. Dong, "DF-SSD: an improved SSD object detection algorithm based on dense-net and feature fusion," *IEEE Access*, vol. 8, Article ID 24344, 2020.

[10] W. Chen, H. Huang, S. Peng, and C. C. Zhou, "YOLO-face: a real-time face detector," *The Visual Computer*, vol. 37, no. 4, pp. 805–813, 2021.

[11] Y. N. Wang and X. L. Wang, "Remote sensing image target detection model based on attention and feature fusion," *Laser & Optoelectronics Progress*, vol. 58, no. 2, pp. 355–363, 2021.
[12] R. Dasgupta, Y. S. Chowdhury, and S. Nanda, "Performance Comparison of Benchmark Activation Function ReLU, Swish and Mish for Facial Mask Detection Using Convolutional Neural Network," in *Intelligent Systems*, pp. 355–367, Springer, Singapore, 2021.
[13] Y. Zhang, X. Zhang, and W. Zhu, "ANC: attention network for COVID-19 explainable diagnosis based on convolutional block Attention module," *Computer Modeling in Engineering and Sciences*, vol. 127, no. 3, pp. 1037–1058, 2021.
[14] N. Van Quang and H. Fujihara, "Revisiting a Single-Stage Method for Face detection," in *Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–8, IEEE, Lille, France, August, 2019.
[15] H. Ku and W. Dong, "Face recognition based on MTCNN and convolutional neural network," *Frontiers in Signal Processing*, vol. 4, no. 1, pp. 37–42, 2020.
[16] S. Fang, Y. Li, and X. Liu, "An improved multi-scale face detection algorithm based on SSD model," *Inf. Technol. Inf*, vol. 2, pp. 39–42, 2019.
[17] Z. Zhang, J. Wang, J. Wang, and Z. Jing, "Small-size face detection with multi-scale and texture feature enhancement," *Computer Application Research*, vol. 38, no. 3, pp. 914–918, 2021.
[18] A. Bochkovskiy, C. Y. Wang, and H. Liao, "YOLOv4: optimal speed and accuracy of object detection," 2020, https://arxiv.org/abs/2004.10934.