Hindawi
Advances in Multimedia
Volume 2022, Article ID 3730529, 11 pages
https://doi.org/10.1155/2022/3730529

Research Article

Research on Face Local Attribute Detection Method Based on Improved SSD Network Structure

Qun Luo[1] **and Zhendong Liu**[2]

[1]*Information Engineering Department, Chongqing City Vocational College, Chongqing 402160, China*
[2]*Big Data Department, Chongqing City Vocational College, Chongqing 402160, China*

Correspondence should be addressed to Zhendong Liu; lzd033353@cqcvc.edu.cn

Received 10 November 2021; Revised 12 December 2021; Accepted 27 December 2021; Published 28 January 2022

Academic Editor: Qiangyi Li

The existing face detection methods usually had the problem of low accuracy of face recognition in the environment of occlusion interference, which was limited when applied to the face detection task in complex scenes. Therefore, in order to realize high precision and real-time local face recognition in a complex environment, a face local attribute detection method based on improved SSD network structure was proposed. Based on the analysis of the face local attribute detection task, SSD was used as the basic detection network structure, and the VGG16 feature extraction model was used as the framework of face local detection. On this basis, by organically connecting different layers of the SSD network and integrating convolution block attention module, the improved SSD network structure was used to realize face local attribute detection. The proposed model was trained and tested using typical public datasets such as Wider Face, MAFA, and COFW. Experimental results showed that this method had high recognition accuracy, can better detect local features of the human face than other models, and can provide some support for local face attribute detection. This method would provide a theoretical basis and technical support for local face attribute detection in complex scenes.

1. Introduction

With the rapid development of the smart city and artificial intelligence technology, face detection and recognition play an important role in the field of digital intelligence [1, 2]. How to analyze various face features accurately and quickly, effectively detect face attributes, and extract useful information has important research value for in-depth research on artificial intelligence. In recent years, research on face detection and recognition has been developed with the in-depth application of face tracking technology. Although the face tracking method is very similar to the face recognition algorithm to a certain extent, there are often differences due to different application emphases. The research on face tracking mainly belongs to a direction in the field of computer vision, which is mainly applied to face monitoring and face recognition. The difficulty mainly focuses on the extraction of face features. Deep learning has achieved great success in solving some popular machine vision problems. The research on face detection is to extract and process the features related to the face from the collected images and analyze the recognized face features. Because there are some differences between individuals, the local features of the human face are also different [3]. In addition, due to different factors such as external ambient light, occlusion position, face shape of different individuals, and collection angle of face features, there are some differences in the extracted face features, which makes the face detection process very complex. With the continuous application of deep convolution neural network in image and feature processing, how to apply deep convolution network to face image detection and recognition has attracted extensive attention.

With the wide application of convolution neural network and transfer learning in the field of artificial intelligence, face detection and recognition methods have also

been deeply studied. People often study face detection and recognition methods based on deep learning, and the constructed model is becoming more and more perfect [4, 5]. Although the model has been improved to a certain extent, the model includes many parameters and a large amount of calculation workload. For example, the trained model is generally difficult to adapt to FPGA mobile devices and cannot meet the requirements of real-time face monitoring due to the lack of necessary resources. In the process of practical application, due to the unpredictability of the external environment, various factors that may interfere with the face cannot complete the specific detection task, which makes the face detection difficult to complete smoothly. In order to achieve a certain recognition accuracy, face detection often requires not only a long time overhead but also some software and hardware support [6]. Therefore, the key to face detection in a natural environment is how to improve the speed and accuracy of real-time detection, which is also a common concern in the field of face detection. Therefore, based on the widely used SSD network structure, this article uses VGG16 to extract face features, uses multilevel feature enhancement, and introduces a convolution block attention block to detect face local attributes.

2. Related Works

Face feature detection is an operation process of extracting relevant features based on face attributes. In recent years, face detection and recognition methods have been widely used in face attribute analysis, face behavior monitoring, face image monitoring, and other tasks, and many research results have been obtained [7]. Face detection has become a very important research branch, which is independent of the target monitoring task. Due to the limitations of existing hardware conditions, imperfect feature recognition algorithms, insufficient datasets for training, and other factors, the traditional face detection methods usually have accuracy and detection rate, which are difficult to meet the needs of practical application. In recent years, people have continuously improved the description of face features, and some of the proposed models have been successfully applied to face detection, which effectively promotes the in-depth application of face detection and recognition methods in different fields. Because deep learning technology is widely used in face detection and recognition, the accuracy and speed of face detection have been improved accordingly. For example, the cascade convolutional neural network (C-CNN) proposed by Li et al. has achieved an average accuracy of 98% on annotated faces in the wild (AFW) [8].

From the progress of artificial intelligence technology and its application research, considering that there are some differences in face attributes and feature extraction methods, the algorithms of face detection and recognition mainly include the methods based on traditional artificial feature construction and convolution neural network and deep learning. Face recognition algorithms based on traditional artificial features generally include a cross-layer algorithm based on AdaBoost and a feature matching algorithm based on DPM [9, 10]. With the in-depth study of deep learning and convolutional neural network, artificial intelligence mainly uses deep neural network technology to realize its application in related fields. At present, convolutional neural networks and deep learning methods can be widely used in many fields, such as natural language processing, intelligent biological recognition, and intelligent driving. Because deep learning and other related methods have strong learning and training ability, they can effectively obtain face features and achieve satisfactory results even in complex environments. The research shows that the method of statistical learning and training can better avoid the interference of artificial feature extraction, and the training dataset is used for feature extraction, and the effective features extracted can better realize face recognition and detection. At present, the common methods include face recognition algorithm based on principal component analysis (PCA), cross-layer algorithm based on AdaBoost, face detection method based on support vector machine (SVM), and convolution neural network. There are many methods used in the field of artificial intelligence, such as face cross-layer classification algorithm based on AdaBoost, which is one of the face classification algorithms based on deep learning [11]. This method adopts the cascade classification method. According to the characteristics of the face, multiple weak classifiers are cascaded to form a strong classifier to detect the face.

FaceNet, an early representative face detection algorithm, mainly extracts effective face features from complex face attributes [12]. The algorithm uses the face attribute information to obtain complex face features, analyzes the face attribute information from the attribute perception layer, and constructs the detection region by correlating the attribute information and the face features. This method realizes the classification of face attribute information and the detection of candidate regions through the training of multitask CNN; that is, the detection region is divided from the spatial structure and the information is classified in order to reduce the interference of the nonface attribute information. Because the extraction and division of face attribute information can improve the robustness of the network model to complex face detection and recognition, when face attribute information is in a complex environment, accurate recognition and detection can still be obtained through model detection. Wang et al. proposed a face-based facial attention mechanism network (FAN) [12, 13], which not only adopts the integrated single shot retina and anchor frame attention model but also adopts the enhancement method of facial features to detect face attribute information. When constructing the anchor level attention model, this method mainly carries out the relevant hybrid operation of the existing feature map and anchor map to improve the resolution of face attribute features and avoid the interference of various external factors [13]. Attention mechanism has been well applied in the fields of image generation, semantic recognition, and so on. The network model based on facial attention mechanism has an obvious effect on complex face detection.

Because single-scale feature mapping is not good at representing face size and shape, using a convolutional

neural network to extract relevant information from different layers can naturally alleviate this contradiction. When the MS-CNN method is used for face detection, the candidate regions for face detection are mainly obtained according to different feature maps, and the inverse function of the feature map is used to replace the upsampling of the original image [14]. This method can significantly speed up the detection speed and provide recognition accuracy [15]. Based on the SSD network structure model, different detection areas are obtained using the function of convolution structure layer and receptive field on the VGG16 platform to obtain a better detection effect. For example, the features of low-level conv4_3 using the SSD network structure model are small and can detect small targets, while the features of high-level conv4_3 are suitable for detecting large targets. Different levels of pyramid structure are used to realize the effective detection of different targets. Because the bottom layer of the model has too small requirements for detection targets, the shallow layer of the network model can hardly obtain smaller target features, which limits the application of SSD based network structure model in small target recognition [16, 17].

Compared with the Fast R-CNN recognition model, the detection model based on the HyperNet network structure is more suitable for small target detection and feature extraction. This is because different layers of the HyperNet network structure model have synergistic effects in processing small targets, and the features extracted by different layers are related to each other. The feature pyramid network (FPN) structure model is used to process the features in turn using different layers from top to bottom. It uses the upper layer to increase the semantic information strength and supplement the semantic information through the connection layer to obtain the required semantic information map. The detection network model based on multilayer feature fusion is mainly adopted to enhance the relevance of context information to improve the speed and accuracy of face detection. Therefore, the multilayer feature fusion method determines that the extraction of top-level features is the key to the top-down network structure model to achieve the effect of face detection.

3. Basic Network for Local Face Attribute Detection

SSD (single shot detector) network model can combine different application requirements for parametric design and structure modification. Therefore, SSD is an object detection network structure commonly used in the field of face detection and recognition [18, 19]. SSD adopts a fast and effective single-stage target detector across multiple feature layers of different scales. This article adopts SSD basic network structure and applies it to face detection. As shown in Figure 1, it is the schematic diagram of the network structure adopted in this article. Compared with the existing R-CNN algorithm, the model can extract face features quickly, does not need too many parameters, and the computational workload is relatively small. Therefore, SSD basic network model has become one of the common structures to realize target detection.

When building the SSD network structure model, using the image processing relationship between the upper and lower layers, the feature maps obtained from different convolution layers are divided; that is, the large feature map and small feature map are processed by different convolution layers in turn, and then each target is detected through different anchor frames. When an arbitrary face image is input, after the basic parameter setting, the SSD network structure model uses the VGG16 network to extract the features of the image, and the associated features of different sizes will be obtained through proportional transformation. The specific proportion transformation formula is shown in the following equations:

$$T_m = f_m(T_{m-1}), \tag{1}$$

$$f_m(T_{m-1}) = f_m(f_{m-1}(\cdots f_1(1))), \tag{2}$$

$$S = T(t_m(T_m), t_{m-1}(T_{m-1}), \ldots, t_{m-n}(T_{m-n})), m > n > 0, \tag{3}$$

where T_m represents the feature map of the m-th layer, f_m denotes the nonlinear operation between two adjacent feature layers, and $f_1(1)$ indicates the operation of obtaining the feature map of the first layer from the input image. t is the detection result of the corresponding scale range of the characteristic map, and the result is represented by S. According to formulas (1) to (3), the characteristic layer of layer m is determined by layer $m-1$. Figure 2 coverages anchor frames of different sizes on the feature map.

4. Detection Method for Local Face Attributes

4.1. Improved Detection Network. Because the layers included in the traditional SSD network structure model are independent of each other during feature detection, the internal operations of each layer do not interfere with each other. By integrating the features containing different semantic information and resolution through cross-layer connection, we can learn and obtain richer feature information and improve the detection performance. Therefore, based on the initial SSD network structure, VGG16 is added to the SSD network model as a feature extraction network, and the full connection layer is replaced. The convolution layer is used to replace the low-level part, and four convolution layers such as Conv8, Conv9, Conv10, and Conv11 are added at the same time. On this basis, the above added feature extraction layer is used to classify the face feature images and predict the relevant positions of different features. Each feature layer of the SSD network model is independent of each other when classifying and predicting features. Usually, if the number of convolution layers is too small, the real face attribute features cannot be reflected only through the underlying features, and the recall rate of features will be too low. Therefore, based on the original SSD network model, this article integrates the relevant

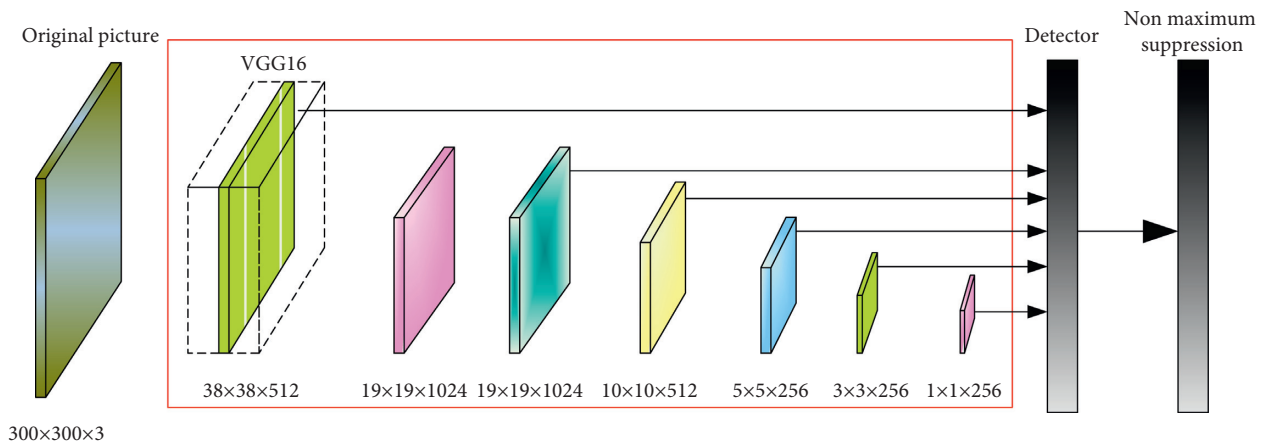

Figure 1: Working diagram of basic SSD network structure.

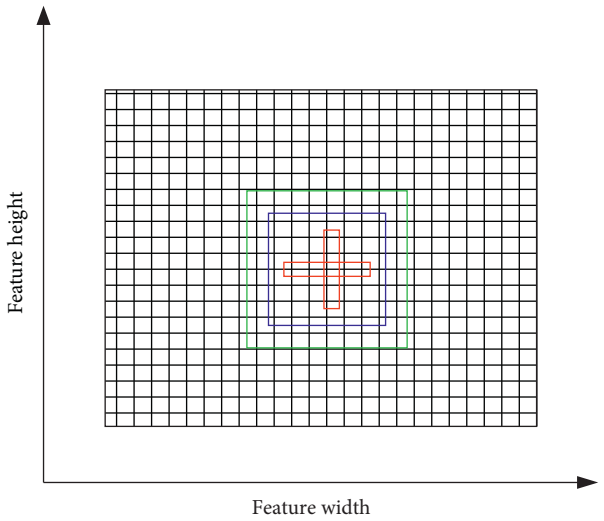

Figure 2: Dimension diagram of feature diagram in the default box.

characteristics of the upper and lower layers by organically connecting each layer. At this time, when using the convolution kernel in SSD to complete the convolution operation, it is necessary to find all attribute features included in the face image; especially, the detection area containing the object should be compared pixel by pixel, while other areas excluding the image do not need to be scanned. Therefore, this article intends to use the channel attention module function to detect the effective area of the network model so as to avoid repeated detection or invalid detection. As shown in Figure 3, feature maps of different sizes can be obtained through VGG16 network structure and convolution layer processing. In order to expand the sampling on the upper feature map, we can associate the information of different feature layers to obtain the lower feature map. For example, for the bottom feature map obtained by channel fusion, after four stitching operations, each feature layer is connected to each other and finally four different feature layers can be obtained, which are, respectively, 38×38, 19×19, 10×10, and 5×5 size feature map.

Each lower feature and upper feature are fused, and they are used as input values for in-depth processing by the channel attention model. After obtaining effective features, each feature point is taken as an object and sampled according to the size of the detection area. Then, the sampled frames are classified, and the qualified frame values are retained after classification. After cross-layer fusion and effective channel attention module, six feature maps with different scales containing rich feature information are finally obtained. Then, the suppression algorithm is used to determine the position and threshold of each frame, and finally the required prediction frame is obtained.

Different feature structures are constructed through Conv4_3, Conv7, and other convolution layers, and different objects are predicted by the SSD network model detection algorithm. Because the number of shallow convolution layers is small, the features extracted by Conv4_3 and Conv7 are generally large, containing the information of small objects that need to be further detected. Due to the detection of small targets, it needs to include a high-resolution feature map and some semantic information to distinguish the target and background information. However, the features obtained after convolution operation of the convolution layer can usually form a larger acceptance domain, which contains large object information that needs to be further detected. SSD network model uses the features obtained by Conv4_3 when detecting small objects. Due to the lack of information, it is difficult to meet the needs of small object detection, which easily produces problems such as false detection and missed detection and also reduces the detection accuracy of the model. Therefore, this article integrates the upper sampling of high-level features and low-level features so as to ensure that the high-level feature attributes are not lost and the shallow feature information is retained at the same time. Due to the differences in the number and size of channels in each feature layer of the SSD network model, feature preprocessing is usually needed to meet the requirements of feature fusion in different layers.

For example, for a size of 19×19 and 5×5, the interpolation method can be used to expand the feature with size 5×5 to make its size become 19×19 and then compare it

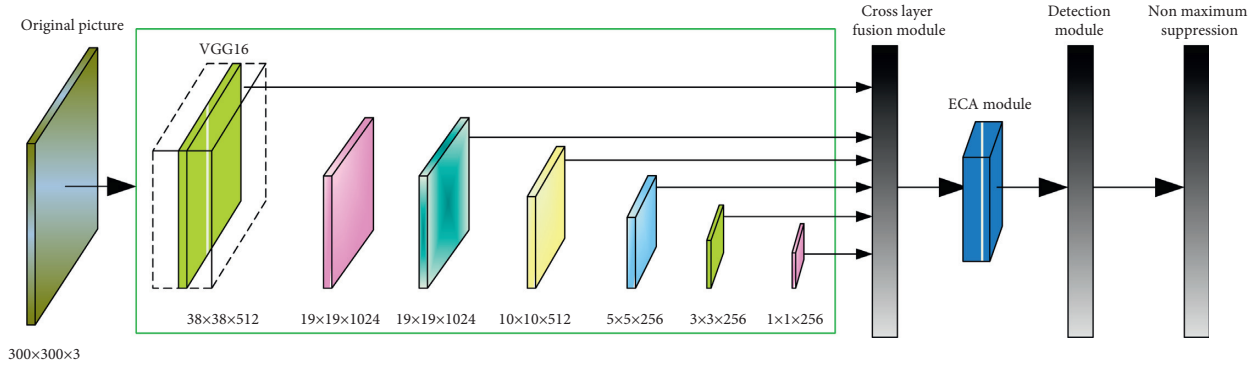

Figure 3: Structure diagram of the improved SSD face detection network.

with the existing size of 19×19, and the corresponding number of channels is obtained by a convolution operation. Connect and fuse the feature layers of different sizes of SSD network model so as to obtain the feature layers of different sizes, as shown in Figure 4.

4.2. Convolution Block Attention Module. As a way of resource reorganization and distribution, the idea of attention mechanism mainly comes from people's observation and thinking process of things. It uses the obtained external information, extracts the focus of attention through analysis and judgment, and allocates its existing resources according to the concerned area. When processing image data, there will also be areas with high-value features that need to be focused on. For face detection, we usually not only get the required global information from the input image but also pay attention to the local characteristics, such as local face attributes or features, so as to verify the effectiveness of the network structure model in face image feature extraction and detection. Attention mechanisms widely used in the field of artificial intelligence mainly include soft attention mechanism and strong attention mechanism model [20]. "Soft attention mechanism" mainly adopts the way of image segmentation to obtain attention frames and channels of different sizes. Different soft attention can be distinguished from each other. Therefore, the soft attention mechanism usually uses a neural network and inverse training to obtain the attention weight parameters. A strong attention mechanism generally expands each information point in the face image features and gets attention. Because it has certain dynamics, it can effectively train attention through reinforcement learning.

Convolutional block attention module is a simple and effective modular attention network. For the input feature map object, the convolution block attention module (CBAM) successively uses the channel attention module and spatial attention module to recombine the input feature map combined with the obtained attention feature map so as to generate a new feature map for attention. Usually, the CBAM module retains the original feature map information. Therefore, the model can be combined with CNN to jointly train all the extracted features. The specific composition of the convolution block attention module is shown in Figure 5.

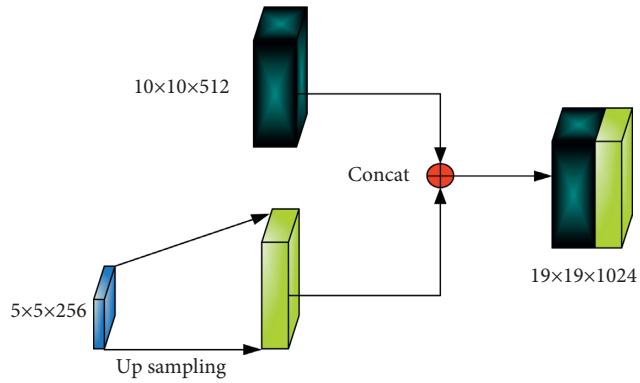

Figure 4: Schematic diagram of cross-layer fusion of features.

The CBAM module adopted in this article is divided into two submodules, namely, channel attention module (CAM) and spatial attention module (SAM), which are mainly used for resource allocation of attention in the channel and spatial dimensions [21]. Set the input intermediate feature map as F and input it into the CAM module to obtain the channel attention map. Multiply it by F to get F'. Similarly, input F′ into SAM module to obtain spatial attention map, and multiply it with F' to obtain the final input characteristic map F''. The overall process can be summarized as the following two formulas:

$$\begin{aligned} F' &= H_C(F) \otimes F, \\ F'' &= H_S(F') \otimes F', \\ F &\in \mathbb{R}^{W \times H \times C}, H_C \in \mathbb{R}^{1 \times 1 \times C}, H_S \in \mathbb{R}^{W \times H \times 1}, \end{aligned} \tag{4}$$

where F is the intermediate characteristic graph and $\otimes$ represents defined as "element-wise" multiplication; that is, the matrix is multiplied element by element, and the attention value is broadcast through multiplication. F' denotes the feature map obtained by combining the CAM module, F'' is the feature map finally obtained by the convolution block attention module through the SAM module, C is the number of feature map channels, H is the height of the feature map, and W is the width of the feature map.

This article uses the channel attention module. The basic idea of attention is to focus on important areas in the visual range and ignore irrelevant information. The spatial

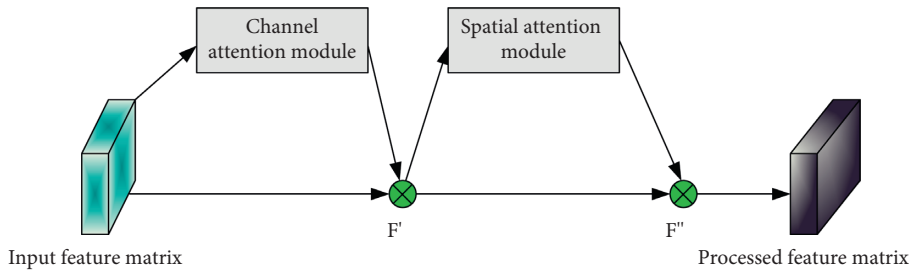

Figure 5: Structure and function diagram of the convolution block attention module.

attention network mainly transforms the image by cutting, scaling, and rotation to solve the image deformation deviation and compensate the position information. As shown in Figure 6, the core of the channel attention network is that each channel is assigned different weight values. The greater the weight, the higher the correlation. In the neural network, each initial picture contains three channels (R, G, B). After multiple convolution operations, a new channel matrix (H, W, C) is formed. H and W represent the height and width of the feature map, respectively. The contribution of C channels to target detection is different, which introduces channel attention and focuses more attention on effective channels. The structure and function diagram of the channel attention module is shown in Figure 6.

The input picture is located on the left side of the channel attention module of the module, as shown in Figure 6. After the convolution operation is used to calculate and process the input image information, a new feature map can be obtained, including the number of channels. In addition, the convolution layer operation can convert each global feature into a global receptive field, reduce the spatial dimension, and output the feature map of $1 \times 1 \times C$. The excitation operation is combined with 1×1 convolution, carries out the learning and training of correlation degree for each channel, uses the obtained correlation degree for feature matching, and outputs the feature matching result according to the channel attention module. Finally, the weighted operation is combined with the original features to enhance the main output features and remove other unimportant information. The excitation operation reduces the dimension of the channel and then expands it back to the original number of channels, which reduces the amount of calculation of the network.

In this article, a more effective channel attention module (ECA) is added on the basis of the SENet network structure. The module improves the detection performance without increasing the complexity of the model. ECA module requires fewer parameters and convolution layers and can interact with the information contained in different convolution layers so as to improve the efficiency of face detection based on a network model. Therefore, compared with the general channel attention module, the ECA module can not only reduce the number of parameters required by the model to a great extent but also greatly improve the classification and detection of image features.

Through the information fusion processing between different layers, the ECA module can effectively learn and train each feature and finally use the feature matching results to obtain the required feature attributes. The feature map generated by the ECA module has rich information and focuses on the effective feature area to reduce various overheads and improve the detection speed and efficiency. The specific area concerned with the ECA module often contains the required objects, so it is more conducive to face detection and recognition.

4.3. Loss Function. The research shows that whether the network structure model can meet the needs of face detection is closely related to the loss function used. A well-run network model needs to repeatedly use the collected dataset for learning and training and then correct the relevant parameters and convolution layer according to the training results. Among them, the loss function is the key to model training and correction. Usually, the loss function is a nonnegative function and determines the relationship between network input and output. The smaller the loss function value obtained after learning and training, the more reasonable the model is. In the process of model modification, face detection often adopts classification and regression analysis methods, so the loss function includes classification loss function and regression loss function [22].

The classification loss function used in this article is the commonly used cross-entropy loss function. It proves the learning effect of the model according to the difference between the output value of the network structure model and the actual value. The smaller the difference, the closer the between the output value and the actual value. There are two kinds of cross-entropy loss functions: one is used to solve binary classification problems, as shown in formula (5), while formula (6) is often used to solve multi-classification cases.

$$M(y,z) = -\sum_{x} y(x)\log z(x), \tag{5}$$

$$M(y,z) = -\sum_{x} (y(x)\log z(x) + (1 - y(x))\log(1 - z(x))), \tag{6}$$

where y represents the real probability distribution of the dataset, z denotes the probability distribution of the network output data, and $M(y,z)$ is the cross-entropy loss function.

The cross-entropy loss function used in this article mainly combines the Sigmoid or Softmax activation

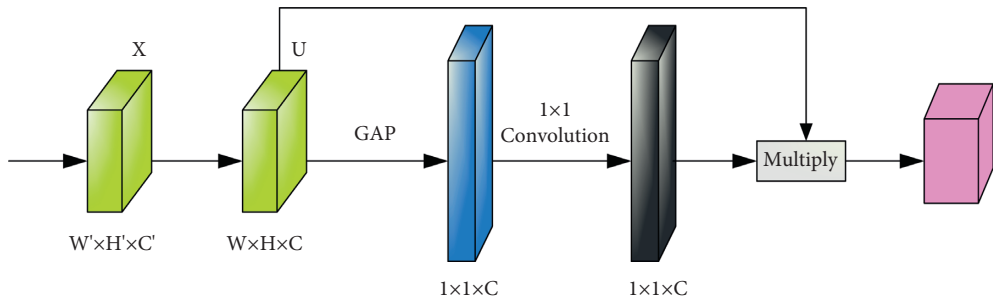

Figure 6: Structure and function diagram of the channel attention module.

function. When dealing with binary classification, the combination of the binary cross loss function and the Sigmoid activation function is used. The combination of the multivalued cross loss function and Softmax activation function can solve the multiclassification situation, while the processing of multilabel situation is mainly the combination of the binary cross loss function and Sigmoid activation function.

It is known from the existing research that the regression loss function is often used in the network structure model for face detection and recognition, such as IOU loss function and focal loss function. Most of these functions are used to solve the sample imbalance caused by up- and down-sampling and the problem that learning samples cannot be distinguished and are difficult to detect. Therefore, in the network model, we can increase the number of these samples to focus on the detection of samples so as to obtain better detection effect, as shown in the following formula:

$$\mathrm{Fl} = -\lambda (1 - z(x))^{\theta} \log z(x). \tag{7}$$

When the classification is wrong, Fl will be very small, $(1 - z(x))$ will be close to 1, and the value of T_m of the loss function will be almost 0, which will not affect the network model. The function of parameter θ is to smooth the weight of difficult samples.

IOU loss function is mainly used to screen detection frames. It screens the detected object according to the difference between the predicted frame and the real frame output by the model. When the difference is greater than the set threshold, the prediction frame output by the model can be used as candidate data, and when the difference is less than the set threshold, the prediction frame is eliminated. That is, the calculated value of the IOU loss function is used to determine the choice of prediction frame. When the value is larger, the prediction frame is closer to the actual frame. If the real box is A and the prediction box is B, the IOU calculation formula is as follows:

$$\begin{aligned} \mathrm{IOU} &= \frac{A \cap B}{A \cup B} \\ &= \frac{A \cap B}{A + B - A \cup B}. \end{aligned} \tag{8}$$

In addition, L1/L2 loss function is often used in face detection. The detection frame in face detection needs to occupy a certain position, so the L1/L2 loss function cannot be used after modifying the position; however, it is normal for the IOU to lose its function. The L1/L2 loss function cannot calculate the specific location, but the IOU loss function can obtain the specific location information. For example, if the position deviation of the detection frame is larger, the calculated value of IOU loss function is smaller; otherwise, the calculated value of IOU loss function is larger. Therefore, in the process of modifying the face detection model, the IOU loss function is used to verify the model, and the effect is better.

5. Experiment and Analysis

5.1. Dataset Description. Datasets and evaluation criteria play a very important role in improving model performance and quantitative evaluation. In order to improve the detection performance and efficiency of the adopted network structure model, the commonly used typical datasets Wider Face, MAFA, and COFW are selected in the experiment so as to strengthen the training of the model using the complete data and clear labels provided by the face dataset so as to make the experiment more objective and fair and reproduce the results.

The Wider Face dataset contains 32203 images and 393703 labeled faces, which is 10 times larger than the largest face detection dataset at present. A large number of label faces, which have certain complexity in shape, size, shape, and interference, are rich in data. They are datasets with the greatest detection difficulty and the highest data richness in the current open-source dataset. Select one part as the training subset, the other part as the verification subset, and the rest as the test subset. According to the detection difficulty of the test dataset, the verification set and the test set are divided into three levels: simple, medium, and difficult.

MAFA belongs to the dataset of mask faces, including 30811 pictures collected by the Internet. Each picture contains at least one face disturbed by the external environment. These disturbed faces are not available in the conventional face dataset. The dataset includes six attributes marked manually, namely, face position, glasses position, occlusion position, face direction (left, middle, right, left front, and right front), occlusion degree (strong, medium, and weak), and occlusion type (simple mask: solid color artificial occlusion; complex mask: complex line artificial occlusion, human body occlusion, and mixed occlusion). The dataset can be well adapted to the training and

Table 1: The training set and testing set of each data set required for the experiment.

Data set	Number of samples		
	Training set	Testing set	Total
Wider Face	4500	3300	7700
MAFA	5500	3500	9000
COFW	5800	2500	8300

Table 2: Effects of different network structures on evaluation parameters.

Network structure	FPS	mAP (%)	AR (%)
SSD	46	76.6	87.9
SSD + SeNet	40	77.5	88.4
SSD + YOLOV	42	78.2	90.5
Improved SeNet	36	77.8	89.2
Improved SSD	39	78.6	88.5
This method	43	80.4	92.6

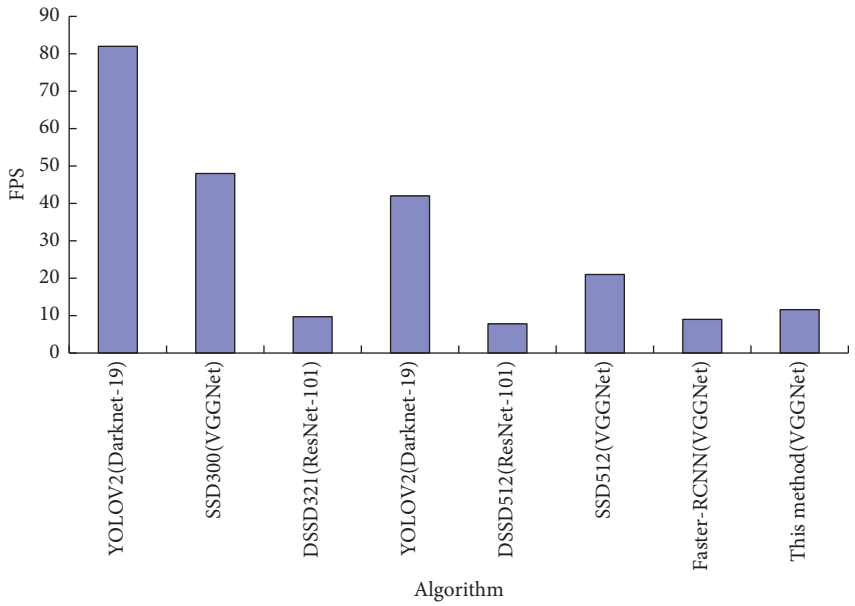

Figure 7: Comparison results of FPS obtained using different network structure models.

optimization of the face detection dataset and deep learning model in a complex environment.

COFW is an earlier partially occluded face detection dataset. Its purpose is to detect the location of face marker points in occlusion environment, including 1852 annotated faces with occlusion. The training set contains 1345 images without occlusion, and the test set contains 507 images with occlusion, with an average occlusion rate of about 23%. Among them, 329 images were occluded by more than 30%, belonging to severe occlusion, and the remaining 178 images were slightly occluded.

Wider Face, MAFA, and COFW datasets are used for experiments, and the datasets are shown in Table 1.

The video contained in Wider Face, MAFA, and COFW datasets is read by Opencv and saved for 30 frames. After the obtained image frames are aligned and cut by Mtcnn face, the image frame size is normalized to 124×124 RGB image.

5.2. Evaluating Indicator. In order to better compare the effects of different methods on face detection, this article uses frames per second (FPS) as the speed evaluation index to represent the number of images that can be processed per second and takes the average recall rate (AR) and average accuracy rate (mAP) of each image category as the evaluation index of face detection accuracy.

5.3. Results and Analysis. The ablation experiment is used to verify the effectiveness of the algorithm in this article. The

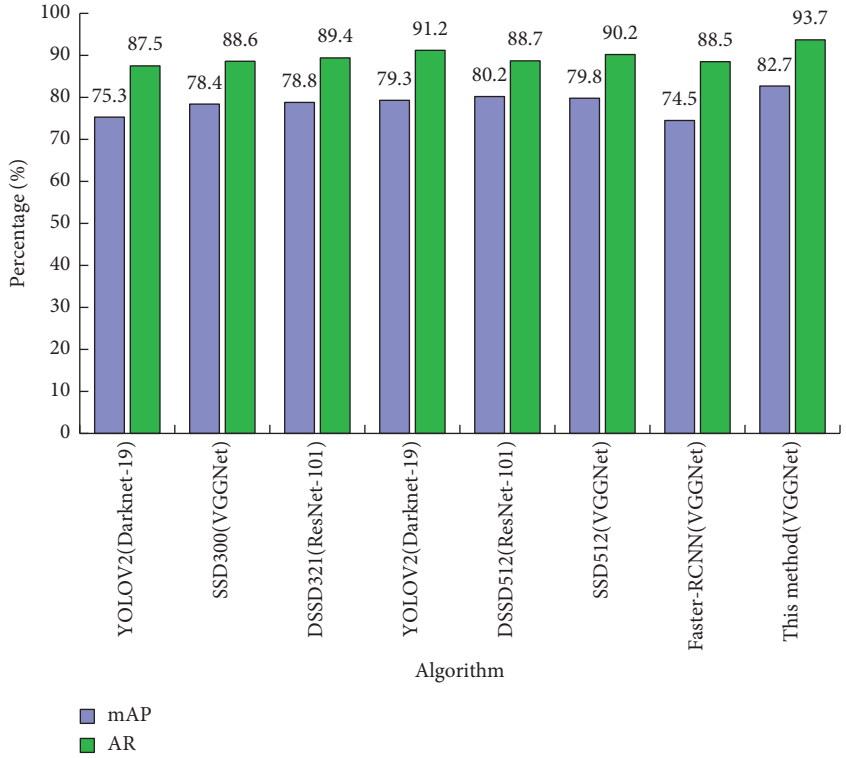

Figure 8: Comparison results of mAP and AR obtained using different network structure models.

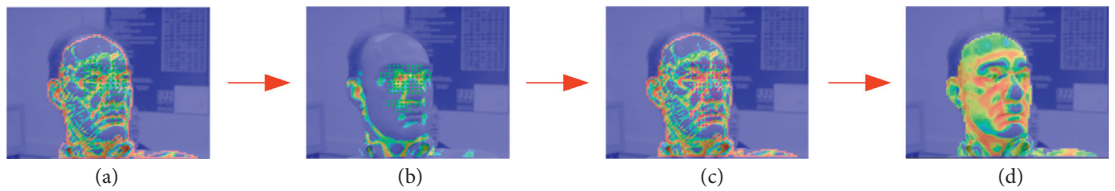

Figure 9: Detection results of local face attributes based on SSD.

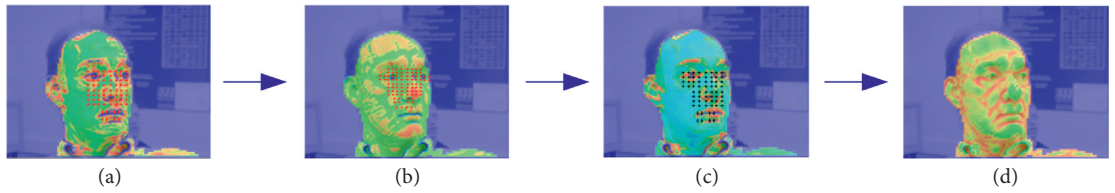

Figure 10: Detection results of local face attributes based on improved SSD.

index values are obtained according to different network structure models, and the results are shown in Table 2.

It can be seen from Table 2 that in the improved method, adding an attention mechanism based on bidirectional feature fusion can improve the average accuracy by 1.3%, the improvement of the anchor can improve the average accuracy by 0.8%, and the overall average accuracy can be improved by 3.5%.

The model in this article is trained on MAFA and COFW training sets. The results on the MAFA test set are compared with the experimental results of mainstream algorithms such as YOLOV2, DSSD321, and Faster R-CNN, as shown in Figures 7 and 8.

It can be seen from Figure 8 that when the input image size is 300×300, the average accuracy of the model proposed in this article is 7.4%, 3.9%, and 4.9% higher than that of YOLOV2, DSSD321, and Faster R-CNN, respectively. At the same time, the SSD model detects 46 images per second, while the improved SSD model detects 33 images per second. Compared with the SSD model, the detection speed of the improved model decreases slightly because the amount of calculation of the model increases when the bidirectional feature fusion is improved, which affects the detection speed of the model.

In order to verify the effectiveness of this algorithm, the target detection results of the original SSD algorithm and

this algorithm are visually displayed, as shown in Figures 9 and 10. In this article, the detection frame fits more closely with the target and improves the missed detection and false detection of small targets to a certain extent.

The results show that the average recall (AR) and mean average precision (mAP) are higher than other methods, and the accuracy is improved. The experimental results of different algorithms on various datasets show that this method is superior to other algorithms in various indexes.

6. Conclusion

Based on SSD and vgg16 detection framework, this article studies complex face detection, including face local attributes, by enhancing feature data at different layers and fusing convolution block attention mechanism. It is verified and tested on the public dataset, and a relatively good balance is achieved between detection accuracy, data rate, and various resource allocation. Faster R-CNN is used for general target detection, but it can still show excellent face detection performance when retrained on appropriate face datasets. By considering the special pattern of the human face, its performance can be further improved. In this article, the Fast R-CNN target detection method is applied to face detection. The recall rate, average accuracy, and detection time of face detection are comprehensively analyzed, and the SSD network model is finally selected. Using the trained model to test in the test set, the AR is 93.7%, the map is 82.7%, and the pixel size is 512×512. The method in this article is applied to face detection in video, which can realize real-time detection.

Data Availability

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

Conflicts of Interest

The author declares no conflicts of interest.

Acknowledgments

This work was supported by the Science and Technology Project of Chongqing Education Commission: "Research on Personalized Learning Video Recommendation Algorithm for Basic Intelligence Education" (KJQN202103901 and 2021YC-JCKX20025).

References

[1] J. Daugman, "Face and gesture recognition: overview," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 675-676, 1997.

[2] Y. Gao and Y. Qi, "Robust visual similarity retrieval in single model face databases," *Pattern Recognition*, vol. 38, no. 7, pp. 1009–1020, 2005.

[3] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: a survey," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–741, 1995.

[4] J. Zhang, X. Hu, Z. Ning et al., "Energy-latency tradeoff for energy-aware offloading in mobile edge computing networks," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2633–2645, 2018.

[5] B. Kepenekci, F. B. Tek, and G. B. Akar, "Occluded face recognition based on Gabor wavelets," in *Proceedings of the IEEE International conference on image processing*, pp. 373–378, Rochester, NY, USA, September 2002.

[6] P. Nagesh and B. Li, "A compressive sensing approach for expression-invariant face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1518–1525, Miami, FL, USA, June 2009.

[7] X. F. Fu, Y. Zhang, and J. Wu, "Joint auxiliary dictionary learning and low rank decomposition face recognition under occlusion expression changes," *Chinese Journal of Image Graphics*, vol. 23, no. 3, 399 pages, 2018.

[8] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5325–5334, Boston, MA, USA, June 2015.

[9] B. Hu, G. P. Gao, L. L. He, X. D. Cong, and J. N. Zhao, "Bending and on-arm effects on a wearable antenna for 2.45 GHz body area network," *IEEE Antennas and Wireless Propagation Letters*, vol. 15, pp. 378–381, 2016.

[10] Z. Zhifeng Li, X. Dahua Lin, and fnm Xiaoou Tang, "Nonparametric discriminant analysis for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 755–761, 2009.

[11] A. Ibarrola and E. Chavez, "A robust entropy-based audio-fingerprint," in *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo*, pp. 1729–1732, Toronto, ON, Canada, July 2006.

[12] S. Yang, P. Luo, C. C. Loy, and X. Tang, "From facial parts responses to face detection: a deep learning approach," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3676–3684, Santiago, Chile, December 2015.

[13] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.

[14] X. Tan, S. Chen, Z. H. Zhou, and J. Liu, "Face recognition under occlusions and variant expressions with partial similarity," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 2, pp. 217–230, 2009.

[15] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," *Computer Vision - ECCV 2016*, pp. 354–370, 2016.

[16] M. Sunagawa, S. I. Shikii, W. Nakai, M. Mochizuki, K. Kusukame, and H. Kitajima, "Comprehensive drowsiness level detection model combining multimodal information," *IEEE Sensors Journal*, vol. 20, no. 7, pp. 3709–3717, 2020.

[17] S. Maciej, "Liveness measurements using optical flow for biometric person authentication," *Metrology and Measurement Systems*, vol. 19, no. 2, pp. 257–268, 2012.

[18] M. Belkin and P. Niyogi, "Towards a theoretical foundation for Laplacian-based manifold methods," *Journal of Computer and System Sciences*, vol. 74, no. 8, pp. 1289–1308, 2008.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[20] X. Hu, J. Cheng, M. Zhou et al., "Emotion-Aware cognitive system in multi-channel cognitive radio ad hoc networks,"

IEEE Communications Magazine, vol. 56, no. 4, pp. 180–187, 2018.

[21] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "Cbam: convolutional block Attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, September 2018.

[22] J. Ma, Y. H. Cai, and N. Sheng, "3D facial expression recognition based on block CBP features and sparse representation," *Computer System Application*, vol. 28, no. 2, 196 pages, 2019.