Identifying Facemask-wearing Condition Using Image Super-Resolution with Classification Network to Prevent COVID-19

BOSHENG QIN, DONGXIAO LI

College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, 310058, China.

Corresponding Author: DONGXIAO LI (lidx@zju.edu.cn).

Abstract

Rapid worldwide spread of Coronavirus Disease 2019 (COVID-19) has resulted in a global pandemic. Correct facemask-wearing is valuable in infectious disease control, but the effectiveness of facemasks has been diminished mostly due to improper wearing. However, there have not been any published reports on the automatic identification of facemask-wearing conditions. In this study, we developed a new facemask-wearing condition identification method in combination with image super-resolution with classification network (SRCNet), which quantified a three categories classification problem based on unconstrained 2D facial image images. The proposed algorithm contained four main steps: image pre-processing, face detection and crop, image super-resolution, and facemask-wearing conditions identification. Our method was trained and evaluated on public dataset Medical Masks Dataset containing 3835 images with 671 images of no facemask-wearing, 134 images of incorrect facemask-wearing, and 3030 images of correct facemask-wearing. Finally, the proposed SRCNet achieved 98.70% accuracy and outperformed traditional end-to-end image classification methods using deep learning without image super-resolution by over 1.5% in kappa. Our findings indicate that the proposed SRCNet could achieve high accuracy identification in facemask-wearing conditions, which have potential application in epidemic prevention involving COVID-19.

Introduction

Coronavirus disease 2019 (COVID-19) is an emerging respiratory infectious disease caused by Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV2) [1]. Currently, COVID-19 had quickly spread to the majority of countries worldwide, affected more than 3.32 million individuals, and caused nearly 236,431 deaths, according to the report of World Health Organization (WHO) on the 3rd of May, 2020. To avoid tragedy in the world, a practical and straightforward approach to prevent the spread of the virus is emergently desired worldwide.

Previous studies found that facemask-wearing is valuable to prevent the spread of respiratory viruses[2-4]. For instance, the efficiency of N95 and surgical masks in blocking the SARS transmission are 91% and 68% respectively[5]. The facemask-wearing can interrupt airborne viruses and particles effectively, such that these pathogens could not enter the respiratory system of a person[6]. As a non-pharmaceutical intervention, facemask-wearing is a noninvasive and cheap method to reduce mortality and morbidity from respiratory infections. Since the outbreak of COVID-19, facemasks are routinely used by the general public to reduce exposure to airborne pathogens in many countries[7]. Patients suspected or actual infection COVID-19 are required to wear facemasks[1]. Facemasks, when fitted properly, effectively disrupt the forward momentum of particles expelled from a cough or sneeze, preventing disease transmission[5]. However, the effectiveness of facemasks in containing the spread of airborne diseases in general public has been diminished mostly due to improper wearing[8]. Therefore, it is necessary to develop an automatic detection approach for facemask-wearing conditions, which could be contributing to personal protection and public epidemic prevention.

The distinctive facial characteristics in facemask-wearing conditions provide an opportunity for automatic identification. Recent advances in computer vision and deep learning present the opportunity for development in many fields[9]. As the main component of deep learning methods, deep neural networks (DNNs) have shown their superior performance in many fields, including object detection, image classification, and image segmentation[10-13]. One primary model of DNNs is convolutional neural networks (CNNs), which are widely applied in computer vision tasks. After training, CNNs can recognize and classify facial images even with slight differences for their powerful feature extraction capability. As one of the CNNs, the image super-resolution (SR) networks can restore image details. Recently, the SR networks go more in-depth, and the idea of auto-encoder and residual learning are added for performance improvement[14,15]. The SR networks are also applied for image processing before classification, which reconstruct images for higher resolution and restore details[16-19]. Moreover, the SR networks improve the classification accuracy significantly, especially for the dataset with low-quality images, and provide a feasible solution to improve facemask wearing conditions identification performance. Therefore, the combination of SR network with classification network (SRCNet) could utilize in facial image classification for accuracy improvement.

To our knowledge, there have not been any published reports related to SR network with classification network in facial image classification for accuracy improvement, especially regarding automatic detection of facemask-wearing conditions. Therefore, we intend to develop a novel method in combining SR network with classification network (SRCNet) to identify facemask-wearing conditions in order to improve classification and accuracy with low-quality facial images.

Our main contributions can be summarized as follows.

1) Develop a new facial image classification method, which combining SR network with classification network (SRCNet) in facial image classification.
2) Improve SR network structure, including activation function and the density of skip connections, which outperformed previous methods.
3) Utilize deep learning method in automatic identification of facemask-wearing conditions.

Methods

This section describes the technology related to the SRCNet and facemask-wearing condition identification, including proposed algorithm, image pre-processing, face detection and crop, SR network, facemask-wearing condition identification network, datasets, and training details. Facemask-wearing condition identification is a kind of three categories classification problem, including no facemask-wearing (NFW), incorrect facemask-wearing (IFW), and correct facemask-wearing (CFW). Our goal is to form a function F(x), which inputs an unprocessed image and output the conditions of wearing masks for all faces in the image. The data was collected from the public dataset Medical Masks Dataset in the Kaggle (https://www.kaggle.com/vtech6/medical-masks-dataset). Informed consent was obtained from the participant to publish the images in an online open-access publication.

Proposed Algorithm

Fig. 1 offered the diagram of the proposed algorithm, which contained three main steps: image pre-processing, face detection and crop, and SRCNet for SR and facemask-wearing condition identification. After the pre-processing of raw images, all the face areas of images were detected using multitask cascaded convolutional neural network. The face areas were then cropped, and the size of cropped images varies. All cropped images were then sent to SRCNet for facemask-wearing condition identification. In SRCNet, all images were judged for the need of SR. As the size of input images of facemask-wearing condition identification network was 224×224, the cropped images which had size no more than 150×150 (width or length no more than 150) were sent to SR network, then for facemask-wearing condition identification. Otherwise, the cropped images were then directly sent for facemask-wearing condition identification. The output was the possibility of the input images with three categories: NFW, IFW, and CFW. After passing the classifier, the pipeline output the final facemask-wearing condition results.

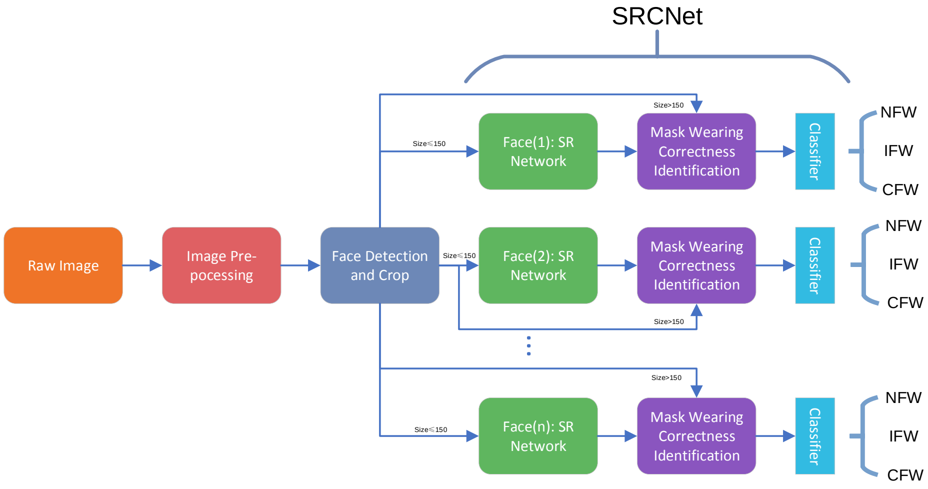

Figure 1. Diagram of the proposed algorithm. NFW = no facemask-wearing, IFW = incorrect facemask-wearing, CFW = correct facemask-wearing.

Image Pre-processing

The goal of image pre-processing was to improve the accuracy in following face detection and facemask-wearing condition identification. The SRCNet was designed to be applied in public for classification, which took uncontrolled 2D RGB images as input. As the raw images taken in real life had a considerable variance in exposure and contrast, image pre-processing was needed for the accuracy of face detection and facemask-wearing condition identification. The raw images were adjusted using MATLAB image processing toolbox by mapping the values of the input intensity image to the new value, in which 1% of the values were saturated at low and high intensities of the input data.

Face Detection and Crop

As the SRCNet need to concentrate on the information from faces rather than the background to improve accuracy, a face detector was needed for the detection of faces and crop face areas. The uncontrolled 2D images had differences in face size, expression, and background. Hence a robust and high accurate face detector was needed. The multitask cascaded convolutional neural network was adopted for face detection, which performed well in getting face areas in real environments[10].

After getting the position of the face, the faces were then cropped from the pre-processed image as the inputs of SR network or facemask-wearing condition identification network depend on image sizes.

SR Network

The first stage of SRCNet was SR network. The cropped face images had a huge variance in size, which could possibly damage the final identification accuracy of SRCNet. Hence, SR was applied before classification. The structure of SR network was inspired by RED[14], which used convolutional layers as auto-encoder and deconvolutional layers for image up-sample. The symmetric skip connections were also applied for preservative of image details. The detailed architectural information of SR network was shown in Fig. 2.

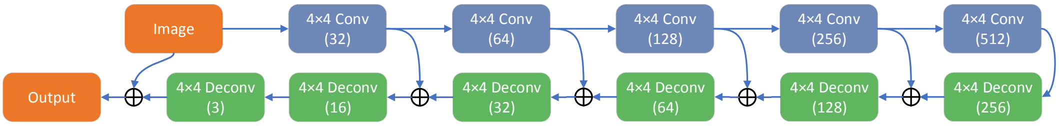

Figure 2. Structure of SR Network.

The SR network had five convolutional layers and six deconvolutional layers. Except for the final deconvolutional layer, which mixed the information from the input image and output three feature maps correspondent to RGB, all other convolutional layers were connected with their corresponding convolutional layers via skip connections. With skip connections, the information was propagated from convolution feature maps to the corresponding deconvolutional layers. The network was then fit for solving the residual of the problem, which is denoted as:

$$F(X) = Y - X \tag{1}$$

Where Y was the expected output image, X was the input image, and $F(X)$ was the function of the SR network.

In convolutional layers, the number of kernels was designed to increase by a factor of 2. With kernels size in 4×4 and stride 2, after passing through the first convolutional layers for feature extraction, every time the image passed through a convolutional layer, the size of feature maps decreases by a factor of ½. Hence, the convolutional layers performed as an auto-encoder and extracted the feature from the input image.

In deconvolutional layers, the number of output feature maps was symmetry to their corresponding convolutional layers to satisfy skip connections. The number of kernels in every deconvolutional network was decreased by a factor of ½ except the final deconvolutional layers, while with kernels size in 4×4 and stride 2, the size of feature maps increases by a factor of 2. After the information combination in the final deconvolutional layer, the output was the image with the same size as the input image. The deconvolutional layers acted as a decoder, which took the output of encoder as input and up-sampling to get a super-resolution image.

It is worth mention that the function used for down and upsampling was stride in convolutional and deconvolutional layers rather than pooling and un-pooling layers, for the aim of SR network was to restore image details rather than learning abstractions. The pooling and un-pooling layers, however, damaged the details of images and deteriorated the restoration performance[14].

The function final deconvolutional layer combined all the information from the previous deconvolutional layer and input image and normalized all pixels to [0,1] for output. The stride for the final deconvolutional layer was set to 1 for information combination without up-sampling, with three output layers corresponding to RGB. The activation function of the final deconvolutional layer was Clipped Rectified Linear Unit, which forced the normalization of output and avoiding error in Loss computing. The definition of Clipped Rectified Linear Unit was as follow:

$$ClippedReLU(x) = min(1, max(0, x)) \tag{2}$$

Where x was the input value, $ClippedReLU(x)$ was the output of Clipped Rectified Linear Unit function.

One main difference between our model and RED was that the improvement of activation function, which was changed from Rectified Linear Unit (ReLU) to Leaky Rectified Linear Unit (LeakyReLU), which were defined as follows:

$$ReLU(x) = max(0, x) \tag{3}$$

$$LeakyReLU(x) = max(0, x) + min(0, \alpha \times x) \tag{4}$$

Where x was the input value, α was scale factor, ReLU(x) was the output of ReLU function, and LeakyReLU(x) was the output of LeakyReLU function.

The reason was that the skip connections propagate the image from input to output. For SR network, the network shall have the capability to subtract or add values for pixels, where the ReLU can only add values for feature maps. LeakyReLU, however, activated neurons with negative values, which improve the performance of the network.

Another difference was the density of skip connections. Rather than using skip connections every a few (two in RED) layers from convolutional layers to their symmetrical deconvolutional feature maps, the density of skip connections increased, and all convolutional layers were connected to their mirrored deconvolutional layers. The reason was to make all layers learn to solve the residual problem, which reduced the loss of information between layers while did not significantly increase network parameters.

The gold of SR network training was to update all learnable parameters to minimize Loss. For SR network, the Mean Squared Error (MSE) was widely used as Loss[14,20-22]. For the reduction of overfitting, a regularization term (weight decay) for the weights was added to cross-entropy loss. Hence, the MES with L_2 regularization was applied as loss function $Loss(w)$, which was defined as:

$$Loss(w) = \frac{1}{N}\sum_{i=1}^{N} ||Y_i - X_i||^2{}_F + \frac{1}{2} \times \lambda \times w^T w \tag{5}$$

Where Y_i was the ground truth, X_i was the output image, $Loss(w)$ was the Loss for collections of given w.

It was worth mentioning the size of the input image can be arbitrary, and the output image had the same size as the input image. The convolutional and deconvolutional layers were symmetric for SR network. Besides, the network is predicted in pixel-wise. Whereas, for better detail enhancement, which needed to dedicate image input size for SR network training, the input images were resized to 224×224×3 with bicubic, which was the same as the input image size of facemask-wearing condition identification network.

Facemask-wearing Condition Identification Network

The second stage of SRCNet was facemask-wearing condition identification. As CNN was one of the most common types of networks for image classification, which performed well in face recognition, CNN was adopted as facemask-wearing condition identification network in the second stage of SRCNet. The goal was to form a function G(x), where x was the input face image, and G(x) output the possibilities with three categories (MFW, IFW, CFW). The classifier then output the classification result based on the output possibilities.

Mobilenet-v2 was applied as facemask-wearing condition identification network, which was a lightweight CNN and achieved high accuracy in image classification. The main features of mobilenet-v2 were residual blocks and depthwise separable convolution[23,24]. The residual blocks contribute to the training of deep network, which solved the problem of gradient vanishing and achieving benefits on back-propagating the gradient to bottom layers. As for facemask-wearing condition identification, there were slight differences between IFW and CFW. Hence, the capability of feature extraction or the depth of network became essential, which contributed to the final identification accuracy. The depthwise separable convolution was applied for the reduction of reducing computation and model size while maintaining the final classification accuracy, which separable convolution splits this into two layers, a separate layer for filtering and a separate layer for combining.

Transfer learning was applied in the network training procedure, which was a kind of knowledge migration between source and target domain. The network was trained in three steps: initialization, form a general face recognition model, and knowledge transfer to facemask-wearing condition identification. The first step was initialization, as initialization contributed to the final identification accuracy and training speed[25,26]. Then was to form a general face recognition model using a large facial image dataset, where the network gained the capability of facial feature extraction. After watching millions of faces, the network then concentrated on facial information rather than interference like backgrounds and the difference because of image shooting parameters. The final step was for knowledge transfer between face recognition to facemask-wearing condition identification. The final fully connected layer was modified to meet with the category requirement of facemask-wearing condition identification.

The reason for adopting transfer learning was the considerable differences in data volumes and their consequences. The facemask-wearing condition identification dataset was relatively small compared with the general face recognition dataset, which might cause overfitting problems and a reduction in identification accuracy in the training process. Hence, the network shall gain knowledge about faces for the reduction of overfitting and improvement of accuracy.

The softmax function was used as the final stage of classifier, which calculated the probability of all classes using outputs of direct ancestor fully connected layer neurons[27]. The definition was:

$$p_i = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{6}$$

where x_i was the total input received by unit i, p_i was the prediction possibility of the image belongs to class i.

For image classification, the cross-entropy was widely used as loss function[28,29]. A regularization term (weight decay) can significantly avoid overfitting. Hence, the cross-entropy with L_2 regularization was applied as loss function $Loss_R$, which was defined as:

$$Loss_R = -\sum_{i=1}^{N}\sum_{j=1}^{K} t_{ij}\log(y_{ij}) + \frac{1}{2}\times\lambda\times w^T w \tag{7}$$

For the cross-entropy term, N was the number of samples, K was the number of classes, t_{ij} was the indicator that the i^{th} sample belongs to the j^{th} class, which was 1 when labels corresponded and 0 when different, and y_{ij} was the output for sample i for class j, which was the output value from softmax layer. For the cross-entropy term, the w was the learned parameters in every learned layer, and the λ was the regularization factor (coefficient).

The training goal of the facemask-wearing condition identification network was to minimize the cross-entropy loss with weight decay. When training, the real classes were given in on-hot vectors, which took the form:

$$V = [x_0, x_1, \dots, x_n] \tag{8}$$

Where V represents the input label vector, x_i $(i \in [1,n])$ represent the real class of the image, n represent the class number. When the image was in i class, then $x_i = 1$ and $x_k = 0 (k \neq i)$.

Datasets

Different facial image datasets were used for different network training for the improvement of the generalization ability of SRCNet. The public facial image dataset CelebA were processed and used for SR network training[30]. As the goal of SR network was for detail enhancement, a large and high-resolution facial image dataset was needed, where the CelebA met with all requirements.

The process of CelebA included three steps: image pre-processing, face detection and crop, and image selections. All raw images were pre-processed, as mentioned above. The face areas were then detected by multitask cascaded convolutional neural network and cropped for training since SR network was designed for restoration detailed information on faces rather than the background. The cropped images, which were smaller than 224×224 (input size of facemask-wearing condition identification network), or non-RGB images, were discarded automatically. All other cropped facial images were inspected manually, and images with blur or dense noise were also discarded. Finally, 70534 high-resolution facial images were split into a training dataset (90%) and a testing dataset (10%) and adopted for SR network training and testing.

The training of the facemask-wearing condition identification network contained three steps. Each step used a different data set for training. For initialization, the goal was for generalization. A large-scale classification dataset was needed for better generalization, and the ImageNet dataset was adopted for network initialization[11]. During this procedure, non-zero values were assigned to parameters, which increased the generalization ability. Besides, proper initialization significantly improved the training speed, informing the general face recognition model.

A large-scale face recognition database was adopted to train the general face recognition model. The images from publicly available CASIA Web-Face datasets were screened manually, and those with poor image quality or the subject containing insufficient images were discarded [31]. Finally, 493750 images with 10562 subjects were split into a training dataset (90%) and testing dataset (10%) and applied for general face recognition model training.

The public facemask-wearing condition dataset Medical Masks Dataset was applied to fine-tuning the network to transfer knowledge from general face recognition to facemask-wearing condition identification. The 2D RGB images were taken in uncontrolled environments, and all faces in the dataset had their position coordinates with facemask-wearing condition labels. The Medical Masks Dataset was processed in four steps: face crop and labeling, label confirmation, image pre-processing, and SR. All faces were cropped and labeled using the given position coordinates and labels. All cropped facial images were then screened manually, and those with incorrect labels were discarded. Then facial images were confirmed and pre-processed using the methods mentioned above. For the final accuracy of SRCNet, the dataset was expanded for the case of not wearing a mask. The resolution of pre-processed images varies, as shown in Table 1. For accuracy improvement of facemask-wearing condition identification network, the facial image shall contain enough details. Hence, SR network was applied to adding details for images with low quality. Those sizes no more than 150×150 (width or length no more than 150) were processed using SR network. Finally, the dataset contained 671 images of NFW, 134 images of IFW, and 3030 images of CFW. The whole dataset was separated into a training dataset (80%) and a testing dataset (20%) for facemask-wearing condition identification network training and testing.

Table 1. Image resolution statistics

Resolution N×N	NFW	IFW	CFW	Total
N≤64	307	34	1126	1467
N≤112	199	33	984	1216
N≤150	73	20	355	448
N≤224	77	33	354	464
N>224	15	14	211	240
Total	671	134	3030	3835

Training Details

The training of SRCNet contained two main steps: SR network training and facemask-wearing condition identification network training.

For SR network training, the training goal was to restore face details, which used the training set of CelebA to achieve. Based on the characteristics of Medical Masks Dataset, the input images were pre-processed as an imitation of low-quality images in Medical Masks Dataset. The high-resolution processed images in CelebA were first filtered with a Gaussian filter with kernel size of 5×5 and a standard deviation of 10. Then down-sampled to 112×112. Since the size of the input and output were the same, the down-sampled images were then up-sampled to 224×224 with bicubic as input, the same size as the input of facemask-wearing condition identification network. The Adam was adopted as optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ [32]. The network was trained for 200 epochs with initial learning rate of 1e−4 and learning rate dropping factor of 0.9 after every 20 epochs. The mini-batch size was 48.

The first step of facemask-wearing condition identification network training was initialization. The network was trained using the ImageNet dataset and training parameters proposed in [23].

The second step was to form a general face recognition model. The output classes were modified to match with the class numbers (10562). For initialization, the weight and bias in the final modified fully connected layer were initialized in a normal distribution with 0 mean and 0.01 standard deviation. The network was trained for 50 epochs, with training dataset shuffled in every epoch. The learning rate was set to be 1e-4, with a learning rate drop factor of 0.9 in every 6 epochs. As we had tested, the loss became stable in around 6 epochs with a constant learning rate, and the learning rate drop factor of 0.9 in every 6 epochs could significantly increase the training speed. The network was trained using Adam as optimizer, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and 1e-4 weight decay for L_2 regularization to avoid overfitting[32].

Transfer learning was applied for fine-tuning the facemask-wearing condition identification network, where the final fully connected layer was modified to match classes (NFW, IFW, and CFW). The weights and biases in the final modified layer were initialized by independently sampling from a normal distribution with zero mean and standard deviation 0.01, which produced superior results compared to other initializers. The Adam was adopted as optimizer with a learning rate of 1e-4, and 1e-4 weight decay for L_2 regularization was also applied to avoid overfitting[32]. The network was trained for 8 epochs to reach high accuracy with a batch size of 16. All the parameters mentioned above were determined using a grid search method for the best identification accuracy on the testing dataset.

Data augmentation was proven to have a significant impact on the final accuracy of the network for the reduction of overfitting[33-35]. For general face recognition network training, the training dataset was randomly rotated in a range of 10 degrees (in norm distribution), shifted vertically and horizontally in a range of 8 pixels, and horizontally flipped in every epoch. At the fine-tuning stage, augmentation was mild, with rotation in 6 degrees (in norm distribution), shift in 4 pixels (vertically and horizontally), and random horizontal flip in every epoch.

Results

The SRCNet was experimented using MATLAB, associated with deep learning toolbox for network training and image processing toolbox for image processing[36]. The SRCNet was trained using a single Nvidia GPU with CUDA and cuDNN enabled.

SR Network Experiment Results

For SR network, the most widely used full-reference quality metrics were peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [37]. The PSNR was metric for quantitatively evaluating image restoration quality, and SSIM compared local patterns of pixel intensities for luminance and contrast.

The comparison with previous state-of-arts methods, including RED[14], SRCNN[20], and bicubic, was made to illustrate the performance of proposed SR network. All the methods were trained on training set of CelebA if needed and tested on the testing set.

As in real application, the quality of images varied. Low-quality images were mainly manifested in resolution and blur. Hence, different quality images were simulated and used for testing the performance of SR network, which were by changing the resolution, the standard deviation σ, and the size of Gaussian filters. For testing with different image resolution, the testing set was first filtered with a Gaussian filter with a kernel size of 5×5 and a standard deviation of 10. Then down-sampled to 64×64, 96×96, 112×112, and 150×150 for evaluations. For testing with different standard deviation of Gaussian filters, the testing set was first filtered with Gaussian filters with kernel size of 5×5 and standard deviation of 5, 10, 15, and 20. Then down-sampled to 112×112 for evaluations. For testing with different kernel size of Gaussian filters, the testing set was first filtered with Gaussian filter with kernel size of 3×3, 5×5, 7×7, 9×9, and standard deviation of 10. Then down-sampled to 112×112 for evaluations.

The size of input images was the same as outputs for SR network. For evaluation of the effect of SR network to facemask-wearing condition identification network, which took 224×224 RGB image as input, all the down-sampled testing sets were up-sampled to 224×224 as the input of SR network. The evaluation results were shown in Table 2, Table 3, and Table 4. Compared with RED, which was the innovation baseline, the proposed SR network outperformed in PSNR and SSIM.

Table 2. Average PSNR and SSIM in Different Down Sampled Images.

PSNR				
Down Sample	**Proposed SR network**	**RED**	**SRCNN**	**Bicubic**
64×64	**25.7461**	25.7343	25.9489	27.1157
96×96	**27.6264**	27.6412	27.8157	27.4728
112×112	**29.1508**	28.8447	28.1890	27.4700
150×150	**28.7863**	28.5277	28.3934	27.3977
SSIM				
Down Sample	**Proposed SR network**	**RED**	**SRCNN**	**Bicubic**
64×64	**0.8984**	0.8974	0.8886	0.9083
96×96	**0.9238**	0.9232	0.9237	0.9166
112×112	**0.9397**	0.9366	0.9294	0.9176
150×150	**0.9311**	0.9293	0.9260	0.9172

Table 3. Average PSNR and SSIM in Different Standard Deviation σ of Gaussian Filters.

PSNR				
σ	**Proposed SR network**	**RED**	**SRCNN**	**Bicubic**
5	**29.3795**	29.0487	28.3087	27.5618
10	**29.1508**	28.8447	28.3934	27.4700
20	**29.0887**	28.7877	28.1599	27.4467
30	**29.0765**	28.7764	28.1544	27.4421
50	**29.0722**	28.7726	28.1525	27.4402
SSIM				
σ	**Proposed SR network**	**RED**	**SRCNN**	**Bicubic**
5	**0.9423**	0.9390	0.9312	0.9187
10	**0.9397**	0.9366	0.9294	0.9176
20	**0.9390**	0.9360	0.9290	0.9173
30	**0.9389**	0.9358	0.9289	0.9172
50	**0.9388**	0.9358	0.9289	0.9172

Table 4. Average PSNR and SSIM in Different Kernel Size of Gaussian Filters.

PSNR				
Kernel Size	**Proposed SR network**	**RED**	**SRCNN**	**Bicubic**
3×3	**31.0008**	30.9300	29.9178	30.5641
5×5	**29.1508**	28.8447	28.3934	27.4700
7×7	**25.3330**	25.2739	25.5768	25.5098
SSIM				
Kernel Size	**Proposed SR network**	**RED**	**SRCNN**	**Bicubic**
3×3	**0.9616**	0.9593	0.9489	0.9482
5×5	**0.9397**	0.9366	0.9294	0.9176
7×7	**0.8820**	0.8808	0.8846	0.8895

As can be observed from Table 2, after the size of the image reached 150×150, the performance of the network decreased. The reason was that the network was trained to restore blurred images with low resolution. As the increase of image resolution, the resolution, and detail of facial image increase, which undermined the condition of using the network. Hence, only images with size no more than 150×150 (width or length no more than 150) were processed with SR network. In this case, the SR network significantly outperformed bicubic.

As the images in Medical Masks Dataset had a considerable variance in resolution, which required the SR network had good performance in different resolution images. Hence, different SR methods were compared and visualized with different resolutions of small and blurred images[38]. The testing image was first blurred with a Gaussian filter with a kernel size of 5×5 and a standard deviation of 10. Then down-sampled to 64×64, 96×96, 112×112, and 150×150 respectively before restoration. The visualized results were shown in Fig. 3. Although all SR methods can enhance face details, the proposed SR network outperformed other methods in all resolutions. The images restored by the proposed SR network were closer to the ground truths, which was due to high PSNR and SSIM.

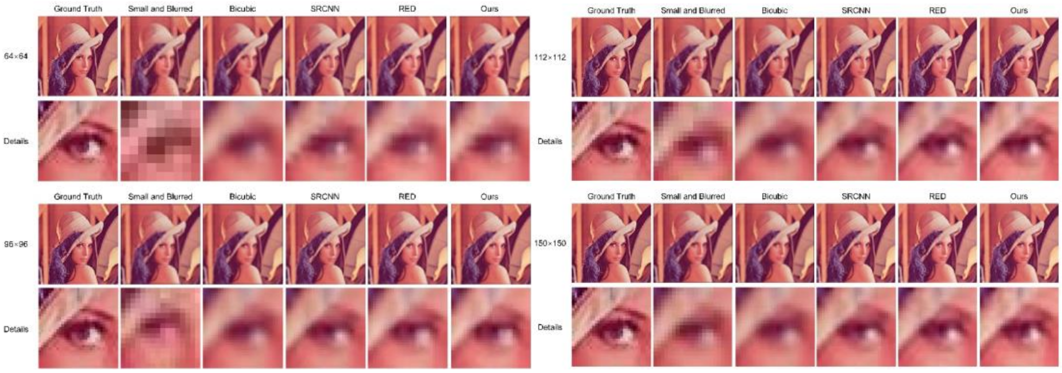

Figure 3. Visualize result with 64×64, 96×96, 112×112, and 150×150 blurred images. The details of images were highlighted.

SRCNet Results

After training, the SRCNet was tested using the testing set of Medical Masks Dataset based on the hold-out method. The proposed algorithm was tested using ablation experiment. The comparison in accuracy and the confusion matrix of SRCNet were reported.

The ablation experiment was designed to illustrate the importance of transfer learning and SR network. The performance of the SRCNet with or without transfer learning or proposed SR network were compared, as shown in Table 5. The transfer learning and SR network increased the identification accuracy considerably by reducing the overfitting problem and increasing facial details. Finally, the SRCNet reached an accuracy of 98.70% and outperformed the mobilenet-v2 without transfer learning or SR network by over 1.5% in kappa.

Table 5. Performance of Facemask-Wearing Condition Identification Network

Method	Accuracy	Mask Wearing Accuracy	Personal Protection Accuracy	κ
1	**98.70%**	**99.09%**	**98.83%**	**96.22%**
2	98.17%	98.57%	98.44%	94.69%
3	97.91%	98.70%	98.17%	93.90%
4	97.26%	98.57%	97.39%	92.12%

1: Proposed SRCNet. 2: Proposed SRCNet without SR network, which was an end-to-end facemask-wearing condition identification network with transfer learning. 3: Proposed SRCNet without transfer learning. 4: Proposed SRCNet without transfer learning or SR network, which was an end-to-end facemask-wearing condition identification network without transfer learning.

Accuracy: Accuracy in three categories classification (NFW, IFW, and CFW). Mask Wearing Accuracy: Accuracy in whether wearing a mask (facemask-wearing, NFW). Personal Protection Accuracy: Accuracy in whether having well personal protection (fail to have personal protection, including NFW and IFW, having personal protection, two categories classification). κ: kappa in three categories classification.

The example of identification results was shown in Fig. 4. Although the face positions and types of facemasks vary, the SRCNet correctly identified all facemask-wearing conditions with high confidence (over 0.9999). As analyzed from failed cases, the critical states (wearing facemask between CFW and IFW), image quality, and blocked faces were the three main reasons for identification errors.

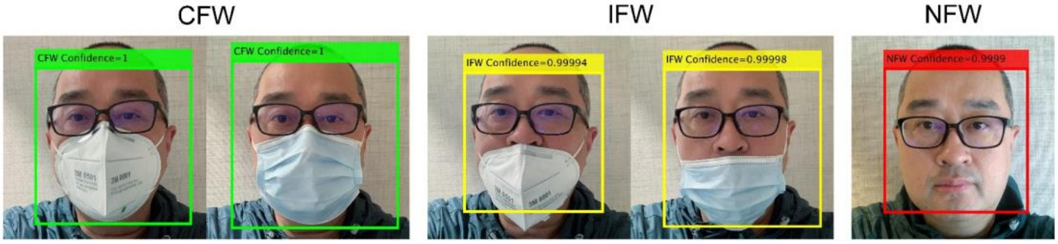

Figure 4. Identification Examples. The labels showed the identification results and confidences of SRCNet. CFW = correct facemask-wearing (green), IFW = incorrect facemask-wearing (yellow), NFW = no facemask-wearing (red).

The confusion matrixes were measured and shown in Fig. 5. The testing dataset contained facial images of NFW, IFW, and CFW. The method we proposed correctly classified 767 images, only 10 prediction errors, which outperformed those without transfer learning or SR network in every category.

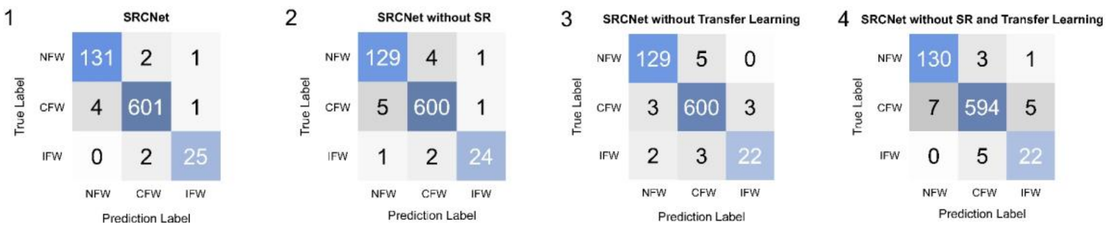

Figure 5. Comparison in Confusion Matrix. NFW = no facemask-wearing, IFW = incorrect facemask-wearing, CFW = correct facemask-wearing.

Discussion and Conclusion

Discussion

Our study presented a novel algorithm to identify facemask-wearing condition, which involved four main steps: image-pre-processing, face detection and crop, SR, and facemask-wearing condition identification. We proposed SRCNet with refined SR network to improve performance with low-quality images.

The study was based on large-scale facial image datasets and Medical Masks Dataset. For SR network, we improved network architecture, including the activation function and density of skip connections. The innovations obtain performance gains considerably in detail enhancement and image restoration compared with previous state-of-art methods, which were evaluated in PSNR and SSIM. For facemask-wearing condition identification, the proposed SRCNet innovatively combined SR network with face identification CNN for performance improvement. Transfer learning was also applied in facemask-wearing condition identification network training. Finally, the SRCNet achieved 98.70% accuracy in three categories classification (NFW, IFW, and CFW), and outperformed traditional end-to-end image classification methods without SR network by over 1.5% in kappa. These findings indicate that by using SR before classification, CNN can achieve higher accuracy. Besides, our experiment proved that deep learning methods could identify facemask-wearing conditions, which have the potential application in epidemic prevention involving COVID-19.

The identification of facemask-wearing conditions has many similarities with facial recognition. However, the development of facemask-wearing condition identification network is challenging for several reasons. The limitation in dataset is one main challenge. The facemask-wearing condition dataset is generally small, and the image quality is not well enough compared with general face recognition dataset. Besides, the various performances of wearing facemask incorrectly largely increase the difficulty of identification. To overcome these challenges, the SRCNet was introduced, which utilized SR network and transfer learning before classification. The SR network solved low-quality image problem, and transfer learning solved the challenge of small dataset with various wearing facemask incorrectly, which get performance improved considerably.

To our knowledge, there have not been any studies in facemask-wearing condition identification with deep learning. In our study, facemask-wearing condition was detected with 98.70% accuracy, which indicated that SRCNet has great potential to support the automatic facemask-wearing condition identification. The design of SRCNet also has consideration in network complexity, which can be applied in public using internet of things (IoTs) and is meaningful to urge the public correctly wearing facemasks for epidemic prevention.

Conclusion

We developed a new facemask-wearing condition identification method in combination with SR network with classification network (SRCNet) in facial image classification. The proposed algorithm contained four main steps: image-pre-processing, face detection and crop, SR, and facemask-wearing conditions identification. The SRCNet achieved 98.70% accuracy and outperformed traditional end-to-end image classification methods using deep learning without SR network by over 1.5% in kappa. Our findings indicate that the proposed that SRCNet could achieve high accuracy in facemask-wearing condition identification, which is meaningful for the prevention of epidemic diseases involving COVID-19 in the public.

Code availability

Code used in the present work, including image processing and network structures, is available at https://github.com/BrightQin/SRCNet.

Acknowledgements

This study was supported by the Student Research Training Program in Zhejiang University.

Author contributions statement

B.Q. designed and performed the experiment, and wrote the manuscript. D.L. guided the study design and provided valuable comments to revise the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

References

1 Chavez, S., Long, B., Koyfman, A. & Liang, S. Y. Coronavirus Disease (COVID-19): A primer for emergency physicians. *Am J Emerg Med*, doi:10.1016/j.ajem.2020.03.036 (2020).

2 Cowling, B. J. *et al.* Facemasks and hand hygiene to prevent influenza transmission in households: a cluster randomized trial. *Ann Intern Med* **151**, 437-446, doi:10.7326/0003-4819-151-7-200910060-00142 (2009).

3 Tracht, S. M., Del Valle, S. Y. & Hyman, J. M. Mathematical modeling of the effectiveness of facemasks in reducing the spread of novel influenza A (H1N1). *PLoS One* **5**, e9018, doi:10.1371/journal.pone.0009018 (2010).

4 Jefferson, T. *et al.* Physical interventions to interrupt or reduce the spread of respiratory viruses. *Cochrane Database Syst Rev*, CD006207, doi:10.1002/14651858.CD006207.pub4 (2011).

5 Sim, S. W., Moey, K. S. & Tan, N. C. The use of facemasks to prevent respiratory infection: a literature review in the context of the Health Belief Model. *Singapore Med J* **55**, 160-167, doi:10.11622/smedj.2014037 (2014).

6 Lai, A. C., Poon, C. K. & Cheung, A. C. Effectiveness of facemasks to reduce exposure hazards for airborne infections among general populations. *J R Soc Interface* **9**, 938-948, doi:10.1098/rsif.2011.0537 (2012).

7 Elachola, H., Ebrahim, S. H. & Gozzer, E. COVID-19: Facemask use prevalence in international airports in Asia, Europe and the Americas, March 2020. *Travel Med Infect Dis*, 101637, doi:10.1016/j.tmaid.2020.101637 (2020).

8 Jefferson, T. *et al.* Physical interventions to interrupt or reduce the spread of respiratory viruses: systematic review. *BMJ* **336**, 77-80, doi:10.1136/bmj.39393.510347.BE (2008).

9 LeCun, Y., Kavukcuoglu, K., Farabet, C. & Ieee. in *2010 Ieee International Symposium on Circuits and Systems IEEE International Symposium on Circuits and Systems* 253-256 (2010).

10 Zhang, K. P., Zhang, Z. P., Li, Z. F. & Qiao, Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *Ieee Signal Proc Let* **23**, 1499-1503, doi:10.1109/Lsp.2016.2603342

(2016).

11 Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the Acm* **60**, 84-90, doi:10.1145/3065386 (2017).

12 Ning, F. *et al.* Toward automatic phenotyping of developing embryos from videos. *IEEE Trans Image Process* **14**, 1360-1371, doi:10.1109/tip.2005.852470 (2005).

13 Cifuentes-Alcobendas, G. & Dominguez-Rodrigo, M. Deep learning and taphonomy: high accuracy in the classification of cut marks made on fleshed and defleshed bones using convolutional neural networks. *Sci Rep* **9**, 18933, doi:10.1038/s41598-019-55439-6 (2019).

14 Mao, X.-J., Shen, C. & Yang, Y.-B. Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections. *arXiv e-prints*, arXiv:1606.08921 (2016). <https://ui.adsabs.harvard.edu/abs/2016arXiv160608921M>.

15 Zhang, Y., Tian, Y., Kong, Y., Zhong, B. & Fu, Y. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2472-2481 (2018).

16 Bruzzone, L. *et al.* in *Image and Signal Processing for Remote Sensing XX* (2014).

17 Thomas, R. & Rangachar, M. J. S. Fractional Bat and Multi-Kernel-Based Spherical SVM for Low Resolution Face Recognition. *International Journal of Pattern Recognition and Artificial Intelligence* **31**, doi:Artn 1756014

10.1142/S0218001417560146 (2017).

18 Liang, M. M. *et al.* Deep Multiscale Spectral-Spatial Feature Fusion for Hyperspectral Images Classification. *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **11**, 2911-2924, doi:10.1109/Jstars.2018.2836671 (2018).

19 Zhu, X. B., Li, Z. Z., Li, X. B., Li, S. S. & Dai, F. Attention-aware perceptual enhancement nets for low-resolution image classification. *Information Sciences* **515**, 233-247, doi:10.1016/j.ins.2019.12.013 (2020).

20 Dong, C., Loy, C. C., He, K. & Tang, X. 184-199 (Springer International Publishing).

21 Dong, C., Change Loy, C. & Tang, X. Accelerating the Super-Resolution Convolutional Neural Network. *arXiv e-prints*, arXiv:1608.00367 (2016). <https://ui.adsabs.harvard.edu/abs/2016arXiv160800367D>.

22 Dong, C., Loy, C. C., He, K. & Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans Pattern Anal Mach Intell* **38**, 295-307, doi:10.1109/TPAMI.2015.2439281 (2016).

23 Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *arXiv e-prints*, arXiv:1801.04381 (2018). <https://ui.adsabs.harvard.edu/abs/2018arXiv180104381S>.

24 Howard, A. G. *et al.* MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv e-prints*, arXiv:1704.04861 (2017). <https://ui.adsabs.harvard.edu/abs/2017arXiv170404861H>.

25 He, K., Zhang, X., Ren, S. & Sun, J. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770-778 (2016).

26 Gurovich, Y. *et al.* Identifying facial phenotypes of genetic disorders using deep learning. *Nat Med* **25**, 60-64, doi:10.1038/s41591-018-0279-0 (2019).

27 Hinton, G. E., Osindero, S. & Teh, Y. W. A fast learning algorithm for deep belief nets. *Neural Comput* **18**, 1527-1554, doi:10.1162/neco.2006.18.7.1527 (2006).

28 Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504-507, doi:10.1126/science.1127647 (2006).

29 Kwak, G. H. *et al.* Automatic mandibular canal detection using a deep convolutional neural network. *Sci Rep* **10**, 5711, doi:10.1038/s41598-020-62586-8 (2020).

30 Yang, S., Luo, P., Loy, C.-C. & Tang, X. in *2015 IEEE International Conference on Computer Vision (ICCV)* 3676-3684 (2015).

31 Yi, D., Lei, Z., Liao, S. & Li, S. Z. Learning Face Representation from Scratch. *arXiv e-prints*, arXiv:1411.7923 (2014). <https://ui.adsabs.harvard.edu/abs/2014arXiv1411.7923Y>.

32 Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, arXiv:1412.6980 (2014). <https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6980K>.

33 Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv e-prints*, arXiv:1409.1556 (2014). <https://ui.adsabs.harvard.edu/abs/2014arXiv1409.1556S>.

34 Lin, M., Chen, Q. & Yan, S. Network In Network. *arXiv e-prints*, arXiv:1312.4400 (2013). <https://ui.adsabs.harvard.edu/abs/2013arXiv1312.4400L>.

35 Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv e-prints*, arXiv:1207.0580 (2012). <https://ui.adsabs.harvard.edu/abs/2012arXiv1207.0580H>.

36 Vedaldi, A. & Lenc, K. in *Proceedings of the 23rd ACM international conference on Multimedia - MM '15* 689-692 (2015).

37 Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* **13**, 600-612, doi:10.1109/tip.2003.819861 (2004).

38 Tai, Y., Yang, J., Liu, X. & Xu, C. in *2017 IEEE International Conference on Computer Vision (ICCV)* 4549-4557 (2017).