IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 July 2022, accepted 14 July 2022, date of publication 18 July 2022, date of current version 22 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3192026

RESEARCH ARTICLE

Fine Segmentation on Faces With Masks Based on a Multistep Iterative Segmentation Algorithm

MIN ZHANG[1,2,3], **KAI XIE**[1,2,3], **YU-HANG ZHANG**[3,4], **CHANG WEN**[3,4], **AND JIAN-BIAO HE**[5]

[1]School of Electronic Information, Yangtze University, Jingzhou 434023, China
[2]National Demonstration Center for Experimental Electrical and Electronic Education, Yangtze University, Jingzhou 434023, China
[3]Western Institute, Yangtze University, Karamay 834000, China
[4]School of Computer Science, Yangtze University, Jingzhou 434023, China
[5]School of Computer Science and Engineering, Central South University, Changsha 410083, China

Corresponding author: Kai Xie (pami2009@163.com)

This work was supported in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region under Grant 2020D01A131, in part by the Fund of Hubei Ministry of Education under Grant B2019039, in part by the Graduate Teaching and Research Fund of Yangtze University under Grant YJY201909, in part by the Teaching and Research Fund of Yangtze University under Grant JY2019011, in part by the Undergraduate Training Programs for Innovation and Entrepreneurship of Yangtze University under Grant 2019099, and in part by the National College Student Innovation and Entrepreneurship Training Program under Grant 202110489003.

ABSTRACT Innovation in facial recognition technology is urgent in the current epidemic situation. In order to extract more face features to perform facial recognition on images of faces with masks, we propose a multi-step iterative face-mask segmentation algorithm (MISA). The improved Mask R-CNN algorithm is used to detect the target in the face region to realize coarse segmentation, and the generalization ability and accuracy of the segmentation are improved. Then, the proposed approach uses an R-Pairwise Differential Siamese Network (R-PDSN) to train a mask occlusion classifier to subdivide the edge blocks of coarse segmentation results. The segmentation accuracy is further improved by optimizing the edge information of the masks step by step. The self-built dataset of faces with masks was used for training and testing. The experimental results showed that the mean pixel accuracy of the proposed method was improved by 2.69% compared with the original Mask R-CNN segmentation algorithm, and the target detection accuracy was more than 98%. These results indicate that the proposed method can achieve good segmentation performance on face images with complex backgrounds, self-occlusion and different types of masks. These results demonstrate that our method can improve the accuracy of segmentation methods for imaged of faces wearing masks.

INDEX TERMS Face segmentation, mask R-CNN, occlusion classifier, R-pairwise differential Siamese network, edge segmentation.

I. INTRODUCTION

The coronavirus pandemic remains a serious threat to human life and health worldwide. Owing to the respiratory transmission mode of the virus, the development of non-contact facial recognition [1] methods operable on people wearing masks has been noted as a problem of the highest priority. To retrieve facial features from images of masked faces more accurately, most methods detect and locate the area occluded by the mask on the exposed parts of a person's face in an image or video. This approach aims to extract more useful face features and eliminate the similarity of masks and other occlusion features that affect the recognition of the inter-class distance.

In recent years, the detection and classification of occlusions have been studied extensively in the field of computer vision. Conventional methods include sparse representation [2] and occlusion dictionary error coding [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Sergio Consoli.

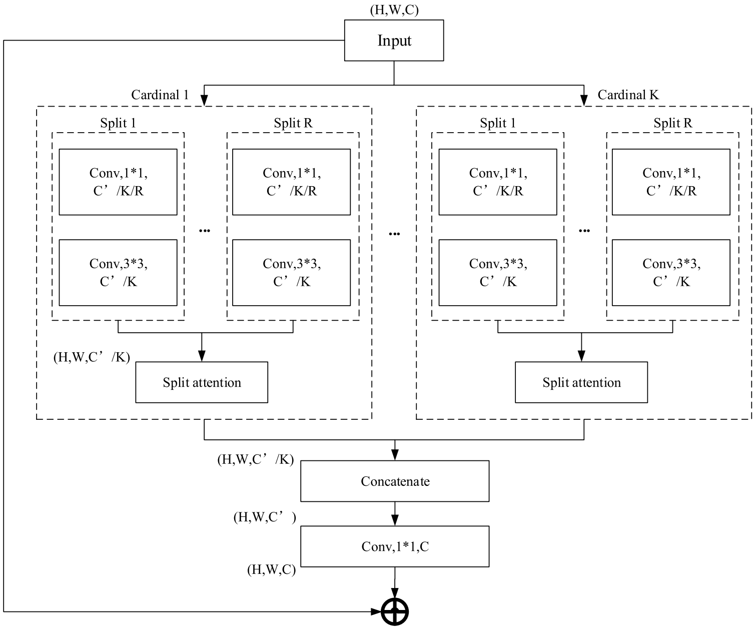


FIGURE 1. ResNeSt module, which contains a feature map group and split attention operation.

These methods locate, limit, and delete occlusions by comparing the structural features of occlusions with global structural facial features. Existing occlusion methods can be classified as either target detection methods or semantic segmentation methods [4]. Target detection methods judge the position of an occlusion based on its contours, geometric features, and the continuity of features of a target face in an image by detecting, positioning, and classifying occlusions in an image space. Yang *et al.* [5] screened the spatial and feature information of each part of the face using Faceness-NET and then reordered the face’s suggestion box, obtained from the feature score of each facial part, to generate a face location box with a high recall rate. This method directly ignores the influence of severe occlusion and different poses, can detect faces effectively. However, for the occlusion area, the target detection algorithm can only locate a rectangular box; it cannot accurately locate irregular and discontinuous occlusion areas. Semantic segmentation algorithm solves this problem by using pixels to classify occlusions. The Deeplab series [6] used conditional random fields to solve the problem of inaccurate edge segmentation and spatial pyramids to solve the problem of fuzzy segmentation at different scales. Semantic segmentation involves some limitations and is unstable in cases of complex backgrounds and uneven light [7]. Lin *et al.* [8] used an improved Mask R-CNN algorithm to integrate face detection and semantic segmentation algorithms to obtain accurate information for face segmentation. Subsequently, they used the generalized intersection over union [9] as a boundary loss function to improve the accuracy of face segmentation and address the shortcomings of semantic segmentation. Meanwhile, owing to the diversity and randomness of occlusions, few semantic segmentation datasets and research methods are available for occluded faces. Moreover, the manual production of datasets inevitably leads to inaccurate labeling of areas owing to subjective factors that cause noise interference in the model. Liang *et al.*, based on the segmentation results of Mask R-CNN instance, used Pool Net to optimize edge details with better edge feature extraction. Pool Net [10] works well for targets with significant non-overlap, while it is difficult to tell apart targets with high foreground overlap where the mask fits tightly to the face.

In this study, we propose a multi-step iterative segmentation algorithm (MISA) to solve this problem. This algorithm uses the improved Mask R-CNN network to perform coarse segmentation of the self-built dataset and incorporates the ResNeSt-Pairwise Differential Siamese Network (R-PDSN) [11], an edge detection algorithm. The algorithm optimizes the segmentation effect of the coarse segmentation results through multiple iterations to improve the segmentation accuracy. Finally, we conducted multiple groups of experiments to verify the generalization and high precision of the proposed MISA.

The remainder of this paper is organized as follows. Section II introduces the theoretical foundations of the models used in the proposed algorithm. Section III presents the details of the training process along with the experimental setup and results, as well as the results of a comparison with other methods. Finally, Section IV presents our conclusions and suggests some possible directions for future research.

TABLE 1. Comparison of advantages and disadvantages among ResNet, ResNeXt and ResNeSt.

	Advantages	Disadvantages
ResNet	eases the problem of network degradation	the receptive field is fixed and single, and cannot integrate features of different scales, and the structure is difficult to expand
ResNeXt	proposed a new module form called cardinality; proposed group convolution, multi-branch and aggregation	features of different scales cannot be extracted when there are many network layers
ResNeSt	inherits the advantages of ResNeXt and adds the multi-path mechanism to focus on local information extraction capability	large amount of calculation

II. RELATED WORK

The basic architecture of some common downstream applications (e.g., semantic segmentation and target detection methods) uses ResNet to perform feature extraction because it solves the problem of network degradation to learn deeper facial characteristics. However, the size of the ResNet receptive field is fixed and single. Therefore, it cannot integrate features of different scales and does not perform well for pixel-level classification tasks such as semantic segmentation. ResNeXt [12] inherited the stack strategy of ResNet but combined the split-transform-merge strategy of GoogleNet and proposed a new module form called ‘cardinality’. In this study, we used ResNeSt [13] as the backbone feature extraction network to improve the Mask R-CNN. The ResNeSt module diagram is shown in Fig. 1. ResNeSt contains a feature map group and split-attention operations. In the feature map group, the feature map itself is divided into multiple groups, with other groups within each group.

The hyperparameters K and R represent the number of feature graph groups and groups in the radix group, respectively. The group in the radix group is called “splits,” and the total number of feature graph groups can be expressed as $G = K \times R$. 1×1 and 3×3 convolutions were carried out for each split in the radix group. After obtaining the R feature graphs, a split attention operation was performed. The output from each cardinality group was concatenated and convolved by 1×1 in the shortcut path. Multiple ResNeSt modules were stacked to form the final ResNeSt network. ResNeSt (combined with ResNeXt’s multibranch, group convolution method, and GoogleNet’s multipath mechanism) expands the receptive field, enhances the interaction of feature layers, and improves the feature expression of different receptive fields. In addition, ResNeSt introduces split-attention blocks and uses channel attention to construct features based on local information. It obtains feature information of R subgroups of feature maps in each cardinality group and realizes information interaction across feature maps, which improves segmentation accuracy without affecting complexity. The advantages and disadvantages of ResNet, ResNeXt, and ResNeSt are presented in Table 1.

Compared to the conventional Faster R-CNN [14] target detection algorithm, Mask R-CNN [15] adds a mask branch to segment a region on interest (RoI), which can simultaneously achieve face detection and face segmentation. Mask R-CNN uses the RoIAlign layer to integrate the pixel values of floating point numbers at the boundary of the entire candidate feature area into continuous pixel values through bilinear interpolation, avoiding the partial pixel loss caused by the two-integer quantization of the traditional RoI pooling layer [16]. A flow chart of the RoIAlign algorithm is presented in Fig.2. This method traverses each candidate box, divides them into $K \times K$ rectangular elements, preserves the floating-point feature points on the edge of the rectangular elements, determines the coordinate values of the four central positions in each rectangular element through bilinear interpolation, and finally outputs them through the pooling layer.

The bilinear interpolation algorithm is illustrated in Fig.3. Assume that the pixel value of the point S to be solved is f(x,y) and the four adjacent unit points are known as $T_{11} = (x_1, y_1)$, $T_{12} = (x_1, y_2)$, $T_{21} = (x_2, y_1)$, and $T_{22} = (x_2, y_2)$.

The algorithm first interpolates linearly in the x direction, resulting in the following formula:

$$f(R_1) \approx \frac{x_2 - x}{x_2 - x_1} f(T_{11}) + \frac{x - x_1}{x_2 - x_1} f(T_{21}) \qquad (1)$$

$$f(R_2) \approx \frac{x_2 - x}{x_2 - x_1} f(T_{12}) + \frac{x - x_1}{x_2 - x_1} f(T_{22}) \qquad (2)$$

where the R_1 and R_2 satisfy $R_1 = (x, y_1)$ and $R_2 = (x,y_2)$ respectively. Then, it interpolates linearly in the y direction, yielding the following formula:

$$f(S) \approx f(x, y) \approx \frac{y_2 - y}{y_2 - y_1} f(R_1) + \frac{y - y_1}{y_2 - y_1} f(R_2) \qquad (3)$$

III. METHOD

The proposed face segmentation algorithm for complex environments is divided into three parts, including instance

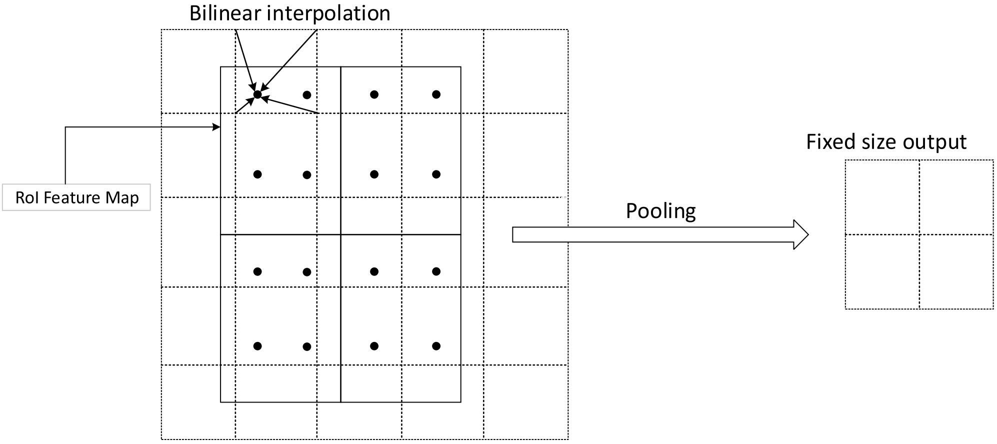


FIGURE 2. Flow chart of RoIAlign algorithm.

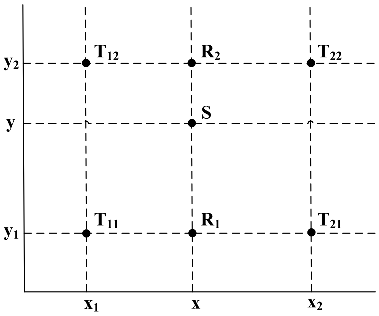


FIGURE 3. Illustration of bilinear interpolation algorithm.

segmentation, occlusion classification, and edge segmentation. In a complex environment, the instance segmentation part uses an improved instance segmentation network to perform coarse face segmentation. An occlusion classifier uses a bilinear difference network to extract the occlusion features of image pairs, and the edge segmentation part uses the edge subblocks of coarse face segmentation to refine the face region. The structure of the algorithm is illustrated in Fig.4.

We first used ResNeSt [17] as the backbone feature extraction network to improve the performance of Mask R-CNN in detecting the target face. This backbone was also used to achieve rough segmentation and improve the generalization ability and accuracy of the segmentation. Then, we used the R-PDSN method to train the occlusion classifier and subdivided the edge blocks of the coarse segmentation results, optimized the edge information continuously to achieve accurate step-by-step segmentation, which improved the segmentation accuracy of the model.

A. MASK R-CNN INSTANCE SEGMENTATION

1) BACKBONE NETWORK IMPROVEMENTS

ResNeXt [18] was designed to be more scalable than ResNet and includes multibranch, group convolution, and aggregation methods. ResNeXt performs a series of transformations on the feature map to fuse the processed outputs. This increases the accuracy and improves the representation ability of the network while only slightly changing or even reducing the complexity of the model. We used deep features to classify regionalized semantic information more easily, whereas shallow features were used to map target details. With an increase in the number of network layers, the feature extraction capability of ResNeXt was insufficient by a small margin. In this study, we used ResNeSt [13] as the backbone feature extraction network to improve the Mask R-CNN. ResNeSt contains a feature map group and split-attention operations. In the feature map group, the feature map itself is divided into multiple groups, with other groups within each group. Finally, the features intercepted by RoIAlign are used to reduce the dimension to the face detection vector form, and the original image size is obtained by upsampling for pixel classification to achieve segmentation.

2) EIOU

The intersection over union (IoU) [19] has two disadvantages as an evaluation index to predict the actual and predicted bounding boxes. First, reverse optimization is difficult when the predicted and actual boxes do not overlap. Second, when the two target boxes intersect in different directions, the IoU value may be the same, which cannot reflect the overlap depth. In this regard, we replaced the EIoU of the original Mask R-CNN with the expected IoU (EIoU). We then directly take the distance, height, and width of the center point of the prediction box as the regression of the penalty term constraint prediction box. The IoU is used as an evaluation index to predict the actual and prediction boxes. There are two disadvantages to using the baseline IoU [20] as regression loss. First, when the prediction box does not overlap with the actual box, the distance between the prediction boxes differs, but the loss value is the same. This leads to difficulty in reversing optimization. Second, when two target boxes intersect in different directions, the IoU value may be the same, which cannot reflect the overlap depth. In this regard, we used the

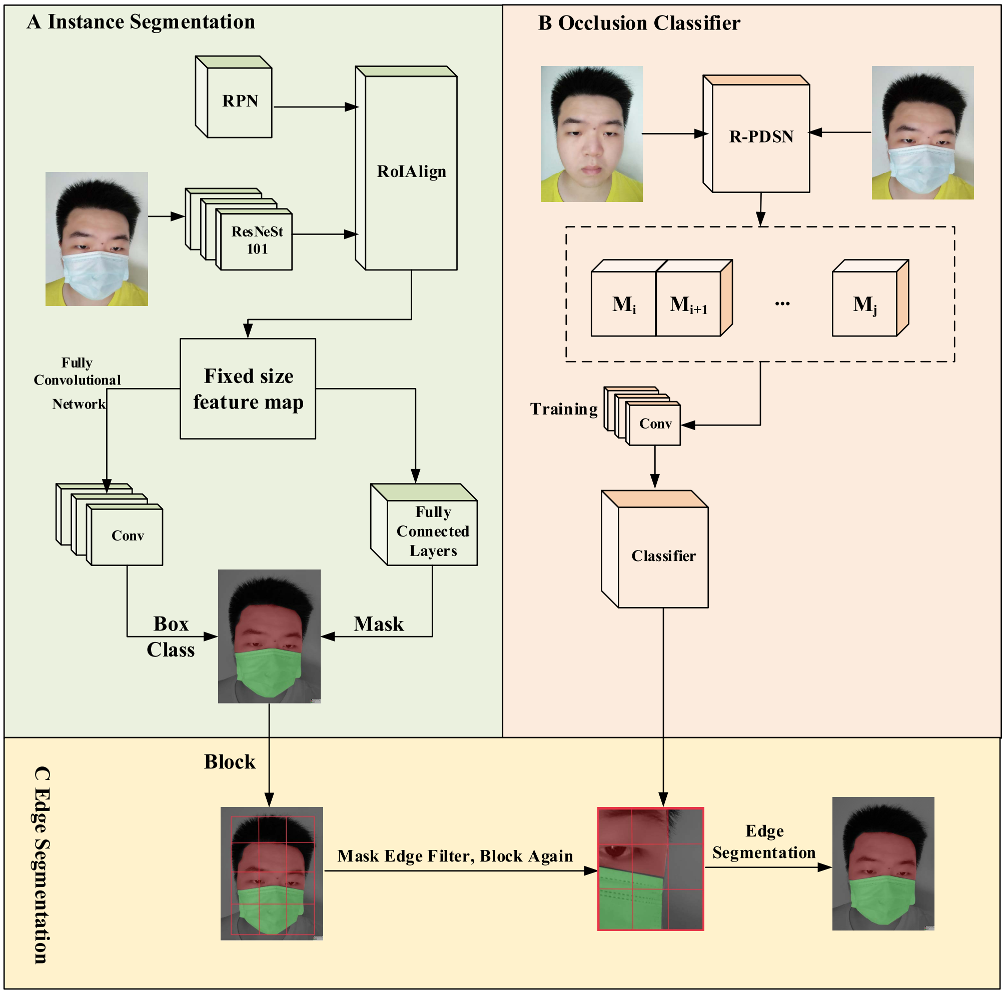


FIGURE 4. **Flow of face segmentation algorithm in complex environment. (A) shows an instance segmentation part using an improved instance segmentation network to perform coarse face segmentation. (B) shows an occlusion classifier using a bilinear difference network to extract the occlusion features of image pairs, and (C) shows the edge segmentation part using the edge sub-blocks of coarse face segmentation to refine the face region.**

improved EIoU to replace the IoU of the benchmark Mask R-CNN and directly restrict the regression of the prediction box by considering the center point distance, overlap degree, and scale of the prediction box as penalty terms. The distance between the center points reflects the distance between the prediction frame and the real frame, which overcomes the problem of non-overlap. The degree of overlap is the ratio of the outer rectangle of the center point to the smallest inner rectangle. Scale loss uses the height and width of the prediction box to optimize the size of the prediction box, making the network converge closer to the target region. EIoU considers the distance, overlap degree, and size of the bounding box between the actual and prediction boxes, which can cause the network to converge faster. The EIoU is shown in Fig.5. The diagonal coordinate of the solid line actual box is $(x_1^p, y_1^p, x_2^p, y_2^p)$, and the coordinate of the dashed prediction box is $(x_1^q, y_1^q, x_2^q, y_2^q)$. Then the distance loss can be expressed as:

$$L_{dis} = \frac{((x_2^q - x_1^q) - (x_2^p - x_1^p))^2 + ((y_2^q - y_1^q) - (y_2^p - y_1^p))^2}{4((x_2^q - x_1^p) + (y_2^q - y_1^p)^2)} \tag{4}$$

The scale loss can be expressed as:

$$L_{asp} = \frac{((x_2^q - x_1^q) - (x_2^p - x_1^p))^2}{(x_2^q - x_1^p)^2} + \frac{((y_2^q - y_1^q) - (y_2^p - y_1^p))^2}{(y_2^q - y_1^p)^2} \tag{5}$$

The IoU loss is:

$$S_q = (x_2^q - x_1^q)(y_2^q - y_1^q) \tag{6}$$

$$S_p = (x_2^p - x_1^p)(y_2^p - y_1^p) \tag{7}$$

$$S_\cap = (x_2^p - x_1^q)(y_2^p - y_1^q) \tag{8}$$

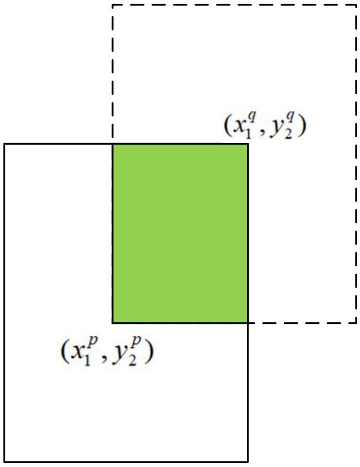


FIGURE 5. EIoU.

$$IoU = \frac{S_\cap}{S_q + S_p - S_\cap} \tag{9}$$

$$L_{IoU} = 1 - IoU \tag{10}$$

Then the bounding box loss function can be defined as:

$$Loss_{box} = L_{IoU} + L_{asp} + L_{dis} \tag{11}$$

B. OCCLUSION CLASSIFIER UNDER PAIRWISE DIFFERENTIAL SIAMESE NETWORK (PDSN)

To locate the face occlusion accurately, we used R-PDSN to map the occlusion feature into a single term for feature extraction. After extracting the deep features of the image pairs, ResNeSt is used to calculate the differences in the deep features of each pair of normal faces and faces with masks as the backbone network of PDSN. It then sends them to the occlusion classifier for training, which learns the nonlinear changes brought by occlusion to the deep features of faces. The components of the R-PDSN are shown in Fig.6. The PDSN consists of a backbone network, ResNeSt, and an occlusion classifier, forming a twin structure. The backbone network ResNeSt was used to extract deep feature information from the image pairs. We used the same backbone network for occluded and unconcluded face images to reduce memory usage and ensure the consistency of the deep feature scale and coarse segmentation. The output consists of two one-dimensional vectors, and the Euclidean distance between the two vectors represents the changes caused by the occlusion of the deep contour features of the face. This method can effectively eliminate the interference of noise and map high-dimensional image features onto a low-dimensional vector nonlinearly. We use $\sum_{i=0}^{j} M_i$ to describe the nonlinear changes in the deep features caused by the masks covering the face.

The training of this method involves a process of optimizing the loss function using the gradient descent method. Assuming that the input images are represented by $X1$ and $X2$, and W is the shared weight, then the output low-dimensional vectors $L1$ and $L2$ are:

$$L1 = (\sum_{f_i}^{N_i} \sum_{k_x}^{k} \sum_{k_y}^{k} W(f_i, m, n) F1(f_i, x + K_x, y + K_y)) + b \tag{12}$$

$$L2 = (\sum_{f_i}^{N_i} \sum_{k_x}^{k} \sum_{k_y}^{k} W(f_i, m, n) F2(f_i, x + K_x, y + K_y)) + b \tag{13}$$

where, f_i represents the input image pixel, $m{*}n$ is the size of the convolution sum K_x, K_y represents the step size, and b represents the offset of the input channel. The energy function of $L1$ and $L2$ can be expressed as:

$$E_i = ||L1| - |L2|| \tag{14}$$

As the energy function, E_i is used to represent the feature change rate after the occlusion feature is added. The higher the value of E, the greater is the influence of occlusion on facial features. However, various occlusions lead directly to differences in the output results of low-dimensional vectors; hence, we chose the simplest convolutional neural network to map the energy function to the interval [0,1]. This process encourages the network to pay more attention to the features of the occlusion and to detect the occlusion and complete the training process of the occlusion classifier.

C. EDGE SEGMENTATION

To solve the problem of inaccurate edge pixel segmentation caused by the artificial labels of segmented datasets, we performed fine segmentation of face edge partial features based on coarse separation. A flow chart of edge segmentation is shown in Fig. 7. We set the background label of the rough segmentation result as 2, the mask label as 1, and the face label as 0. After the first block processing, the blocks with sub-block category numbers greater than 1 are defined as the face edge blocks according to the result of coarse segmentation. These edge blocks involve seven types of circumstances, namely, (0,0), (1, 1), (2,2), (0, 1), (0,2), (1, 2), and (0,1,2), which indicate blocks with faces only, blocks with masks only, blocks with background only, blocks with faces and masks, blocks with faces and the background, blocks with masks and background and blocks with faces, and masks and background, respectively.

After dividing the edge blocks into blocks, {s11, s12, s13, . . . , sij} and $i{*}j$ sub-blocks are obtained. Each sub-block, such as *Mij*, is sent to the feature extraction network, which then distinguishes between blocks and face blocks according to the occlusion classifier. To obtain more face region information, we find the face edge blocks of the sub-blocks again and repeat the above operation. **Algorithm 1** lists the specific steps of the edge segmentation algorithm. Considering that different backgrounds cause random noise in occlusion classification, we set the course segmented background part to 1 to ensure that $E(x)$ of the background pixel is always 0.

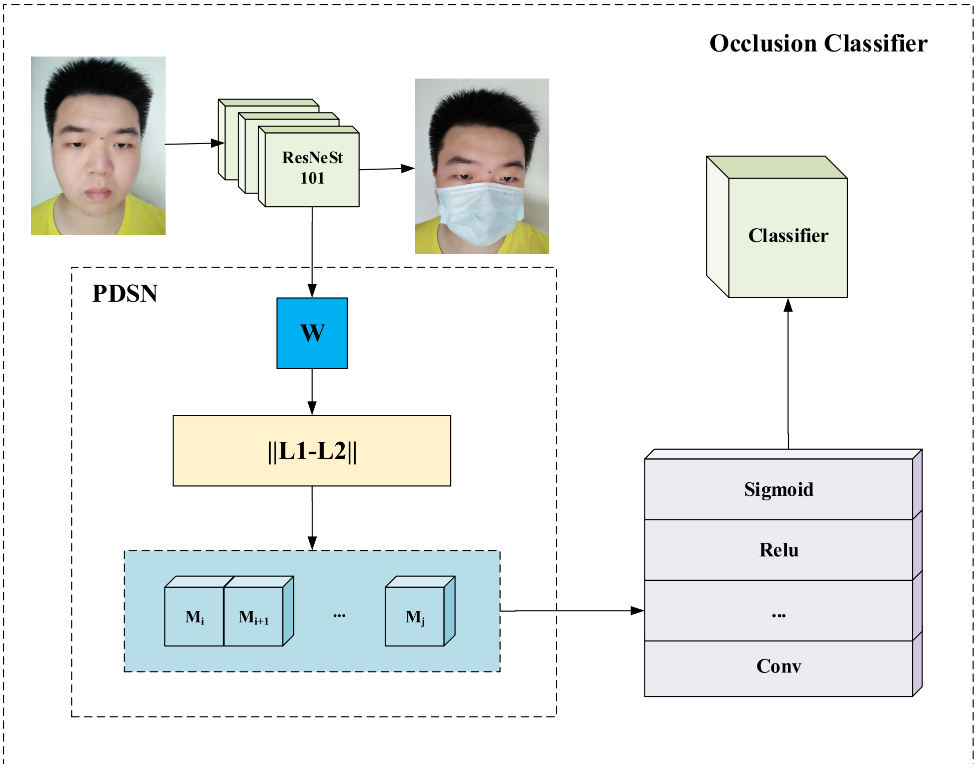


FIGURE 6. Occlusion classifier structure diagram. In the PDSN part, W is the shared weight, and L_1 and L_2 are the output low-dimensional vectors.

Through the classification of occlusion features by the classifier, we performed binarization of *Mij* and obtained the following formula:

$$\delta(i,j)=f(x)=\begin{cases}1, & E(x)>R\\ 0, & E(x)\leq R\end{cases} \tag{15}$$

where R represents the threshold value greater than 0 for occluded and complete faces.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present the experimental setup and results of an investigation conducted to evaluate the performance of the proposed approach.

A. EXPERIMENT SETTING

1) EXPERIMENTAL PLATFORM AND DATASET

The hardware device used in the instance segmentation performance comparison experiment is a Legion Y7000P workstation, with the Windows10 operating system, an Intel Core I7-10875h processor, a CPU with a clock rate of 2.30 GHz, 16 GB of memory, and a GeForce RTX 2060 GPU. For software, we used the PyCharm editor version 2020.1.2, along with the Python 3.7 programming language and TensorFlow as integrated libraries. We downloaded the public dataset of faces wearing masks from RMFD [21] and selected 200 unoccluded face images and 200 face images with masks from the dataset. Additionally, several images from video streams of each of several research assistants at a frame rate of 24 frames/s were edited as the basic dataset for this experiment. We then used a virtual mask to enhance the data of each open face image, and 1050 face data were obtained for the experiment. Images from some of the datasets are shown in Fig.8.

For multi-angled faces, we performed face correction [22] on the faces, and we used the MTCNN [23] face detection [24] model to position two key point coordinates of the eye and face frame information, where (x, y) and (h, w) represent the face frame at the upper-left corner coordinates. We then used the relationship between the left and right eyes and the position of the center point to calculate the deflection angle and corrected the image processing according to the angle. We set the position coordinates of the left eye as $\mathrm{L}(\mathrm{x}_1, \mathrm{y}_1)$ and the coordinates of the right eye as $\mathrm{R}(\mathrm{x}_2, \mathrm{y}_2)$. Then, the geometric center of the image is:

$$Center=(x+\frac{p}{2},y-\frac{q}{2}) \tag{16}$$

The Euclidean distance between the left and right eyes is:

$$D=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2} \tag{17}$$

The distance from the eye to the center point is:

$$D_L=\sqrt{(x_1-x-\frac{p}{2})^2+(y_1-y+\frac{q}{2})^2} \tag{18}$$

$$D_R=\sqrt{(x_2-x-\frac{p}{2})^2+(y_2-y+\frac{q}{2})^2} \tag{19}$$

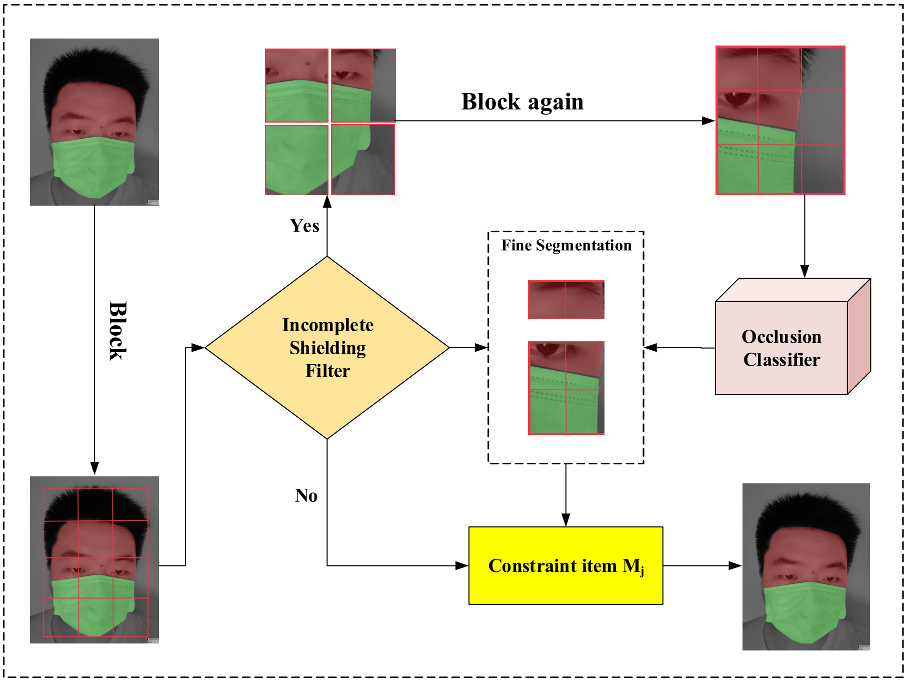


FIGURE 7. Edge segmentation calculation process.

Then, we obtained the face and face deflection angle as:

$$\theta = \arctan(\frac{y_2 - y + \frac{q}{2}}{x_2 - x - \frac{p}{2}}) - \arccos(\frac{D^2 + D_L^2 - D_R^2}{2DD_L}) \quad (20)$$

Based on the facial key points, we calculated the deflection angle and then aligned the face. Because the PDSN algorithm requires multiple pairs of faces in masks and unmasked faces, we amplified the dataset of unmasked faces wearing virtual masks. The faces and masks were divided into left and right halves, and the masks were scaled according to the deflection angle of the left and right faces to ensure that the masks could be worn more naturally.

After face pretreatment, owing to the irregular shape of the face and mask regions, we used Labelme software to label faces, masks, and background in 1050 experimental data images at the pixel level. Then, the face and mask region pixels in each image were divided. The pixel region of faces was called face, and the pixel region of masks was called mask. We represented different faces or masks in each image using different mask values and stored them in JSON files.

B. THE EXPERIMENTAL DETAILS

1) COMPARATIVE ANALYSIS OF COARSE SEGMENTATION ACCURACY

According to the proposed MISA algorithm, to verify the performance of the model quantatively, our experiment evaluated the experimental method from two aspects: target detection and segmentation. We used the ROC (receiver operating characteristic) curve and AP (Average-Precision) value to evaluate target detection performance, as shown in Fig.9. For a dichotomous problem, instances are classified into positive or negative classes. However, in actual classification, four situations are possible, including true positive (TP), false negative (FN), false positive (FP), and true negative (TN). The calculation formulas for the horizontal and vertical axes can be obtained as given below:

(1) The true-positive rate is:

$$TPR = \frac{TP}{TP + FN} \quad (21)$$

(2) The false-positive rate is:

$$FPR = \frac{FP}{FP + TN} \quad (22)$$

(3) The true-negative rate is:

$$TNR = \frac{TN}{FP + TN} \quad (23)$$

We used two evaluation indices, mean pixel accuracy (mPA) and segmentation running time, to verify the initial segmentation model. We set k as the number of categories, and the value of k in this experiment was three, including the background, mask, and face. P_{ii} stands for true positives (i.e., the number of samples of class i and were predicted to be class i), and P_{ij} stands for false positives (i.e., the number of samples of class i and were predicted to be class j). The formula for the PA (pixel accuracy) evaluation index is:

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \quad (24)$$

$$mPA = \frac{\sum_{i=0}^{k} AP_i}{k} \quad (25)$$

Algorithm 1 Edge Segmentation

```
divide the coarse segmentation image into M*N small blocks, respectively denoted by $P = \{P_{11}, P_{12}, \ldots, P_{MN}\}$
for p in P
    if A(a,b) = (0,0):
        E(x) = R + 1;
update the partition block with $P'_{(0,0)} = f(x) \cdot p$, according to Formula (14) after the constraint item
        else if A(a,b) == (1,1):
            E(x) = R
update the partition block with$P'_{(1,1)} = f(x) \cdot p$
according to Formula (14) after the constraint item
        else:
            for i in Lenth(A(0,1)):
                while m|n == 0:
                divide the separated A(0,1) into m*n
                blocks, denoted as P respectively
                    for $P^{(i)}_{\mathrm{mn}}$ in $P_{\mathrm{i}}$:
                        $E(x) = M(P^{(i)}_{mn})$
update the partition block with $P'_{mn} = f(x) \cdot p^{(i)}_{mn}$, according to Formula (14) after the constraint item
                        j = j + 1; $P'_i = \sum_{0}^{Lenth(P_i)} P'_{mn}$
                    end for
                end while
            end for
        $P' = \sum_{0}^{I} P'_i$
        end if
end for
$\widetilde{P} = p'_{(0,0)} + p'_{(1,1)} + p'$
return $\widetilde{P}$
```

We also used the Dice coefficient to measure the segmentation performance of the various methods. The Dice coefficient is a measure of similarity often used to compare two samples in the scope [0, 1]:

$$DiceConfficient = \frac{2\,|X \cap Y|}{|X| + |Y|} \tag{26}$$

where $|X \cap Y|$ is the intersection between X and Y and |Y| and |X| are respectively the number of elements of X and Y, and a molecular coefficient of 2.

The target detection accuracy of Mask R-CNN affects the segmentation direction and accuracy. In the test set of this experiment, DeeplabV3+, Mask R-CNN, Faster R-CNN, and the method in this study were used to carry out the comparison test of target detection, as shown in Fig.9:

To adapt to the uniform size of the network and hardware, the image and label were normalized to a size of 255 × 255 pixels and compared with DeeplabV3+ [25], Mask R-CNN with ResNeSt and EIoU, the original Mask R-CNN, and RefineNet, and all backbone networks used ResNeSt101. The datasets we constructed was used to divide the training, testing, and verification sets with a ratio of 8:1:1. There were 840 samples in the training set (700 faces wearing masks and 140 faces without shielding) and 105 photographs in the testing set (85 faces wearing masks and 20 faces without shielding). The validation set consisted of 105 photographs (80 faces with masks and 25 without masks). We used the training set to train the four segmentation algorithms, set the early stopping method to train the networks, and used a learning rate of 0.0001 to iterate and

FIGURE 8. Images from the datasets, divided into three parts including faces with masks, faces and faces with virtual masks.

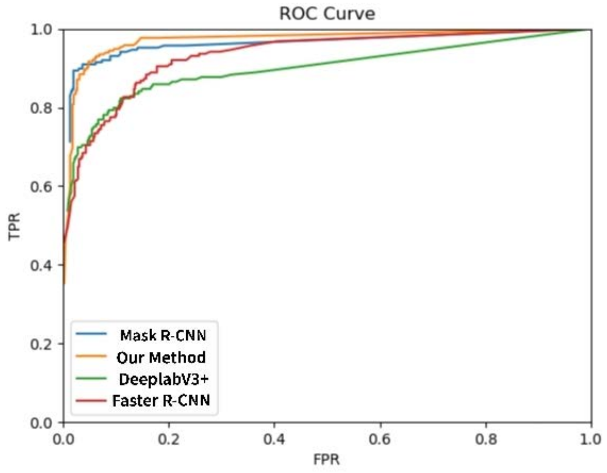


FIGURE 9. ROC curve, the face and mask detection accuracy of the improved Mask R-CNN segmentation model (yellow solid line) was close to that of Mask R-CNN before the improvement (blue solid line), which was more than 2% higher than that of Fast R-CNN (red solid line), and close to 95%.

TABLE 2. Comparison of mPA values of different segmentation networks.

Comparison of mPA values of different segmentation networks	mPA(%)	times(ms)
DeeplabV3+	81.5	195
RefineNet	80.2	281
Mask R-CNN	89.9	238
Our Method	**91.1**	**249**

train the models 500 times. Only prediction boxes with a confidence greater than 0.9 were retained after the maximum suppression method. The validation set was used to verify and calculate the mean pixel accuracy and average running time of the four segmentation algorithms. A comparison of the results is presented in Table 2 and Fig.10. Fig.10 shows the comparison of segmentation accuracy between IoU and Dice Coefficient [26].

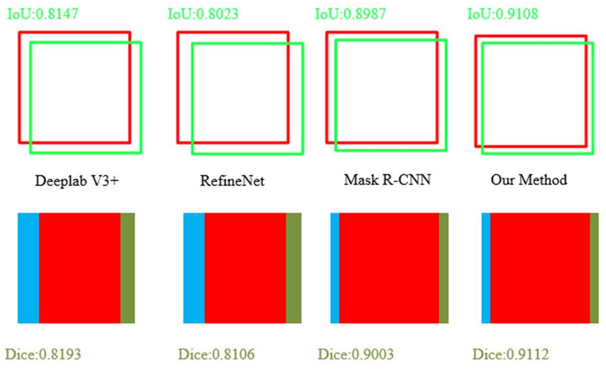


FIGURE 10. Comparison of IoU and dice of different segmentation networks.

TABLE 3. Edge segmentation ablation experiment.

Networks	PA		
	background	mask	face
Mask R-CNN+ResNet	94.39	90.02	86.65
Mask R-CNN +ResNeSt	94.88	91.06	88.27
Mask R-CNN + ResNeSt +edge segmentation	**95.42**	**94.51**	**89.39**

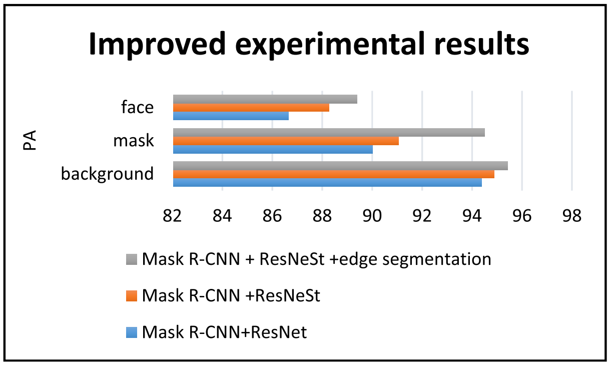


FIGURE 11. Improved experimental results.

2) EXPERIMENTAL ANALYSIS OF EDGE SEGMENTATION ABLATION

To verify the effectiveness of the proposed edge segmentation algorithm, ablation experiments were performed before and after the edge segmentation algorithm was added, as presented in Table 3 and Fig.11. Specifically, for the first two lines in Table 3, after incorporating the basic network of distributed attention mechanisms, the average segmentation accuracy improved by 1.02%, which benefited from the basic network through a distributed attention mechanism for mining local characteristics and the correlation of the local characteristics of different scale channels, to strengthen the characteristics of the different channel powers of expression. After the edge segmentation algorithm was added, finer segmentation was achieved through the re-classification and re-segmentation of edge pixels, and the segmentation accuracy was improved by 2.69% compared to the original instance segmentation.

C. RESULTS ANALYSIS

1) COMPARISION AND ANALYSIS OF DETECTION AND SEGMENTATION ACCURACY

As shown in Fig.9, the face and mask detection accuracy of the improved Mask R-CNN segmentation model

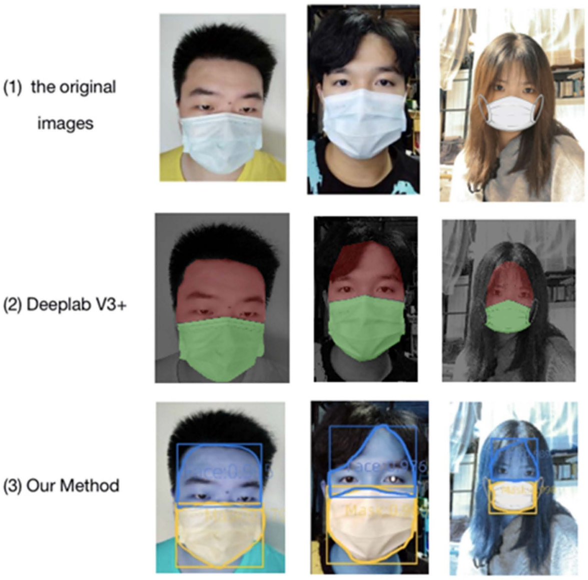


FIGURE 12. The segmentation result of partial images. The figures showed that the average segmentation accuracy of the face part was more than 98.6%, and the average segmentation accuracy of the mask part was more than 98.8%.

(yellow solid line) was close to that of Mask R-CNN before the improvement (blue solid line), which was more than 2% higher than that of Fast R-CNN (red solid line), and close to 95%, which benefits from the deep feature extraction of the target detection branch and the rapid convergence of EIoU. As shown in Fig.10, the running time of the DeeplabV3+ [27] model was the shortest owing to its relatively simple structure, whereas the method in this study is similar to Mask R-CNN in that its running time does not increase significantly because of the complexity of the network structure, which meets the requirements of real-time operation. In addition, the improved Mask R-CNN exhibited a slightly higher time but achieved a higher accuracy and obtained richer face information through segmentation after precise target positioning. This performance mainly benefited from the fact that the feature extraction network integrates the details of faces and masks of different feature layers after integrating the grouping attention mechanism, thus enriching the feature expression.

2) INSTANCE SEGMENTATION EFFECT COMPARISION

Fig.12 shows the segmentation performance of the proposed method and different network models for face profiles, multiple targets, and different types of masks. In DeeplabV3+, the red areas show the result of dividing images into masks, whereas the green areas show the result of dividing images into faces. In the MISA algorithm, to distinguish the face and mask regions of different targets more clearly, we used different fill colors, and the target boxes were marked with their degree of confidence. As shown in Fig.12, the size of the prediction box is closer to the inner rectangle of the real segmentation result after the addition of EIoU, and the box exhibited a higher target detection accuracy. The segmentation effect of the edge details of the MISA mask segmentation algorithm was optimized with the addition of the edge segmentation algorithm.

V. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a multi-step iterative segmentation algorithm in which the traditional backbone network of Mask R-CNN is replaced with ResNeSt to enhance the feature extraction capability of the basic network. We also used the improved EIoU as the loss function to enable the prediction box to regress to the target region more accurately and quickly. In addition, we adopted an occlusion classifier trained by R-PDSN to subdivide the occlusion edge blocks of the coarse segmentation results and optimized the edge information continuously for accurate segmentation, which improves the generalization ability of segmentation and the accuracy of face mask segmentation. The experimental results have shown that the proposed approach can achieve better results than other deep learning methods in terms of both generalization and accuracy.

In future work, we will consider extending the method not only to mask shielding but also to the broad category of random shielding. Further, we will consider segmented face features as the main feature source of face recognition models to avoid the problem of insufficient face information caused by various shielding types. We also consider the segmentation accuracy in other complex environments, such as the impact of the tilt of the camera on segmentation. Considering that mask detection is a real-time detection, our proposed method should also make more tests on real-time segmentation. At the same time, we considered the limitations of the dataset, so we considered expanding our dataset and testing it on some other common datasets. In the paper, we compared with some segmentation algorithms such as Deeplab V3+ and Faster RCNN, and in the future, we will consider comparing with other excellent algorithms such as SOLO and YOLACT.

VI. AUTHOR CONTRIBUTIONS

Min Zhang conceived the algorithms and designed the experiments, Kai Xie reviewed the paper, Yu-Hang Zhang conducted a comparative experiment on images and labeled the datasets, and Chang Wen checked the spelling and made suggestions.

APPENDIX

Datasets: 1050 images

Learning rate: 0.0001

Training times: 500

ACKNOWLEDGMENT

(Min Zhang and Kai Xie contributed equally to this work.)

REFERENCES

[1] K.-F. Li, J.-Y. Lu, and Q.-Z. Huang, "Face recognition based on sub-image dividing and multi-class support vector machine," School Elect. Inf. Eng., Henan Inst. Eng., Tech. Rep. TP391.41, 2018, doi: 10.13774/j.cnki.kjtb.2018.08.032.

[2] W. Wan and J. Chen, "Occlusion robust face recognition based on mask learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3795–3799.

[3] X.-X. Li, D.-Q. Dai, X.-F. Zhang, and C.-X. Ren, "Structured sparse error coding for face recognition with occlusion," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1889–1900, May 2013, doi: 10.1109/TIP.2013.2237920.

[4] S.-Q. Chang and Y.-C. Li, "A facial recognition method based on face segmentation," *Ind. Control Comput.*, vol. 31, no. 4, pp. 80–81, 2018, doi: 10.3969/j.issn.1001-182X.2018.04.033.

[5] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Faceness-Net: Face detection through deep facial part responses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1845–1859, Aug. 2018, doi: 10.1109/TPAMI.2017.2738644.

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.

[7] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[8] K. Lin, H. Zhao, J. Lv, C. Li, X. Liu, R. Chen, and R. Zhao, "Face detection and segmentation based on improved mask R-CNN," *Discrete Dyn. Nature Soc.*, vol. 2020, pp. 1–11, May 2020, doi: 10.1155/2020/9242917.

[9] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666, doi: 10.1109/CVPR.2019.00075.

[10] J. J. Liu, Q. Hou, M. M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3917–3926.

[11] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential Siamese network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 773–782.

[12] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "ResNeSt: Split-attention networks," 2020, *arXiv:2004.08955*.

[13] S. Wang, Y. Liu, Y. Qing, C. Wang, T. Lan, and R. Yao, "Detection of insulator defects with improved ResNeSt and region proposal network," *IEEE Access*, vol. 8, pp. 184841–184850, 2020, doi: 10.1109/ACCESS.2020.3029857.

[14] J. Shi, Y. Zhou, and W. J. X. Q. Zhang, "Target detection based on improved mask RCNN in service robot," *Chin. Assoc. Autom.*, pp. 8519–8524, 2019.

[15] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 386–397, Feb. 2020, doi: 10.1109/TPAMI.2018.2844175.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[17] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500, doi: 10.1109/CVPR.2017.634.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[19] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "UnitBox: An advanced object detection network," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 516–520, doi: 10.1145/2964284.2967274.

[20] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," 2021, *arXiv:2101.08158*.

[21] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, H. Chen, Y. Miao, Z. Huang, and J. Liang, "Masked face recognition dataset and application," 2020, *arXiv:2003.09093*.

[22] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 98–105, doi: 10.1109/FG.2018.00024.

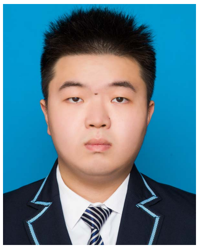
[23] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.

[24] I. Kalinovskii and V. Spitsyn, "Compact convolutional neural network cascade for face detection," 2015, *arXiv:1508.01292*.

[25] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818, doi: 10.1007/978-3-030-01234-2_49.

[26] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571, doi: 10.1109/3DV.2016.79.

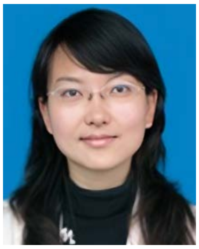
[27] S. C. Yurtkulu, Y. H. Sahin, and G. Unal, "Semantic segmentation with extended DeepLabv3 architecture," in *Proc. 27th Signal Process. Commun. Appl. Conf. (SIU)*, Apr. 2019, pp. 1–4, doi: 10.1109/SIU.2019.8806244.



MIN ZHANG was born in Jangsu, China, in 2000. He joined the National Demonstration Center for Experimental Electrical and Electronic Education, in 2019, with the intent to research deep learning and image processing. He has been conducting research projects on speech emotion recognition. He is currently an Assistant Researcher with Yangtze University, Jingzhou, China. His research interests include image processing and machine learning.

KAI XIE received the M.S. degree in electronic engineering from the National University of Defense Technology, Changsha, China, in 2003, and the Ph.D. degree in pattern recognition and intelligent system from Shanghai Jiao Tong University, Shanghai, China, in 2006. He is currently a Professor with the School of Electronic Information, Yangtze University, Jingzhou, China. He is also working in the field of image processing and signal processing.

YU-HANG ZHANG was born in Hubei, China, in 2001. He joined the National Demonstration Center for Experimental Electrical and Electronic Education, in 2020, with the intent to research deep learning and 3D visualization. He is currently an Assistant Researcher with Yangtze University, Jingzhou, China. He is committed to research in the laboratory, image processing, and other scientific research projects. His research interests include 3D visualization and artificial intelligence.



CHANG WEN received the B.S. degree in computer science from the Naval University of Engineering, Wuhan, China, in 2002, and the M.S. degree in computer science from Yangtze University, Jingzhou, China, in 2008. She is currently an Assistant Professor with the School of Computer Science, Yangtze University. She is also working in the field of image processing and signal processing.

JIAN-BIAO HE received the B.S. and M.S. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 1986 and 1989, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, Central South University. His research interests include artificial intelligence, the Internet of Things, pattern recognition, mobile robots, and cloud computing.