Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 8586016, 8 pages
https://doi.org/10.1155/2021/8586016

Research Article

Masked Face Detection Algorithm in the Dense Crowd Based on Federated Learning

Rui Zhu,[1] **Kangning Yin**,[2] **Hang Xiong**,[1] **Hailian Tang**,[2] **and Guangqiang Yin**[2]

[1]*School of Information and Communication Engineering, University of Electronic Science and Technology of China, Sichuan 611731, China*
[2]*School of Information and Software Engineering, University of Electronic Science and Technology of China, Sichuan 611731, China*

Correspondence should be addressed to Guangqiang Yin; yingq@uestc.edu.cn

Received 20 July 2021; Accepted 15 September 2021; Published 4 October 2021

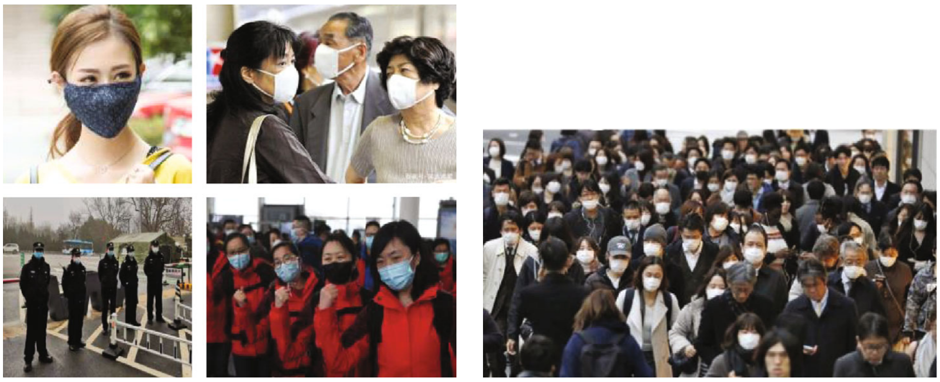
Academic Editor: Yan Huang

Wearing masks is an effective and simple method to prevent the spread of the COVID-19 pandemic in public places, such as train stations, classrooms, and streets. It is of positive significance to urge people to wear masks with computer vision technology. However, the existing detection methods are mainly for simple scenes, and facial missing detection is prone to occur in dense crowds with different scales and occlusions. Moreover, the data obtained by surveillance cameras in public places are difficult to be collected for centralized training, due to the privacy of individuals. In order to solve these problems, a cascaded network is proposed: the first level is the Dilation RetinaNet Face Location (DRFL) Network, which contains Enhanced Receptive Field Context (ERFC) module with the dilation convolution, aiming to reduce network parameters and locate faces of different scales. In order to adapt to embedded camera devices, the second level is the SRNet20 network, which is created by Neural Architecture Search (NAS). Due to privacy protection, it is difficult for surveillance video to share in practice, so our SRNet20 network is trained in federated learning. Meanwhile, we have made a masked face dataset containing about 20,000 images. Finally, the experiments highlight that the detection mAP of the face location is 90.6% on the Wider Face dataset, and the classification mAP of the masked face classification is 98.5% on the dataset we made, which means our cascaded network can detect masked faces in dense crowd scenes well.

1. Introduction

COVID-19 spreads rapidly among the population and has a serious impact on society, economy, and people's normal lives. The weekly epidemiological update of the World Health Organization (WHO) [1] presented that the cumulative number of cases reported globally is now over 186 million, and the number of deaths exceeds 4 million. Fortunately, wearing masks is an effective and simple method to prevent the spread of COVID-19 [2], and almost everyone is obligated to wear a face mask in public places. Relying solely on manpower for inspections inevitably has disadvantages, such as high work intensity, low efficiency, and timeliness, but using detection algorithms to complete this task can save many human resources. Using computer vision technology to detect whether people wear masks and to give corresponding reminders can achieve the purpose of noncontact detection, preventing the spread of the virus and ensuring people's safety.

Moreover, most of the existing algorithms train the model by collecting the data together, but the reality is that videos captured by the cameras in public places will not be easily obtained because of personal privacy [3]. The surveillance video data of public place belong to different departments, which make the data form an isolated island and difficult to be concentrated together for model training. As a new distributed machine learning method, federated learning, with the help of the storage and computing capacity of the device itself, can cobuild the model without data out of the local, so as to protect data privacy and effectively solve the problem of data island [4].



(a) Examples in public dataset

(b) Expected dataset

Figure 1: Comparing images.

Therefore, the task is decomposed into two subnetworks. The first network is used for the general face location, and the second is used for the masked face classification. The main contributions of our paper are listed below:

(1) The DRFL network is proposed and trained on the Wilder Face dataset to locate faces in dense crowds

(2) The SRNet20 network is designed with NAS and trained by methods of federated learning to classify masked faces

(3) A masked face dataset is created and contains 18,000 images in the train set and 1,751 images in the test set. In order to facilitate other researchers, this dataset is also published on the GitHub: https://github.com/woshizr/masked-Face

2. Related Work

2.1. Face Detection Algorithms. Face detection is closely related to general object detection. In recent years, object detection algorithms have developed rapidly, which are mainly divided into two categories: single stage object detection algorithms, represented by YOLO [5] and RetinaNet [6], divide the image into regions and predict bounding boxes and probabilities for each region simultaneously. Therefore, this kind of algorithm is faster. The two-stage object detection algorithms, represented by RCNN [7] and FPN [8], generate a large number of proposal regions, which then classify the proposals into foreground classes or background. Therefore, the accuracy of this kind of algorithm is higher. Based on the object detection algorithms, a large number of face detection algorithms and masked face detection algorithms have been developed: MTCNN [9] uses 3 cascaded networks to achieve face detection; Face RCNN [10] is based on Faster RCNN [11] for face detection; SSH [12] enhances the feature extraction of convolutional layers with different depths to achieve multiscale face detection; PyramidBox [13] uses the context information of the face to improve the detection of occluded faces; Didi company proposes a mask wearing detection algorithm based on DFS [14], the algorithm detects the face region first, expands the face area based on the face features, and then uses the attention mechanism to find the mask area, and finally detects whether the face is wearing a mask; AIZOO proposes a lightweight mask wearing detection algorithm [15] based on SSD and improves the network structure; RetinaMask [16] detects the face with mask by adding attention mechanism in context module.

Many efforts have also been made in society to help with masked face detection. In [17], three kinds of masked face datasets are proposed, including masked face detection dataset (MFDD), real-world masked face recognition dataset (RMFRD), and simulated masked face recognition dataset (SMFRD). Among them, RMFRD is currently the world's largest real-world masked face dataset, which provides the correct masked face dataset (CMFD) and the incorrectly masked face dataset (IMFD), and some sample images are shown in Figure 1(a); however, the dataset in dense scene is often shown as Figure 1(b). Therefore, the performance of the algorithm in Figure 1(b) can better illustrate the advantages and disadvantages of the algorithm.

2.2. Federated Learning. The development of artificial intelligence technology has encountered two main challenges: one is that data exists in the form of data islands in most industries; the other is that training models require a lot of data, and improper collection of data will make it difficult to protect the privacy and security of data. In the traditional centralized machine learning method, the data collected from different devices need to be uploaded to the cloud [18], and the central server in the cloud uses the data to train the model, as shown in Figure 2. Data are directly exposed in the cloud, which is difficult to protect user privacy [19].

To solve the above problems, in 2016, Google proposed federated learning [20], a machine learning framework based on user privacy protection. Their main idea is to build machine learning models based on data distributed on multiple devices and prevent user privacy from being leaked. Federated learning allows the device to use local data to train the model, after the training, the local device does not need to send sensitive data to the cloud, but only needs to upload the model parameters [21]. The central server of federated learning then aggregates the collected model parameters,

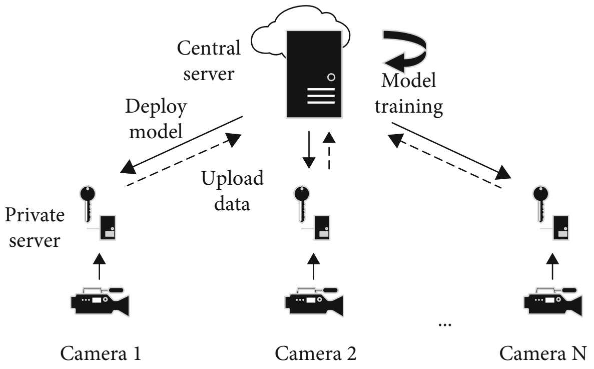


Figure 2: Centralized training.

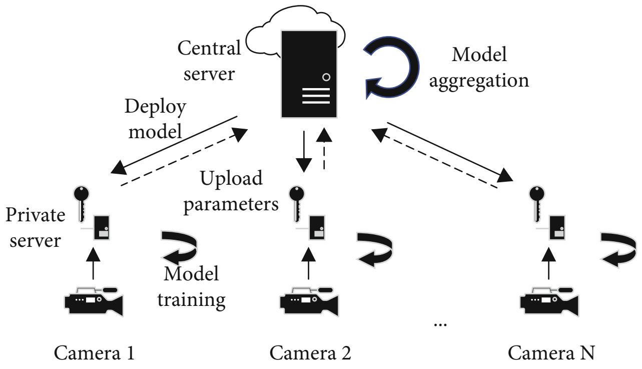


Figure 3: Federated training.

and this process continues until the joint training models reach the expected accuracy, as shown in Figure 3.

The data of the provider are kept locally, and the leakage of data privacy is suppressed from the source. Of course, federated learning also involves many aspects; in this paper, it mainly involves the use of multiparty data sources for federated training.

3. Models and Improvements

The whole algorithm is divided into two cascaded subnetworks: a general face location network and a face classification network with masks. The algorithm process is shown in Figure 4.

All face boxes are found in the input image through the face location network, and then whether each face box is wearing a mask is determined through the classification network. Especially, a federated training method is used to keep the data locally, and only the model parameters are transferred between clients, when training the classification network.

3.1. Dilation RetinaNet Face Location Network. The DRFL network is inspired by RetinaNet. In order to solve the problems of occlusion and multiscale faces in the masked face detection task in dense crowd, the backbone of the DRFL network uses ResNet50 [22] as the feature extraction network. C3, C4, and C5 represent the low-level feature, middle-level feature, and high-level feature extracted for the image. P3, P4, and P5 are feature fusion in the FPN network through upsampling and residual connection. The fused features are used to enhance feature extraction, increase the scope of the receptive field, and enhance the robustness of small-scale face detection through independent Enhanced Receptive Field Context (ERFC) module. The DRFL network structure is shown in Figure 5.

The entire feature extraction network combines top-down and bottom-up feature fusion strategies to improve the multiscale prediction network. Finally, a multitask loss function is used to fully consider the central point distance between the face and the detection frame, overlap rate, and key point information, thereby improving the accuracy of face detection.

3.2. Enhanced Receptive Field Context Module. The ERFC module with special dilation convolution is used to extract the feature output by the FPN. The advantage of using dilation convolution is that it can increase the receptive field while avoiding the loss of information caused by the pooling operation. Each convolution output contains a larger range of information and captures multiscale context information. As shown in Figure 6, (a) corresponds to 3×3 convolution with dilation rate 1, which is the same as ordinary convolution operation, (b) corresponds to 3×3 convolution with dilation rate 2, and the receptive field has increased to 5×5.

The specific operation of ERFC module is to first compute the input features by the 3×3 convolution, and then one of them is to enhance the extraction of context information through the parallel 3×3 convolution with dilation rate 1 and 3×3 convolution with dilation rate 2, in order to improve the detection robustness of occluded faces. At the same time, the local parameters are reduced by 16.7% without changing the receptive field and detection accuracy. Finally, all the outputs are concatenated as the output of the entire ERFC module and transmitted to the next network as shown in Figure 6(c).

3.3. Masked Face Classification Network. The significance of NAS is to solve the parameter adjustment problem of deep learning models, which is a cross-research that combines optimization and machine learning. Before deep learning, the traditional machine learning models might also encounter the problem of parameter adjustment. Because the structure of the shallow model is relatively simple, most studies unify the structure of the model as a super parameter to search, such as the number of hidden neurons in the three-layer neural network. The methods for optimizing these hyperparameters are mainly black box optimization methods, such as evolutionary optimization, Bayesian optimization, and reinforcement learning.

However, in deep learning, with the expansion of the model scale, the number of super parameters also increases, which brings new challenges to the optimization problem. The search space of NAS directly affects the difficulty of optimization. A simple search strategy [23] in neural network search is to multiply each branch by a weight during training and to send the result to the next level. After

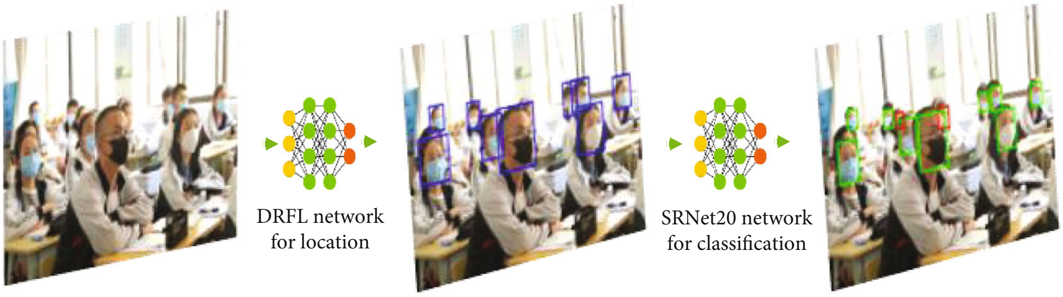


Figure 4: Two cascaded subnetworks.

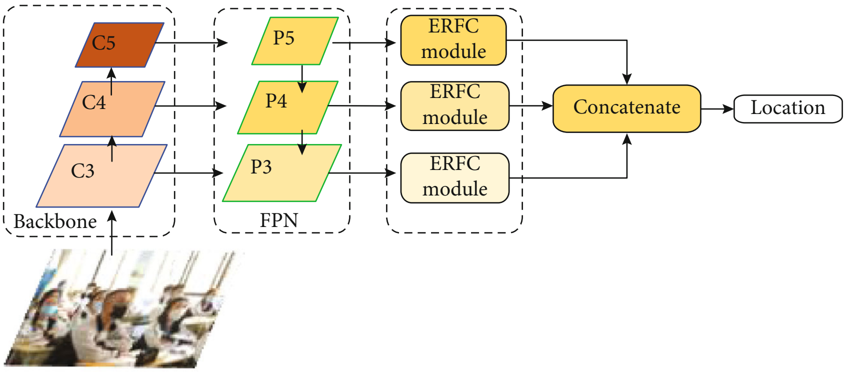
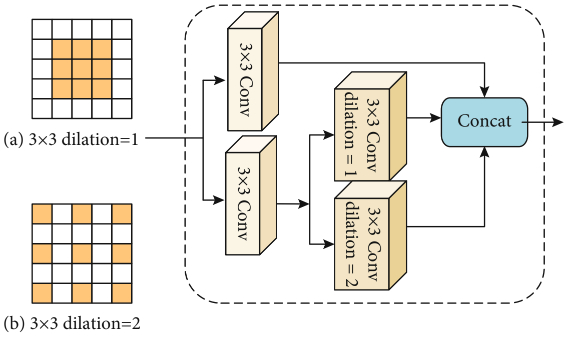


Figure 5: The DRFL network architecture.



(a) 3×3 dilation=1

(b) 3×3 dilation=2

Figure 6: The ERFC module with different dilation rates.

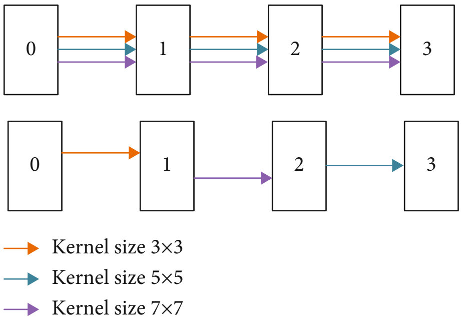


Figure 7: The search strategy in this paper.

training, the branch with the largest weight is retained. The working principle of the search is shown in Figure 7.

Specifically, in this paper, we designed the SRNet20 network based on ResNet18 network, the convolution kernel of 3×3 is replaced by the parallel structure of 3×3, 5×5, 7×7, and the NAS method is used to find the most suitable branch of the task. Then, in the experimental part, we train the searched classification network on our own dataset and compare the results with the original ResNet results on the dataset.

3.4. Model Training Method of Federated Learning.

In this paper, the dataset is divided into 10 disjoint parts, representing 10 independent clients, which simulates the real situation of training the classification network. Client C_i has a local private dataset D_i, and model M_0 is published from the central server.

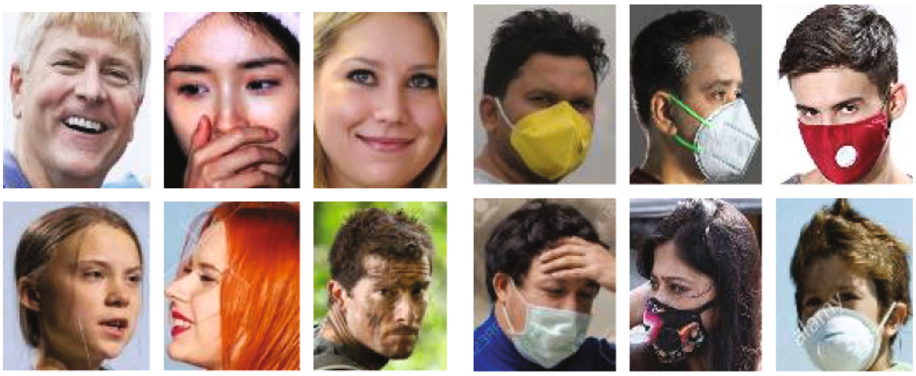
The steps in the training stage are as follows:

(1) Client C_i receives model M_0 from the central server

(2) Client C_i trains the model based on the local dataset D_i and obtains a new model M_i

(3) Client C_i calculates the model parameter difference $M_{\Delta i}$, where $M_{\Delta i} = M_i - M_0$, and uploads the parameter difference $M_{\Delta i}$ to the central server

(4) The central server aggregates the parameter differences uploaded by users, updates the model M_0,



(a) Unmasked face

(b) Masked face

Figure 8: Some images in our dataset.

and resends it to clients participating in federated learning

After a round of update is completed, we check whether the accuracy of the local model meets the requirements. If it meets the requirements, stop training; otherwise, prepare for the next round of training.

4. Experiments and Results

4.1. Dataset. First, the general face location network uses the public Wider Face [24] dataset. It is a benchmark dataset in the field of face detection. It contains 32,203 images and a total of 393,703 annotated faces, of which 158,989 annotated faces are in the training set and 39,496 are in the validation set. Each subset contains 3 levels of detection difficulty: easy, medium, and hard. These different faces have a wide range of changes in terms of scale, posture, illumination, expression, and occlusion. Using this dataset to train the DRFL network will have better detection and location capabilities for faces of different scales.

Second, the masked face classification network is trained on self-made dataset. The training set contains 18,000 images, including 9,000 faces with masks and 9,000 faces without masks. The test set contains 1,751 images, including 656 faces wearing masks and 1,095 faces without masks. The dataset contains face data of different ages, genders, and orientations, which can prevent the network from overfitting the data of a single pose and improve the generalization ability of the network. Some images are shown in Figure 8.

4.2. Loss Function. Based on the loss function of RetinaFace [25], the feature pyramid is adopted to realize the fusion of multiscale information, which plays an important role in the detection of small faces. Its multitask loss function for any training anchor i is shown in the following equation.

$$L = L_{cls}(p_i, p_i^*) + \lambda_1 p_i^* L_{\text{box}}(k_i, k_i^*) + \lambda_2 p_i^* L_{pts}(q_i, q_i^*). \quad (1)$$

There are three parts of the loss function:

(1) Face classification loss $L_{cls}(p_i, p_i^*)$, where p_i is the predicted probability of anchor i which has a face and p_i^* is 1 for the positive anchor and 0 for the negative anchor. L_{cls} is the softmax loss for binary classes

(2) Face box regression loss $L_{\text{box}}(k_i, k_i^*)$, where $k_i = \{k_x, k_y, k_w, k_h\}_i$ and $k^*{}_i = \{\mathrm{k}_{\mathrm{x}}^*, \mathrm{k}_{\mathrm{y}}^*, \mathrm{k}_{\mathrm{w}}^*, k_h^*\}_i$ represent the coordinates of the predicted box and groundtruth box in the positive anchor. $L_{\text{box}}(k_i, k_i^*) = R(k_i - k_i^*)$, where R is smooth L1 defined in [26]

(3) Facial landmark regression loss L_{pts}, where $q_i = \{q_{x_1}, q_{y_1}, \cdots, q_{x_5}, q_{y_5}\}_i$ and $q_i^* = \{q_{x_1}^*, q_{y_1}^*, \cdots, q_{x_5}^*, q_{y_5}^*\}_i$ represent the predicted five facial landmarks and groundtruth associated with the positive anchor. The loss is similar to the box centre regression. The loss-balancing parameters $\lambda_1$and λ_2 are set to 0.25 and 0.1

In the face classification network, we use CrossEntropy loss shown in the following equation.

$$L_{CE} = -\sum_{i=1}^{n} p(x_i) \log (q(x_i)). \quad (2)$$

The $p(x_i)$ represents the real label of x_i, and $q(x_i)$ represents the possibility of x_i measured through the network.

4.3. Setup for Experiments

4.3.1. Data Augmentation. When training the deep learning network, the specific operation randomly cropped the image in the mini-batch to 0.8-1.0 times the size of the original image, and at the same time perform a horizontal flip with a 50% probability, and finally use the resize operation to adjust to a uniform size. Before entering the network, normalize each channel of the image.

The images are randomly cropped and randomly flipped to achieve data augmentation, which improves the accuracy and robustness of the model to a certain extent.

4.3.2. Anchors. The DRFL network uses different anchor boxes in different feature pyramid layers from P3 to P5. In the lower feature layer, small-scale anchor points are tiled to capture small facial features. The high feature layer corresponds to a large area in the original image, so large facial features are captured in the high-level feature layer. The sizes of anchors are shown in Table 1.

4.3.3. Optimization Strategy. In the experiment, the optimization strategy for training the network is to use Adam for the first 10 epochs and SGD for the subsequent epochs. At the 20th epoch, the learning rate decays to 0.1 times, and at 40 epochs, it decays to 0.01 times.

4.4. Tests and Results. In order to test the performance of this network, there are the following three experiments. Experiment 1 tests mAP of the DRFL network. Experiment 2 tests the ERFC module with dilation convolution and without dilation convolution. Experiment 3 compares mAP of original ResNet with SRNet (ResNet after NAS) and verifies the feasibility of federated learning.

(1) *Experiment 1.* Train the DRFL to realize face location, and test the results on the Wider Face validation set. The comparison with other algorithms is shown in Table 2

The results show that our network has advantages in the easy part and the medium part of this validation set. The performance of our algorithm is similar to other algorithms and basically meets the actual needs.

(2) *Experiment 2.* In order to verify the effectiveness of the dilation convolution in the ERFC module, using one 3 × 3 convolution kernel with dilation rate 2 to replace two 3 × 3 convolution kernels with dilation rate 1, we train an unreplaced DRFL network on the same dataset as the baseline and compare it with the replaced network. The test results on the Wider Face validation set are shown in Table 3

The results show that the ERFC module using dilation convolution hardly affects performance while reducing 16.7% parameters, and it is suitable for deploying on embedded cameras.

(3) *Experiment 3.* First, select the appropriate classification model. The convolution kernels of SRNet20, which is created by NAS, are shown in Table 4

Comparing the mAP of the searched network and the original network on the face classification dataset is shown in Table 5.

Comparing with the masked face classification accuracy, the SRNet20 is 8.8% higher than ResNet18, and the SRNet50 is 5.4% higher than the original ResNet50, which proves the effectiveness of the NAS for classification network.

Second, in order to verify the feasibility of federated learning, we simulate a total of 10 clients, and n represents the number of clients who really participate in the training. The model is SRNet20 network, and the number of clients and accuracy are shown in Figure 9.

The result shows that the model quickly overfits when the number of participating clients is small. As the number of participating clients increases, the accuracy gradually rises. After sufficient training, the results of federated training are shown in Table 6.

Finally, masked face detection in the dense crowd is completed by cascade network. The mAP of the face location is 90.6%, and the mAP of the masked face classification is 98.5%. We input the test images into the cascade network, and the results are shown in Figure 10. The red box represents the person without the mask, and the green box represents the person with the mask. The near faces can be correctly detected even with slight occlusion, but the blurred faces in the distance are still missed, and this is also the direction for future improvements.

Table 1: Anchor size in DRFL network.

Feature pyramid	Anchor
P3 (80 × 80 × 64)	16, 20.16, 25.40
P4 (40 × 40 × 64)	64, 80.3, 101.59
P5 (20 × 20 × 64)	256, 322.54, 406.37

Table 2: Accuracy on the Wider Face validation set.

Method		Difficulty	
	Easy	Medium	Hard
MTCNN	84.8%	82.5%	59.8%
Face R-CNN	93.7%	92.1%	83.1%
SSH	93.1%	92.1%	84.5%
DRFL (ours)	94.7%	93.0%	84.2%

Table 3: Results on the Wider Face validation set.

Model		Difficulty	
	Easy	Medium	Hard
DRFL (without dilation)	94.7%	93.1%	84.4%
DRFL (with dilation)	94.7%	93.0%	84.2%

Table 4: Kernel sizes of layers.

Model	Layer 1	Layer 2	Layer 3
SRNet 20	3 × 3, 5 × 5, 7 × 7	7 × 7, 3 × 3, 5 × 5	7 × 7, 7 × 7, 7 × 7

Table 5: mAP of the original network and the searched network.

Model	mAP
ResNet18 (pretraining)	90.0%
ResNet50 (pretraining)	93.0%
SRNet20	98.8%
SRNet50	98.4%

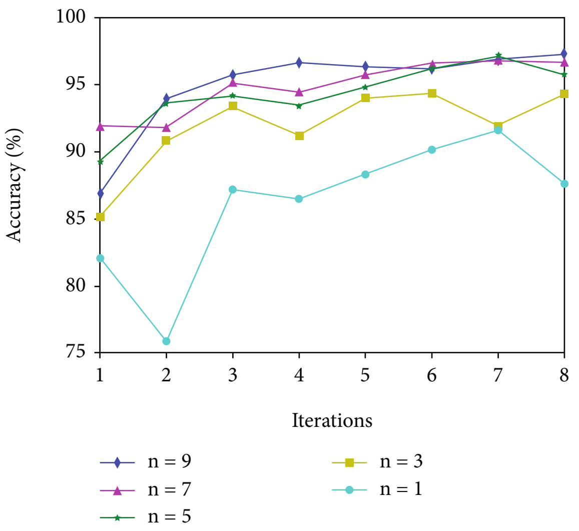


FIGURE 9: Influence of different number of clients on accuracy.

TABLE 6: The mAP of centralized training and federated learning.

Method	mAP
SRNet20 (centralized training)	98.8%
SRNet20 (federated learning)	98.5%

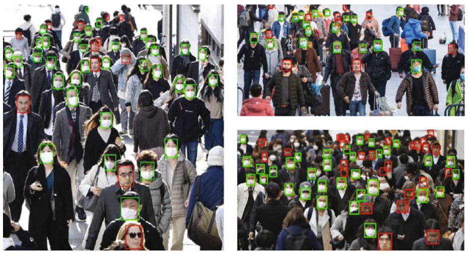


FIGURE 10: Results of our algorithm.

5. Conclusions

In this paper, we create the DRFL network to implement multiscale face location and create SRNet20 network by NAS to classify masked faces. For privacy protection, we introduce federated learning to provide a joint training solution for multiparty data sources in the real world. By cascading the two networks, the purpose of masked face detection in dense crowds is achieved. From the effect of the test images, our DRFL network has good performance. But for long-distance faces that are blurred or severely occluded, the detection effect needs to be further improved. In the future, we can increase the dataset or adjust the network structure to enhance the network detection robustness. Or we may use a lightweight backbone network to achieve real-time detection in dense crowd scene and apply it to actual life scenarios.

Data Availability

Data is available at https://github.com/woshizr/masked-Face.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was supported by Centre for Public Security Information and Equipment Integration Technology of UESTC (University of Electronic Science and Technology of China). Thanks are due to the computing platform provided by the laboratory.

References

[1] World Health Organization, *COVID-19 weekly epidemiological update, 13 July 2021*, WHO, 2021.

[2] N. H. L. Leung, D. K. W. Chu, E. Y. C. Shiu et al., “Respiratory virus shedding in exhaled breath and efficacy of face masks,” *Nature Medicine*, vol. 26, no. 5, pp. 676–680, 2020.

[3] Z. Cai, Z. He, X. Guan, and Y. Li, “Collective data-sanitization for preventing sensitive information inference attacks in social networks,” *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 577–590, 2018.

[4] Z. Cai, Z. Xiong, H. Xu, P. Wang, W. Li, and Y. Pan, “Generative adversarial networks: a survey towards private and secure applications,” *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–38, 2021.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: unified, real-time object detection,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, Las Vegas, NV, USA, 2016.

[6] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, Venice, Italy, 2017.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, 2014.

[8] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2117–2125, Honolulu, HI, USA, 2017.

[9] 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.

[10] H. Wang, Z. Li, X. Ji, and Y. Wang, “Face r-cnn,” 2017, https://arxiv.org/abs/1706.01061.

[11] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: towards real-time object detection with region proposal networks,” *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.

[12] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, “Ssh: single stage headless face detector,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4875–4884, Venice, Italy, 2017.

[13] X. Tang, D. K. Du, Z. He, and J. Liu, “Pyramidbox: a context-assisted single shot face detector,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 797–813, Munich, 2018.

[14] Y. WANG, X. ZHANG, J. YE, H. SHEN, Z. LIN, and W. TIAN, “Mask-wearing recognition in the wild,” *SCIENTIA SINICA Informationis*, vol. 50, no. 7, pp. 1110–1120, 2020.

[15] W. Liu, D. Anguelov, D. Erhan et al., “Ssd: Single shot multibox detector,” in *Computer Vision – ECCV 2016*, pp. 21–37, Springer, Cham, 2016.

[16] M. Jiang, X. Fan, and H. Yan, “Retinamask: a face mask detector,” 2020, https://arxiv.org/abs/2005.03950.

[17] Z. Wang, G. Wang, B. Huang et al., “Masked face recognition dataset and application,” 2020, https://arxiv.org/abs/2003.09093.

[18] Z. Cai and X. Zheng, “A private and efficient mechanism for data uploading in smart cyber-physical systems,” *IEEE Transactions on Network Science and Engineering (TNSE)*, vol. 7, no. 2, pp. 766–775, 2020.

[19] Z. Xu and Z. Cai, “Privacy-preserved data sharing towards multiple parties in industrial IoTs,” *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 38, no. 5, pp. 968–979, 2020.

[20] B. Mcmahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Artificial intelligence and statistics*, pp. 1273–1282, PMLR, 2017.

[21] Z. Cai and Z. He, “Trading Private Range Counting over Big IoT Data,” in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, Dallas, TX, USA, 2019.

[22] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016.

[23] H. Liu, K. Simonyan, and Y. Yang, “Darts: differentiable architecture search,” 2018, https://arxiv.org/abs/1806.09055.

[24] S. Yang, P. Luo, C. C. Loy, and X. Tang, “Wider face: a face detection benchmark,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5525–5533, Las Vegas, NV, USA, 2016.

[25] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, “Retinaface: single-stage dense face localisation in the wild,” 2019, https://arxiv.org/abs/1905.00641.

[26] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, Santiago, Chile, 2015.