Hindawi
Discrete Dynamics in Nature and Society
Volume 2020, Article ID 9242917, 11 pages
https://doi.org/10.1155/2020/9242917

Research Article

Face Detection and Segmentation Based on Improved Mask R-CNN

Kaihan Lin, Huimin Zhao, Jujian Lv, Canyao Li, Xiaoyong Liu, Rongjun Chen, and Ruoyan Zhao

School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China

Correspondence should be addressed to Huimin Zhao; zhaohuimin@gpnu.edu.cn and Jujian Lv; jujianlv@gpnu.edu.cn

Received 18 December 2019; Accepted 11 March 2020; Published 1 May 2020

Guest Editor: Zheng Wang

Deep convolutional neural networks have been successfully applied to face detection recently. Despite making remarkable progress, most of the existing detection methods only localize each face using a bounding box, which cannot segment each face from the background image simultaneously. To overcome this drawback, we present a face detection and segmentation method based on improved Mask R-CNN, named G-Mask, which incorporates face detection and segmentation into one framework aiming to obtain more fine-grained information of face. Specifically, in this proposed method, ResNet-101 is utilized to extract features, RPN is used to generate RoIs, and RoIAlign faithfully preserves the exact spatial locations to generate binary mask through Fully Convolution Network (FCN). Furthermore, Generalized Intersection over Union (GIoU) is used as the bounding box loss function to improve the detection accuracy. Compared with Faster R-CNN, Mask R-CNN, and Multitask Cascade CNN, the proposed G-Mask method has achieved promising results on FDDB, AFW, and WIDER FACE benchmarks.

1. Introduction

Face detection is a key link of subsequent face-related applications, such as face recognition [1], facial expression recognition [2], and face hallucination [3], because its effect directly affects the subsequent applications performance. Therefore, face detection has become a research hotspot in the field of pattern recognition and computer vision and has been widely studied in the past two decades.

Large amounts of approaches have been proposed for face detection. The early research on face detection [4–9] mainly focused on the design of handcraft feature and used traditional machine learning algorithms to train effective classifiers for detection and recognition. Such approaches are limited in that the efficient feature design is complex and the detection accuracy is relatively low. In recent years, face detection methods based on deep convolutional neural network [10–13] have been widely studied, which are more robust and efficient than handcraft feature methods. Besides, a series of efficient object detection frameworks are used for face detection to improve detection performance [14–18], including R-CNN [19], Fast R-CNN [20], and Faster R-CNN [21]. These methods mainly implement face detection and the location of the face bounding box, which may have some drawbacks such as the extracted face features have background noise, spatial quantization is rough and cannot be accurately positioned. These drawbacks will directly affect the follow-up subsequent face-related applications, such as face recognition, facial expression recognition, and face alignment [22]. Therefore, it is necessary to study a face detection and segmentation method.

Mask R-CNN [23], an improved object detection model based on Faster R-CNN, has an impressive performance on various object detection and segmentation benchmarks such as COCO challenges [24] and Cityscapes dataset [25]. Unlike traditional R-CNN series methods, Mask R-CNN adds a mask branch for predicting segmentation masks on each Region of Interest (RoI), which can fulfil both detection and segmentation tasks. In order to fulfil both face detection and segmentation tasks from the image to overcome the drawbacks of the existing methods, a face detection and segmentation method based on improved Mask R-CNN (G-Mask) is proposed in this paper. In particular, our scheme introduces Generalized Intersection over Union (GIoU) [26] as the loss function for bounding box regression to improve detection accuracy of face detection. The main contributions of this paper are as follows:

(1) A new dataset was created (more details are described in Section 4.1), which annotated 5115 images randomly selected from the FDDB [27] and ChokePoint datasets [28].

(2) A face detection and segmentation method based on improved Mask R-CNN was proposed, which can detect faces correctly while also precisely segmenting each face in an image. Furthermore, the proposed method improves the detection performance by introducing GIoU as a bounding box loss function. The experimental results verify that our proposed G-Mask method achieves promising performance on several mainstream benchmarks, including the FDDB, AFW [29], and WIDER FACE [30].

The remainder of this paper is organized as follows. Section 2 briefly reviews the related work. The G-Mask framework for face detection and segmentation is described in detail in Section 3. Section 4 presents the experiment and discussion of the proposed method. In the last section, the work is summarized and the direction of future work is proposed.

2. Related Work

Face detection as one of the important research directions of computer vision has been extensively studied in recent years. From the development process of face detection, we can simply classify previous work as handcraft feature based and neural networks based methods.

2.1. Handcraft Feature Based Methods. With the appearance of the first real-time face detection method called Viola-Jones [4] in 2004, face detection has begun to be applied in practice. The well-known Viola-Jones can perform real-time detection using Haar feature and cascaded structure, but it also has some drawbacks, such as large feature size and low recognition rate for complex situations. To address these concerns, a lot of new handcraft features are proposed, such as HOG [5], SIFT [6], SUFT [7], and LBP [8], which have achieved outstanding results. Apart from the above methods, one of the significant advances was Deformable Part Model (DPM), proposed by Felzenszwalb et al. [9]. In the DPM model, the face is represented as a set of deformable parts, and the improved HOG feature and SVM are used for detection, achieving remarkable performance. In general, the advantages of handcraft features are that the model is intuitive and extensible, and the disadvantage is that the detection accuracy is limited in the face of multi-objective tasks.

2.2. Neural Networks Based Methods. As early as 1994, Vaillant et al. [10] first proposed using neural network to detect faces. In this work, Convolutional Neural Networks (CNN) is used to classify whether each pixel is part of a face and then determine the location of the face through another CNN. After that, the researchers did a lot of research based on this work. In recent years, the deep learning approaches has significantly promoted the development of the computer vision technology, including face detection. Li et al. [11] proposed a cascade CNN network architecture for rapid face detection, which is a multiresolution network structure that can quickly eliminate background regions in the low-resolution stage and carefully evaluate challenging candidates in the last high resolution stage. Ranjan et al. [12] proposed a deformation part model based on normalized features extracted by deep convolutional neural network. Yang et al. [13] proposed a method called Convolutional Channel Feature (CCF) by combining the advantages of both filtered channel features and CNN, which has a lower computational cost and storage cost than the general end-to-end CNN method.

Recently, witnessing the significant advancement of object detection using region-based methods, researchers have gradually applied the R-CNN series of methods to face detection. Qin et al. [14] proposed a joint training scheme for CNN cascade, Region Proposal Network (RPN), and Fast R-CNN. In [15], Jiang et al. trained the Faster R-CNN model by using WIDER dataset and verified performance on the FDDB and IJB-A benchmarks. Sun et al. [16] improve the Faster R-CNN framework through a series of strategies such as multiscale training, hard negative mining, and feature concatenation. Wu et al. [17] proposed a different scales face detection method based on Faster R-CNN for the challenge of small-scale face detection. Liu et al. [18] proposed a cascaded backbone branches fully convolutional neural network (BB-FCN) and used facial landmark localization results to guide R-CNN-based face detection. The neural networks based methods are already the mainstream of face detection because of its high efficiency and stability. In this work, we propose a G-Mask scheme, which achieves fairly progress in face detection task compared to the original architecture.

3. Improved Mask R-CNN

3.1. Network Architecture. The proposed method is extended from the Mask R-CNN [23] framework, which is the state-of-the-art object detection scheme and demonstrated impressive performance on various object detection benchmarks. As stated in Figure 1, the proposed G-Mask method consists of two branches, one for face detection and the other for face and background image segmentation. In this work, the ResNet-101 backbone is used to extract the facial features of the input image, and the Region of Interest (RoI) is rapidly generated on the feature map through the Region Proposal Network (RPN). We also use the Region of Interest Align (RoIAlign) to faithfully preserve exact spatial locations and output the feature map to a fixed size. At the end of the network, the bounding box is located and classified in the

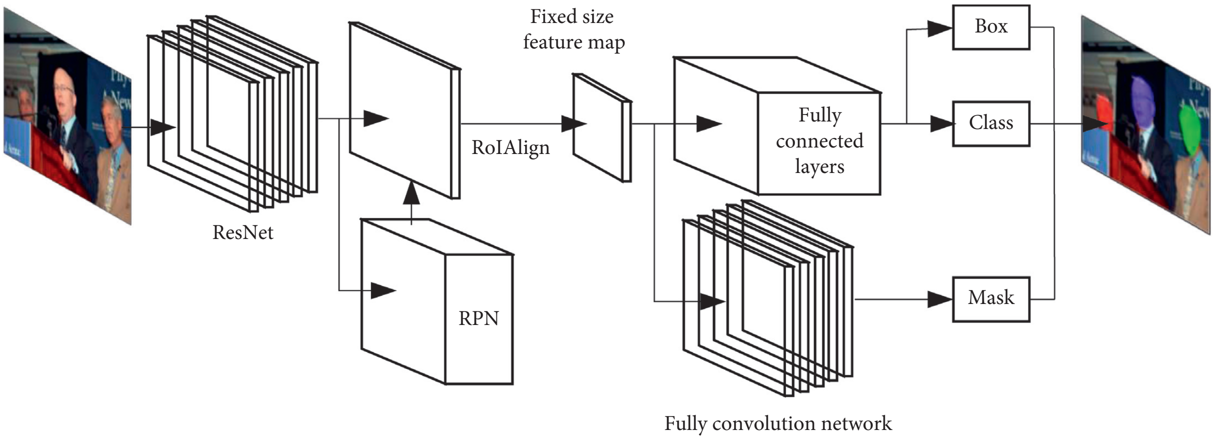


Figure 1: Network architecture of the G-Mask.

detection branch, and the corresponding face mask is generated on the image in the segmentation branch through the Fully Convolution Network (FCN) [31]. In the following, we will introduce the key steps of our network in detail.

3.2. Region Proposal Network. For images with human faces in our daily life, there are generally some face objects with different scales and aspect ratios. Therefore, in our approach, Region Proposal Network (RPN) generates RoIs by sliding windows on the feature map through anchors with different scales and different aspect ratios. Details are shown in Figure 2. The largest rectangle in the figure represents the feature map extracted by the convolutional neural network, and the dotted line indicates that the anchor is the standard anchor. Assume that the standard anchor size is 64 pixels, and the three anchors it contained represent three anchors with aspect ratios of 1 : 1, 1 : 2, and 2 : 1. The dot-dash line and the solid line represent the anchors of 32 and 128 pixels, respectively. Similarly, each of them also has three aspect ratios anchors. For traditional RPN, the above three scales and three aspect ratios are used to slide on the feature map to generate RoIs. In this paper, we use 5 scales (16^2, 32^2, 64^2, 128^2, and 256^2) and 3 aspect ratios (1 : 1, 1 : 2, and 2 : 1), leading to 15 anchors at each location, which was more effective in detecting objects of different scales.

3.3. RoIAlign Layer. G-Mask, unlike the general face detection methods, has a segmentation operation, which requires more refined spatial quantization for feature extraction. In the traditional region-based approaches, RoIPool is the standard operation for extracting small feature map from RoIs, which have two quantization operations that result in misalignments between the RoI and the extracted features. For traditional detection methods, this may not affect classification and localization, while for our approach, it has a great impact on prediction of pixel-accurate masks, as well as for small object detection.

In response to the above problem, we introduced the RoIAlign layer, following the scheme of [23]. As shown in Figure 3, suppose the feature map is divided into 2×2 bins.

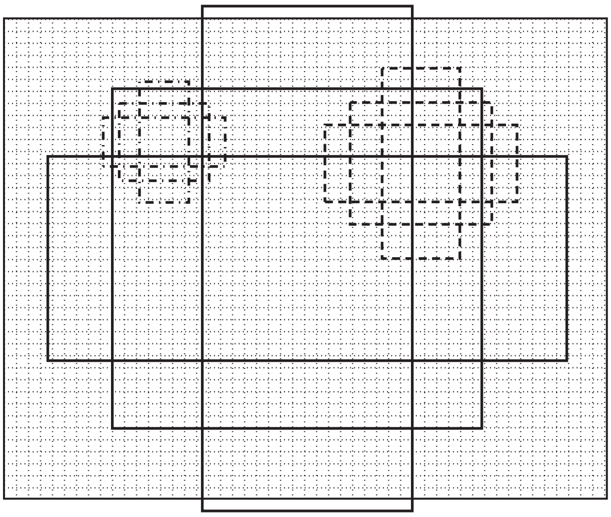


Figure 2: Illustration of RPN network.

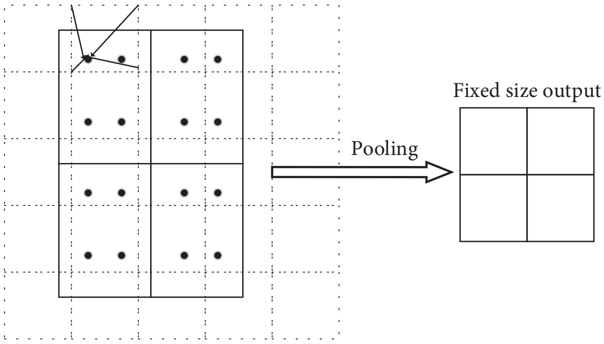


Figure 3: Bilinear interpolation in RoIAlign, where the dashed background grid represents the feature map, the solid grid represents an RoI (with 2×2 bins in this example), and the dots represent the four sample points in each bin.

It can be seen that the RoIAlign layer cancels the harsh quantization operations on the feature map and uses bilinear interpolation to preserve the floating-number coordinates, thereby avoiding misalignments between the RoI and the extracted features. The bilinear interpolation function has two steps, which are defined as follows:

Interpolate on the x-axis direction as follows:

$$f(R_1) \approx \frac{x_2 - x}{x_2 - x_1} f(Q_{11}) + \frac{x - x_1}{x_2 - x_1} f(Q_{21}), \quad R_1 = (x, y_1), \tag{1}$$

$$f(R_2) \approx \frac{x_2 - x}{x_2 - x_1} f(Q_{12}) + \frac{x - x_1}{x_2 - x_1} f(Q_{22}), \quad R_2 = (x, y_2). \tag{2}$$

Interpolate on the y-axis direction as follows:

$$f(P) = f(x, y) \approx \frac{y_2 - y}{y_2 - y_1} f(R_1) + \frac{y - y_1}{y_2 - y_1} f(R_2), \tag{3}$$

where $f(x, y)$ is the value of the sampling point P, $f(Q_{11})$, $f(Q_{12})$, $f(Q_{21})$, and $f(Q_{22})$ are the values of the four nearby grid points $Q_{11} = (x_1, y_1)$, $Q_{12} = (x_1, y_2)$, $Q_{21} = (x_2, y_1)$, and $Q_{22} = (x_2, y_2)$, and $f(R_1)$, $f(R_2)$ are the value obtained by interpolating in the x-axis direction.

3.4. Mask Branch. The mask branch realizes the segmentation of face object and background image in G-Mask model, which predicts the segmentation mask in a pixel to pixel manner by applying Full Convolutional Network (FCN) [31] to each RoI. The FCN scheme is one of the solutions for instance segmentation, which originates from CNN but is also different from general CNN. For the traditional CNN network architecture, in order to obtain the feature vector of fixed dimensions, the convolutional layer is generally connected with several full connection layers, and finally the output is a numerical description of the input, which is generally applicable to tasks such as image recognition and classification, object detection, and positioning. The FCN framework is similar to the traditional CNN network, which also includes the convolutional layer and the pooling layer. In particular, the FCN uses the deconvolution to up-sample the feature map in the end convolution layer so that the output image size can be restored to the original image size, and finally uses the Softmax classifier to predict the category of each pixel.

3.5. Generalized Intersection over Union. Bounding box regression, as one of the fundamental components of many computer vision tasks, deserves further study by researchers [32]. However, unlike the architecture and feature extraction strategy improvement researches, which have made great progress in recent years [33], the research of bounding box regression has lagged behind somewhat. The Generalized Intersection over Union (GIoU) [26], as the latest metric and bounding box regression method, demonstrates state-of-the-art results on various object detection benchmarks by incorporating with the general object detection frameworks. For traditional IoU, there are two weaknesses when it is used as a metric or a bounding box regression loss: (a) the IoU value is zero when two objects do not overlap, making it difficult to optimize the nonoverlapping bounding boxes; (b) the IoU value may be the same when two objects intersect in different orientations, so the IoU function does not reflect how the two objects overlap. To overcome these drawbacks, GIoU not only focuses on the situation where two objects overlap but also considers the situation of nonoverlapping. The details of the GIoU metric are shown in Figure 4. Suppose $B_p = (x_1^p, y_1^p, x_2^p, y_2^p)$ and $B_g = (x_1^g, y_1^g, x_2^g, y_2^g)$ are the coordinates of an object's predicted bounding box and the ground-truth bounding box, where $x_2 > x_1$ and $y_2 > y_1$ in B_P and B_g; then, the area of them is

$$A_p = \left(x_2^p - x_1^p\right) \times \left(y_2^p - y_1^p\right), \tag{4}$$

$$A_g = \left(x_2^g - x_1^g\right) \times \left(y_2^g - y_1^g\right). \tag{5}$$

The coordinates and area of intersection I of B_P and B_g can be calculated as

$$\begin{aligned} x_1^i &= \max\left(x_1^p, x_1^g\right), \\ x_2^i &= \min\left(x_2^p, x_2^g\right), \end{aligned} \tag{6}$$

$$\begin{aligned} y_1^i &= \max\left(y_1^p, y_1^g\right), \\ y_2^i &= \min\left(y_2^p, y_2^g\right), \end{aligned} \tag{7}$$

$$A_i = \begin{cases} \left(x_2^i - x_1^i\right) \times \left(y_2^i - y_1^i\right), & \text{if } x_2^i > x_1^i,\ y_2^i > y_1^i, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Similarly, the smallest enclosing box B_c can be found through

$$\begin{aligned} x_1^c &= \min\left(x_1^p, x_1^g\right), \\ x_2^c &= \max\left(x_2^p, x_2^g\right), \end{aligned} \tag{9}$$

$$\begin{aligned} y_1^c &= \min\left(y_1^p, y_1^g\right), \\ y_2^c &= \max\left(y_2^p, y_2^g\right), \end{aligned} \tag{10}$$

and the area of B_c can be computed as

$$A_c = \left(x_2^c - x_1^c\right) \times \left(y_2^c - y_1^c\right). \tag{11}$$

The IoU between B_P and B_g is defined as

$$\text{IoU} = \frac{A_i}{A_p + A_g - A_i}. \tag{12}$$

Therefore, GIoU can be calculated by the definition of

$$\text{GIoU} = \text{IoU} - \frac{A_c - \left(A_p + A_g - A_i\right)}{A_c}. \tag{13}$$

3.6. Loss Function. The proposed G-Mask model consists of two stages, which are the same as the general region-based model. In the first stage, RPN proposes the candidate bounding boxes of the object face. The second stage, follow the Fast R-CNN architecture, extracts features from each candidate box and then performs classification and

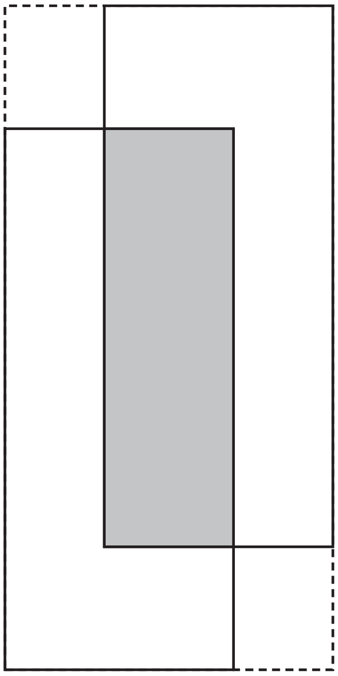


Figure 4: Illustration of GIoU metric. The solid line indicates the prediction box and ground truth box, the dotted line indicates the smallest enclosing box, and the shaded portion indicates the intersection of the prediction box and the ground truth box.

bounding box location. In addition, like the Mask R-CNN, we added a mask branch parallel to the classification branch and the bounding box location branch. Therefore, we define a multitasking objective function, which includes classification loss L_{cls}, bounding box location loss L_{box}, and segmentation loss L_{mask}. Our loss function for each image is defined as

$$L = L^*_{\mathrm{cls}} + L^*_{\mathrm{box}} + L^*_{\mathrm{mask}}. \quad (14)$$

In (14), the classification loss L_{cls} and segmentation loss L_{mask} are defined the same as in Mask R-CNN. For the bounding box loss, we found that GIoU can better respond to face detection tasks through several experiments compared with the traditional bounding box regression method. Therefore, in this paper, we introduced GIoU as a bounding box loss function. In more detail, the classification loss is defined as in

$$L^*_{\mathrm{cls}} = (\{p_i\}) = \frac{1}{N_{\mathrm{cls}}} \sum_i L_{\mathrm{cls}}(p_i, p_i^*), \quad (15)$$

where N_{cls} is the minibatch size, i is the index of an anchor in a minibatch, and p_i is the prediction probability of whether anchor i is a face target. The ground-truth label $p_i^* = 1$ if the anchor is positive, and $p_i^* = 0$ when the anchor is negative. The classification loss L_{cls} of each anchor is log loss of whether an object is a face, which is defined as

$$L_{\mathrm{cls}}(p_i, p_i^*) = -[p_i^* \log p_i + (1 - p_i^*) \log(1 - p_i)]. \quad (16)$$

For bounding box loss, we introduce GIoU as the loss function, and the definition of GIoU metric is described in (13), so the loss bounding box function is defined as follows:

$$L^*_{\mathrm{box}} = 1 - \mathrm{GIoU}. \quad (17)$$

For segmentation box loss, we adopt the average binary cross-entropy loss, which is defined in

$$L^*_{\mathrm{mask}} = -\frac{1}{m^2} \sum_{1 \le i,j \le m} \left[y_{ij} \log \hat{y}^k_{ij} + \left(1 - y_{ij}\right) \log\left(1 - \hat{y}^k_{ij}\right) \right], \quad (18)$$

where y_{ij} is the label value of a cell (i, j) for the region of size $m \times m$ and $\hat{y}^k_{ij}$ is the predicted value of the k-th class of this cell. L^*_{mask} is only defined on a specific mask, which is related to the ground-truth class k, and other mask outputs do not affect the loss.

4. Experiments

4.1. Experimental Setup. Unlike object detection and generic face detection, there are no off-the-shelf face datasets with masks annotation that can be employed to train our model [34]. Therefore, the first step of our work is to create a new dataset with mask annotations. In order to enhance the reliability of the samples, we selected 5115 samples from FDDB and ChokePoint datasets and annotated them with masks labels. After the annotation work, we trained the G-Mask model on this dataset.

For implementation, we adopt Keras [35] framework to train the G-Mask model in Ubuntu 16.04. ResNet-101 [36] is used as the backbone network architecture in our work. In the training phase, the G-Mask model is train on aforementioned dataset for 150,000 iterations (where the epoch is 50 and the steps of per epoch are 3000) with the learning rate set to 0.001 and the weight decay rate set to 0.0001. We randomly sample one image per batch for training [37], in which the short side of each image was resized to 800 and the long side was resized to 1024. In the RPN part, RoIs is generated by sliding the window on the feature map through anchors of different scales and different aspect ratios. It will have 2000 RoIs kept after nonmaximum suppression, and the RoIs will only be considered as foreground if its IoU with the ground truth is greater than 0.5. The testing phase settings are the same as the training phase, and the region proposal is considered to be a face only if the confidence score is greater than 0.7. The training and testing process is carried out on the same server, which is a Xeon E5 CPU of 128 GB flash memory and NVIDIA GeForce GTX 1080Ti GPU.

4.2. Experimental Results. In this work, G-Mask model not only realized the bounding box localization of the face target but also separated the face information from the background image by binary mask, so that more detailed face information could be obtained through the above process. The comparison experiment was carried out on three popular face benchmark datasets, including FDDB, AFW, and WIDER FACE.

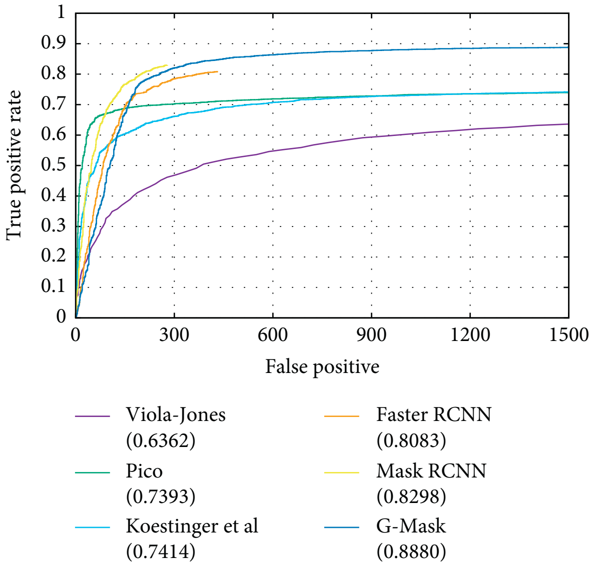


Figure 5: Comparisons of face detection with other methods on FDDB benchmark.

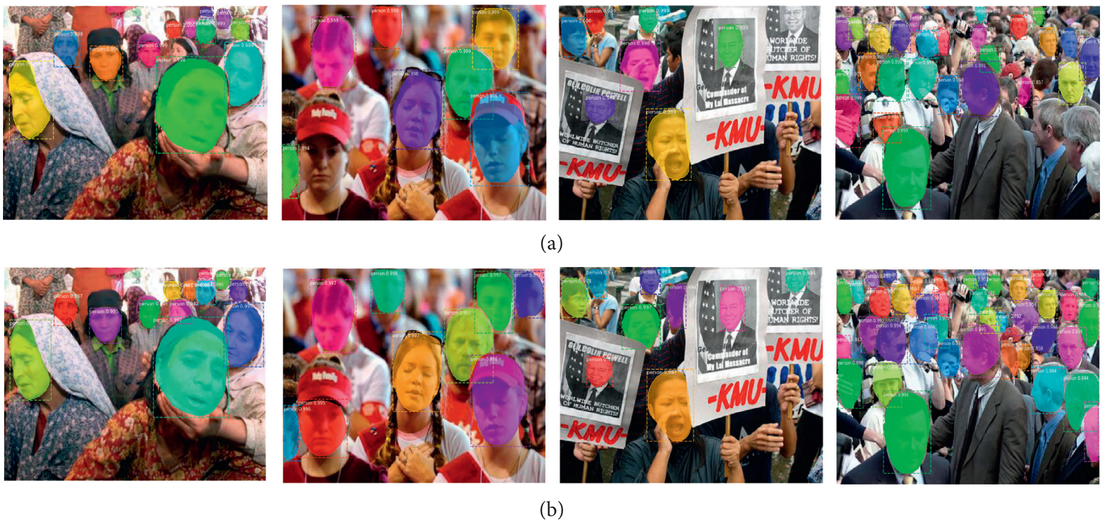


Figure 6: Different detection results of Mask R-CNN and G-Mask in the complex scene of FDDB dataset. (a) Mask R-CNN model and (b) G-Mask model.

The FDDB [27] dataset is a well-known face detection evaluation dataset and benchmark, which contains 2845 images of 5171 human faces. In this dataset, the faces of each image come from different scenes, which is quite challenging. We compared several methods on the FDDB dataset, including Faster R-CNN [15], Mask R-CNN [23], Pico [38], Viola-Jones [39], and Koestinger [40]. For effective comparison, the training data of the G-Mask, Mask R-CNN, and Faster R-CNN models are the same, which is the dataset constructed in this work. We compared the true positive rates at 1500 false positives, and the results are shown in Figure 5. It can be seen from Figure 5 that G-Mask performs better than Faster R-CNN when there are more than 160 false positives. When there are more than 280 false positives, the performance of G-Mask is better than that of Mask R-CNN. Furthermore, our method can achieve 88.80% true positive rate in 1500 false positives, which exceeded all the comparison methods. The comparison results of the FDDB dataset show that our proposed G-Mask method has achieved promising results, demonstrating that our method can segment face information while detecting effectively. Some detection results of the Mask R-CNN and G-Mask models in the complex scenario of FDDB dataset are shown in Figure 6. It is obvious that

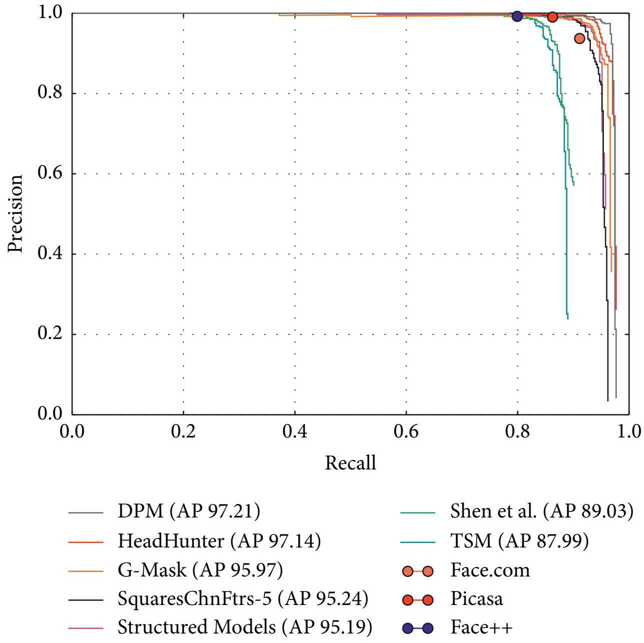
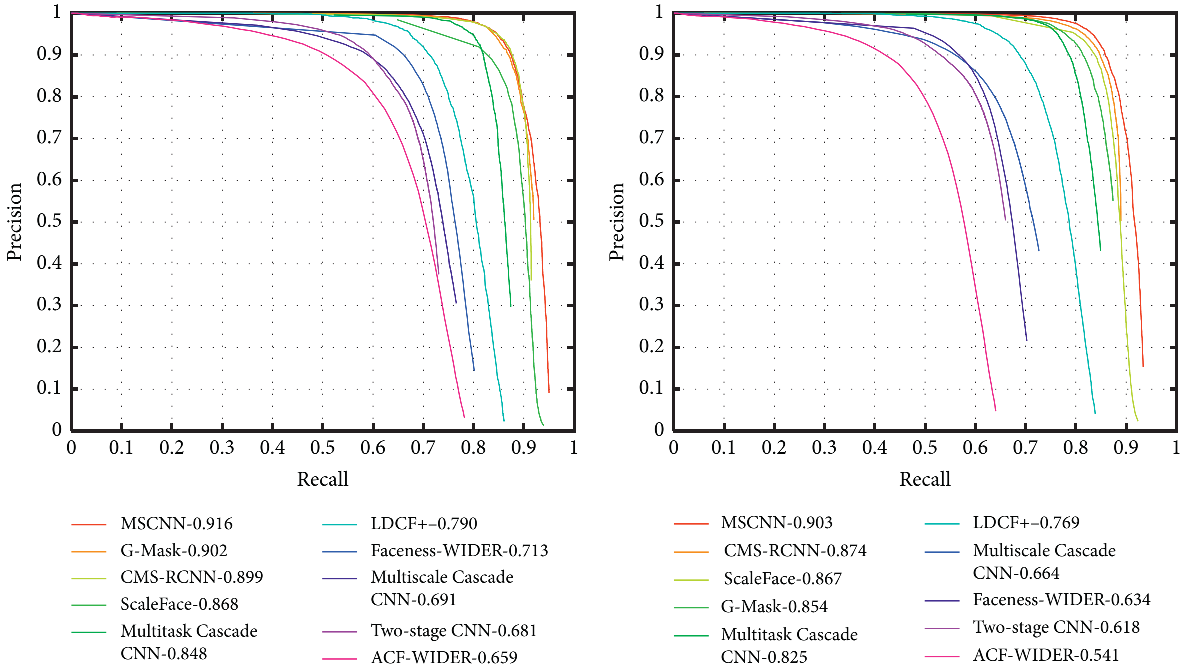


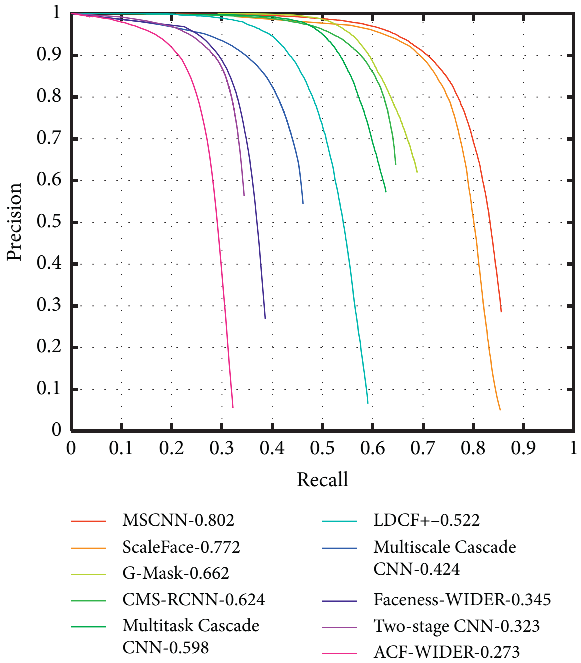
Figure 7: The precision-recall curve of our method on the AFW benchmark. Data of other models and evaluation code are derived from [41].



(a)

(b)

Figure 8: Continued.



(c)

Figure 8: The precision-recall curve on the WIDER FACE benchmark: (a) on the easy subset, (b) on the medium subset, and (c) on the hard subset.

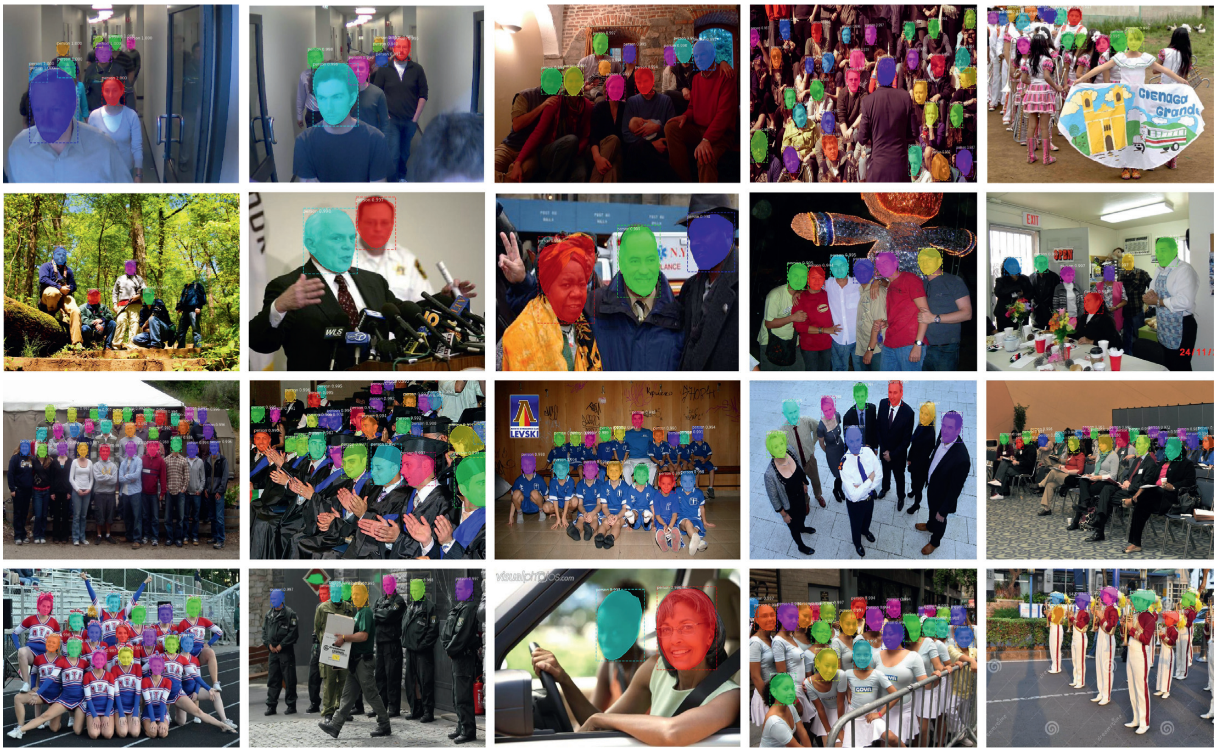


Figure 9: More results of G-Mask method.

the G-Mask model performs better in the multiscale face task, which demonstrates the effectiveness of the proposed method in face detection.

The AFW dataset [29] is a face dataset and benchmark established by using Flickr image, which contains 205 images with 473 labeled faces. The precision-recall curve of our method

Table 1: Running time of different region-based methods.

Method	Running time (s)		
	FDDB	AFW	ChokePoint
R-CNN	14.75	15.32	14.51
Fast R-CNN	3.12	3.08	2.84
Faster R-CNN	**0.30**	**0.32**	**0.28**
Mask R-CNN	0.32	0.35	0.33
G-Mask	0.35	0.42	0.33

on the AFW benchmark is shown in Figure 7, and it can be seen that the G-Mask method achieved 95.97% average precision (AP). Although our dataset has a different label format from the AFW benchmark, as well as the moderately sized training dataset, we also demonstrate the generalization of our method.

WIDER FACE [30], one of the largest and most challenging face detection datasets in the open source data, has 32,203 images and 393,703 labeled faces. In this dataset, various changes in the face size, pose, and occlusion have brought great challenges to face detection, and the dataset is divided into easy, medium, and hard subsets according to the difficulty level. To further demonstrate the detection performance of our proposed method, we trained the G-Mask model on WIDER FACE dataset and verified it on the validation dataset. The proposed method is compared with several major methods including MSCNN [42], CMS-RCNN [43], ScaleFace [44], Multitask Cascade CNN [45], and Faceness-WIDER [46]. The precision-recall curves of G-Mask method on the WIDER FACE benchmark are shown in Figure 8. It can be seen that our method obtained 0.902 AP in the easy subset, 0.854 AP in the medium subset, and 0.662 AP in the hard subset, which exceeded most of the comparison methods. Compared with the state-of-the-art MSCNN method, the AP value of the proposed method is only 0.014 lower in the easy subset and 0.049 lower in the medium subset. There are some gaps between G-Mask and MSCNN methods on hard subset. The reason may be that the MSCNN method uses a series of strategies for small-scale faces detection and thus they can deal with more challenging cases. Nevertheless, the G-Mask method still achieves promising performance, which demonstrates the effectiveness of the G-Mask method.

We further demonstrate more qualitative results of G-Mask method in Figure 9. It can be observed that the proposed method can detect faces correctly while also precisely segmenting each face in an image.

We also compared the running time of different region-based methods in the a series of dataset such as FDDB, AFW, and ChokePoint. The WIDER FACE dataset was not used for testing because the running time of the hard and easy subset on the WIDER FACE was quite different. We randomly selected 100 images from each of the above datasets to test and calculate their average time, and the results are reported in Table 1. We can clearly see that Faster R-CNN has the shortest running time because of its relatively simple structure, while the proposed method has a running time similar to Mask R-CNN. Compared with Faster RCNN method, G-Mask adds a segmentation branch, which leads to an increase in computational complexity. However, the G-Mask method can achieve higher accuracy with less time consumption compared with other region-based methods and can also obtain more detailed face information through segmentation branches while accurately locating.

5. Conclusions

In this paper, a G-Mask method was proposed for face detection and segmentation. The approach can extract features by ResNet-101, generate RoIs by RPN, preserve the precise spatial position by RoIAlign, and generate binary masks through the full convolutional network (FCN). In doing so, the proposed framework is able to detect faces correctly while also precisely segmenting each face in an image. Experimental results with self-built face dataset as well as public available datasets have verified that our proposed G-Mask method achieves promising performance. For the future work, we will consider improving the speed of the proposed method.

Data Availability

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

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partly supported by Innovation Team Project of the Education Department of Guangdong Province (2017KCXTD021), Key Laboratory of the Education Department of Guangdong Province (2019KSYS009), Foundation for Youth Innovation Talents in Higher Education of Guangdong Province (2018KQNCX139), Project for Distinctive Innovation of Ordinary Universities of Guangdong Province (2018KTSCX120), and the Ph.D. Start-Up Fund of Natural Science Foundation of Guangdong Province (2016A030310335).

References

[1] J. Deng, J. Guo, N. Xue et al., “Arcface: additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, Long Beach, CA, USA, June 2019.

[2] N. Zeng, H. Zhang, B. Song et al., “Facial expression recognition via learning deep sparse autoencoders,” *Neurocomputing*, vol. 273, pp. 4690–4699, 2018.

[3] Y. Shi, L. I. Guanbin, Q. Cao et al., “Face hallucination by attentive sequence optimization with reinforcement learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[4] P. Viola and M. J. Jones, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 886–893, San Diego, CA, USA, June 2005.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[8] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester et al., "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2009.

[10] R. Vaillant, C. Monrocq, and Y. Le Cun, "Original approach for the localisation of objects in images," *IEEE Proceedings—Vision, Image, and Signal Processing*, vol. 141, no. 4, pp. 245–250, 1994.

[11] H. Li, Z. Lin, X. Shen et al., "A convolutional neural network cascade for face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5325–5334, Boston, MA, USA, June 2015.

[12] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2017.

[13] B. Yang, J. Yan, Z. Lei et al., "Convolutional channel features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 82–90, Boston, MA, USA, June 2015.

[14] H. Qin, J. Yan, X. Li et al., "Joint training of cascaded CNN for face detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3456–3465, Las Vegas, NV, USA, July 2016.

[15] H. Jiang and E. Learned-Miller, "Face detection with the faster R-CNN," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 650–657, Washington, DC, USA, June 2017.

[16] X. Sun, P. Wu, and S. C. H. Hoi, "Face detection using deep learning: an improved faster RCNN approach," *Neurocomputing*, vol. 299, no. 1, pp. 42–50, 2018.

[17] W. Wu, Y. Yin, X. Wang, and D. Xu, "Face detection with different scales based on faster R-CNN," *IEEE Transactions on Cybernetics*, vol. 49, no. 11, pp. 4017–4028, 2019.

[18] L. Liu, G. Li, Y. Xie et al., "Facial landmark machines: a backbone-branches architecture with progressive representation learning," *IEEE Transactions on Multimedia*, vol. 21, no. 9, 2019.

[19] R. Girshick, J. Donahue, T. Darrell et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, Columbus, OH, USA, June 2014.

[20] R. Girshick, "Fast r-CNN," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1440–1448, Boston, MA, USA, June 2015.

[21] S. Ren, K. He, R. Girshick et al., "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. 91–99, 2015.

[22] W. Wu, C. Qian, S. Yang et al., "Look at boundary: a boundary-aware face alignment algorithm,," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2129–2138, Salt Lake City, UT, USA, June 2018.

[23] K. He, G. Gkioxari, P. Dollár et al., "Mask r-CNN," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2961–2969, Honolulu, HI, USA, July 2017.

[24] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: common objects in context," *Computer Vision—ECCV 2014*, Springer, Berlin, Germany, pp. 740–755, 2014.

[25] M. Cordts, M. Omran, S. Ramos et al., "The cityscapes dataset for semantic urban scene understanding,," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, Las Vegas, NV, USA, July 2016.

[26] H. Rezatofighi, N. Tsoi, J. Y. Gwak et al., "Generalized intersection over union: a metric and a loss for bounding box regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 658–666, Long Beach, CA, USA, June 2019.

[27] V. Jain and E. Learned-Miller, "FDDB: a benchmark for facedetection in unconstrained settings," Technical report UM-CS-2010-009, 2010.

[28] Y. Wong, S. Chen, S. Mau et al., "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 74–81, Colorado Springs, CO, USA, June 2011.

[29] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2879–2886, Providence, RI, USA, June 2012.

[30] S. Yang, P. Luo, C. C. Loy et al., "Wider face: a face detection benchmark,," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5525–5533, Las Vegas, NV, USA, June 2016.

[31] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, Boston, MA, USA, June 2015.

[32] J. Ren, A. Hussain, J. Han, and X. Jia, "Cognitive modelling and learning for multimedia mining and understanding," *Cognitive Computation*, vol. 11, no. 6, pp. 761-762, 2019.

[33] J. Tschannerl, J. Ren, P. Yuen et al., "MIMR-DGSA: unsupervised hyperspectral band selection based on information theory and a modified discrete gravitational search algorithm," *Information Fusion*, vol. 51, pp. 189–200, 2019.

[34] K. Lin, H. Zhao, J. Lv et al., "Face detection and segmentation with generalized intersection over union based on mask R-CNN," in *Proceedings of the International Conference On Brain Inspired Cognitive Systems*, pp. 106–116, Guangzhou, China, July 2019.

[35] F. Chollet, "Keras, github repository," 2015, https://github.com/fchollet/keras.

[36] K. He, X. Zhang, S. Ren et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, July 2016.

[37] P. Wan, C. Wu, Y. Lin et al., "Driving anger states detection based on incremental association markov blanket and least

square support vector machine," *Discrete Dynamics in Nature and Society*, vol. 2019, Article ID 2745381, 17 pages, 2019.

[38] N. Markuš, M. Frljak, I. S. Pandzic et al., "A method for object detection based on pixel intensity comparisons organized in decision trees," 2013, https://arxiv.org/abs/1305.4537.

[39] D. Hefenbrock, J. Oberg, N. T. N. Thanh et al., "Accelerating viola-jones face detection to Fpga-level using gpus," in *Proceedings of the IEEE Annual International Symposium on Field-Programmable Custom Computing Machines*, pp. 11–18, Charlotte, NC, USA, May 2010.

[40] M. Köstinger, P. Wohlhart, P. M. Roth et al., "Robust face detection by simple means," in *Proceedings of the DAGM 2012 CVAW Workshop*, Graz, Austria, August 2012.

[41] M. Mathias, R. Benenson, M. Pedersoli et al., "Face detection without bells and whistles," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 720–735, Zurich, Switzerland, September 2014.

[42] Z. Cai, Q. Fan, R. S. Feris et al., "A unified multi-scale deep convolutional neural network for fast object detection," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 354–370, Amsterdam, Netherlands, October 2016.

[43] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "Cms-rcnn: contextual multi-scale region-based cnn for unconstrained face detection," *Deep Learning for Biometrics*, Springer, Berlin, Germany, pp. 57–79, 2017.

[44] S. Yang, Y. Xiong, C. C. Loy et al., "Face detection through scale-friendly deep convolutional networks," 2017, https://arxiv.org/abs/1706.02863.

[45] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[46] S. Yang, P. Luo, C. C. Loy et al., "Faceness-net: face detection through deep facial part responses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1845–1859, 2017.