

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Masked face recognition with mask transfer and self-attention under the COVID-19 pandemic

MENG, ZHANG^{1,2}, RUJIE LIU², DAISUKE DEGUCHI¹, (Member, IEEE), and HIROSHI MURASE¹, (Life Fellow, IEEE)

¹Graduate School of Informatics, Nagoya University, Nagoya, Japan

²Fujitsu Research and Development Center, Beijing, China

Corresponding author: Meng, Zhang (zhang.meng@f.mbox.nagoya-u.ac.jp).

ABSTRACT Face masks bring a new challenge to face recognition systems especially against the background of the COVID-19 pandemic. In this paper, a method mitigating the negative effects of mask defects on face recognition is proposed. Firstly, a low-cost, accurate method of masked face synthesis, i.e. mask transfer, is proposed for data augmentation. Secondly, an attention-aware masked face recognition (AMaskNet) is proposed to improve the performance of masked face recognition, which includes two modules: a feature extractor and a contribution estimator. Therein, the contribution estimator is employed to learn the contribution of the feature elements, thus achieving refined feature representation by simple matrix multiplications. Meanwhile, the end-to-end training strategy is utilized to optimize the entire model. Finally, a mask-aware similarity matching strategy (MS) is taken to improve the performance in the inference stage. The experiments show that the proposed method consistently outperforms on three masked face recognition datasets: RMFRD [1], COX [2] and Public-IvS [3]. Meanwhile, qualitative analysis experiments using CAM [4] indicate that the contribution learned by AMaskNet is more conducive to masked face recognition.

INDEX TERMS Masked face recognition, mask transfer, face synthesis, attention mechanism.

I. INTRODUCTION

The COVID-19 pandemic has created a global disaster: the World Health Organization (WHO) and Centers for Disease Control and Prevention (CDC) have suggested everyone should wear a mask in a public setting especially when other social distancing measures are difficult to maintain [4]. Face recognition is non-contact, highly efficient, user friendly, and so forth, and has been widely used in access control and security authentication in public places; however, masks bring a new challenge to existing commercial face recognition techniques. Face recognition becomes more difficult when a large part of the face is covered by a mask. Therefore, it is essential to study the effect of wearing face masks on the behavior of face recognition systems and design mitigation techniques to offset the inevitable performance loss.

Deep-learning-based approaches have predominated in the field of face recognition due to the emergence of advanced convolutional neural networks and large-scale datasets [5], [6], [7]. Despite the success of deep learning models under general face recognition scenarios, the deep features still show imperfect invariance to wearing a mask, where the

whole face image can't be provided for description. Therefore, the use of face masks triggers a significant research challenge: firstly, it is necessary to collect a large-scale training dataset, which includes the different types of faces with masks. To collect such a large-scale training dataset, on the one hand, it is time consuming and incurs higher labor cost, on the other hand, maintaining the diversity of data in such datasets is a slow process. Therefore, a low-cost, convenient face data augmentation method is needed as a matter of urgency. Secondly, it is necessary to mitigate the performance loss from the perspective of model design according to the characteristics of face masks.

Some methods of simulating a masked face [8], [9], [10], [11] have been proposed for face data augmentation. MaskTheFace [10] used a Dlib-based [12] face landmark detector to identify facial tilt and six key features of the face necessary for applying a mask. MaskedFace-Net [8], defined a mask-to-face deformable model and applied homographic transformation to map mask pixels over the targeted facial areas. However, these methods only use affine transformation, and the added mask often looks unnatural; furthermore,

those methods ignore pose and illumination consistency thus leading to biased masked face augmentation. Recently, Generative Adversarial Networks (GAN) has become a powerful technique to data augmentation [13]. However, on the one hand, GAN-based method suffers from mode collapse deeply, which usually manifests as that the images generated by generators tend to have high similarity amongst them, even though their corresponding latent vectors have been very different. On the other hand, GAN-based methods are generally slow and difficult to run online in recognition. On the contrary, our method can quickly collect various types of mask images and can be transferred to the face image in run-time for the mask-aware similarity matching strategy in inference stage.

Occlusion is a common problem in computer vision, and numerous approaches have been proposed to handle this problem [2], [14]. Wearing a mask is considered the most difficult facial occlusion challenge since it covers most of the face including the mouth and nose. Anwar et al. [10] came up with an open-source tool (MaskTheFace), to create a large dataset of masked faces, and then re-train existing face recognition systems to improve their accuracy. To decrease the negative influence of masks, in [15], the masked region is directly discarded when extracting deep features. Mundial, et al. [16] used a supervised learning approach and an in-depth neural network to recognize masked faces and extract individual facial features, with which an SVM classifier was built for classification purposes.

The covered facial areas contain a lot of salient features that make it useful for face recognition, such as the nasal region [17], but the extracted features of the covered facial areas are damaged due to occlusions caused by mask. Therefore, compared with uncovered facial areas such as the eye region, intuitively speaking, the mask area does not contain much discriminative information useful in recognizing a face, which gives us the hint that more attention should be paid to the uncovered region in feature extraction. Recently, attention mechanisms have been introduced to video face recognition systems [18], [19], where an attention mechanism is used to mimic human's perception to focus on important information. Inspired by those works, an attention-aware network (AMaskNet) is proposed for masked face recognition, which puts more weight to useful features while (in relative terms) ignoring those corrupted by the face mask by learning a contribution matrix. Indeed, AMaskNet can localize the salient facial areas and extract more discriminative features for non-masked face image and alleviates the performance degradation of the non-masked scene. Even in low-quality face recognition scenes, the recognition performance of without wearing a mask is even slightly improved.

Herein, in this paper, we qualitatively and quantitatively analyse the effect of wearing a mask on face recognition, and then propose a method for mitigating the negative effects of mask defects on face recognition. Firstly, a low-cost, accurate method of mask transfer is proposed for masked face synthesis by considering pose and illumination consistency.

Secondly, an attention-aware masked face recognition model (AMaskNet) is designed to improve the performance of masked face recognition. The AMaskNet includes two modules, feature extractor and contribution estimator, wherein the latter is used to learn the contribution of each spatial region which is then combined with the feature to improve its representation capability. An end-to-end training strategy is adopted to optimize the whole network. Finally, a mask-aware similarity matching strategy is put forward to improve the performance in the inference stage. The experiments show that the proposed method consistently outperforms on three masked face recognition datasets: RMFRD [1], COX [2] and Public-IvS [3]. Meanwhile, qualitative analysis experiments using CAM [20] indicate that the contribution learned by the AMaskNet is beneficial to masked face recognition. The main contribution can be summarized thus:

(1) By the consideration of pose and illumination consistency, a low-cost, accurate mask transfer method for masked face data augmentation was put forward, which can add the mask from any face image with a mask to any face image without a mask.

(2) Qualitative and quantitative experiments were conducted to analyze the effect of wearing face masks on the behavior of face recognition systems.

(3) An attention-aware masked face recognition (AMaskNet) was proposed to improve the performance of masked face recognition.

(4) A mask-aware similarity matching strategy was proposed for the inference stage, which can be applied to any face recognition scene in which one image with a mask and the other without a mask are present.

The paper is organized as follows: in Section 2 related work on masked face recognition is reviewed. Section 3 describes the proposed method of mask transfer, AMaskNet, and mask-aware similarity matching strategy. Section 4 presents the experimental results and conclusion is given in Section 5.

II. RELATED WORK

A. METHODS OF SIMULATED MASKED FACE IMAGE

Recently, some methods of simulated masked face image have been proposed [8], [9], [10], [11]. MaskTheFace [10] used a Dlib-based [12] face landmark detector to identify the face tilt and six key features of the face necessary for applying a mask. MaskedFace-Net [8] defined a mask-to-face deformable model and applied homographic transformation for mapping mask pixels over the targeted facial areas. Firstly, feature-based cascade classifiers are applied to detect a region of interest in the facial image, with which a key-point detector is applied to realize automatic detection of 68 landmarks representing the facial structure. Besides, an image of a conventional face mask is selected as a reference image for the mapping where 12 key points are manually annotated for delineating the mask area. Finally, a homographic transformation is applied to map mask pixels over the targeted facial areas relying on the defined point-to-point

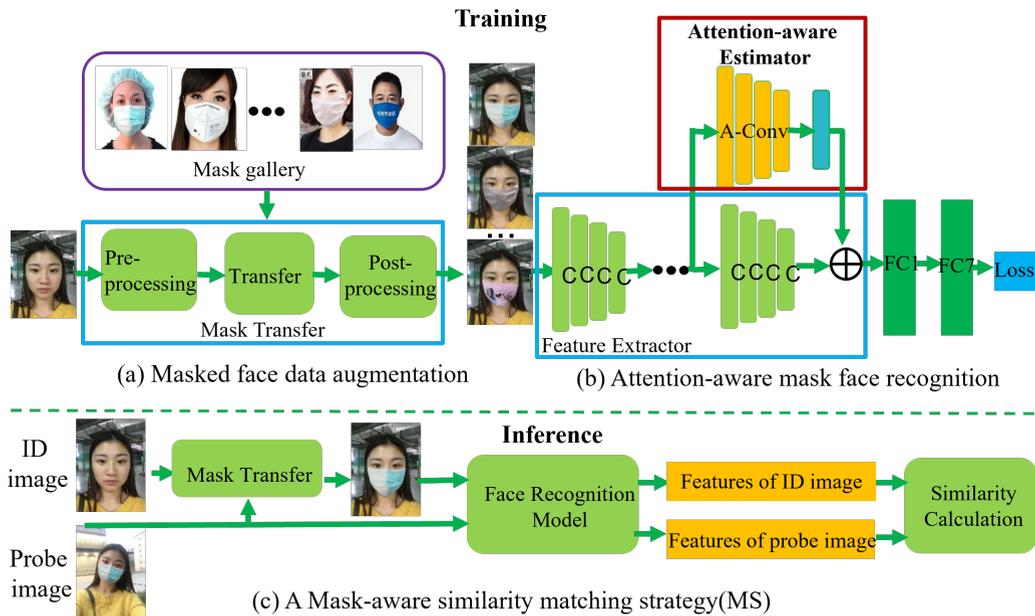


FIGURE 1. The architecture of our proposed method, which include the method of mask transfer, attention-aware mask face recognition (AMaskNet), and a mask-aware similarity matching strategy for inference. (Best viewed in color)

correspondence of landmarks between mask image and face image.

B. MASKED FACE RECOGNITION

Many different deep-learning-based approaches have been proposed to solve the occlusion problem. In 2014, Sun et al. [21] discovered that the features learned by DeepID2+ show certain robustness to image corruption in face verification tasks, and the combination of DeepID2+ features extracted from 25 face patches may further improve the robustness. Daniel et al. [14] used the augmenting training data with synthetic occluded faces to tackle the occlusion problem. Recently, the Masked face recognition have attracted more and more attention during the COVID-19 pandemic. Aqeel Anwar et al. [10] proposed an open-source tool, named Mask-TheFace, to create masked face dataset from face dataset with extended feature support, and then used this dataset to re-train existing face recognition engines to improve their accuracy. Hariri and Walid [15] developed a reliable method based on occlusion removal and deep learning-based features in order to address the problem of the masked face recognition process. The first step is to remove the masked face region. Next, three pre-trained deep Convolutional Neural Networks (CNN) namely, VGG-16, AlexNet, and ResNet-50, are used to extract deep features from the obtained regions (mostly eyes and forehead regions). The Bag-of-features paradigm is then applied to the feature maps of the last convolutional layer in order to quantize them and to get a slight representation comparing to the fully connected layer of classical CNN. Finally, Multilayer Perceptron (MLP) is applied for the classification process. Mundial et al. [16] used a supervised

learning approach for masked face recognition together with in-depth neural network-based facial features. First, a CNN model was trained to generate an embedding of features of an image. Then a dataset which helps in building classifier for masked face was focused, which consists of three images of a person, two masked face images, and one without a face mask. In the end, the Support Vector Machine (SVM) is used for classification.

C. ATTENTION MECHANISMS IN FACE RECOGNITION

An attention mechanism is used to mimic human attention, which can concentrate on important information [22], [23]. Recently, attention mechanisms have been introduced to video face recognition [18], [19]. A meta attention-based aggregation scheme is adopted in [19], to fine-grain the weights in an adaptive manner along each feature dimension among all frames to handle the features on a dimensional level. Rao et al. [18] used an attention-aware deep reinforcement learning method to discard misleading and confounding frames and find the focus of attention in video footage of faces.

III. PROPOSED APPROACH

In this section, the mask transfer method for masked face synthesis is firstly described by two stages, one is the construction of a mask gallery and another is the generation of synthetic masked face images (Figs. 1(a) and 2). Secondly, the proposed AMaskNet is introduced. As shown in Fig.1(b), AMaskNet incorporates feature extractor and contribution estimator, and the contribution estimator further consists of two sub-modules, i.e., a self-spatial contribution estimator and a self-channel contribution estimator. The two sub-modules are

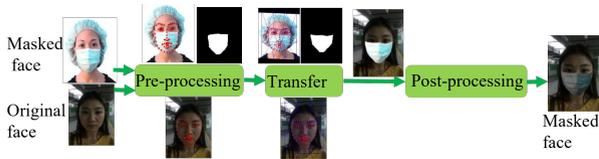


FIGURE 2. The flowchart of mask transfer. The masked face is a photo randomly selected from the mask gallery.

used to learn the spatial contribution and channel contribution respectively, and the refined features are obtained by combining the two contributions. Moreover, the optimization of the entire network can be realized through an end-to-end training procedure. Finally, a mask-aware similarity matching strategy is introduced for inference purposes, as illustrated in Fig. 1(c).

A. MASK TRANSFER (MT)

To generate the synthetic masked face image, a gallery of different masks should be firstly constructed. Given a non-masked face image and one mask from the gallery, the masked face image is obtained by transfer the mask, including pre-processing, transfer, and post-processing, as illustrated in Fig. 2.

1) Collection of mask gallery

The construction of mask gallery is to obtain a face image set covering versatile masks, such as different mask color, shape, and texture. For one mask type, only one face image with this mask is enough, therefore, it is an easy thing to build this dataset through collection from website. Generally, any face image with a mask is qualified for the collection. However, to improve the quality of synthetic masked face images, the frontal view face images are preferable.

2) Mask transfer for masked face synthesis

Pre-processing. Dlib [12] is used to detect 68 landmark points in both masked and non-masked face images. With these landmarks, a Triangulated Irregular Network (TIN) is constructed and the facial area is thus divided into multiple triangular regions. Meanwhile, the grab-cut method [24] is employed to segment mask areas from masked face images.

Transfer. Transfer mask from masked face images to non-masked ones should be implemented based on the geometric relationship of the two faces. For each triangular piece in the TIN model, the affine transformation between the two images is calculated, and the mask region contained in this piece is transformed directly to a non-masked image. The whole mask will be transferred after all triangular pieces have been transformed.

Post-processing. Due to the lighting and contrast difference of the two images, directly transfer the mask usually leads to inconsistency of illumination in the target image. For this problem, two post-processing steps are performed after the mask is transferred to the target image: (1) Alpha-

matting is used to make the boundary more natural in terms of the transition across the boundary; (2) histogram specification is adopted to make the transferred mask region more illumination-consistent with the original non-masked face image. More specifically, the gray-scale distribution of the mask region is adjusted according to the gray-scale histogram of the original non-masked image, as shown in Fig. 7.

B. ATTENTION-AWARE NET FOR MASKED FACE RECOGNITION (AMASKNET)

The network architecture consists of two modules: a feature extractor and a contribution estimator, as illustrated in Fig. 3. A conventional face recognition scheme, e.g. ResNet34 [5], may be used for feature extraction, yielding a feature vector as the initial face representation. As mentioned previously, the face mask does not contain much subject-related information, so the extracted feature is less discriminative. For this problem, the contribution estimator is designed to learn a contribution matrix to assign more weight to useful features while (in relative terms) ignoring those corrupted by the face mask. The contribution estimator is an attention scheme and is implemented via a branch structure, which includes both spatial and channel components

1) Feature extractor

A conventional face recognition scheme (e.g., ResNet) can be utilized as a backbone for extraction of features, which is used as the baseline model. Recently, the ArcFace [5]) has achieved the state-of-the-art performance and has been widely used in many papers. Here, we firstly adopt the Arc-Face34 as our backbone, where BN-Conv-BN-PReLU-Conv-BN module as the residual bottleneck and all the convolution kernel size in residual bottlenecks have a size of 3×3 . Then, the output feature of all models is fixed to 512-dimension by a fully connected layer.

2) Attention-aware contribution estimator

The contribution estimator covers both spatial- and channel-wise measurements. They adaptively aggregate the feature maps in both channel and spatial domains to learn the inter-channel relationship and interspatial relationship matrixes. The two matrixes are then multiplied with the initial feature representation to produce refined face features. The sigmoid function is utilized instead of the ReLU [25] to map the output onto the interval $[0,1]$, which is used as a contribution coefficient for weighting features. An end-to-end training strategy is used to optimize the entire network.

Spatial contribution estimator. A branch architecture is added to the backbone of the feature extractor model as the spatial contribution estimator, attempting to estimate the contribution for each component of the feature map, as shown in Fig. 3(b). In this process, a contribution matrix is learnt which can assign greater weights to useful features while (in relative terms) ignoring those corrupted by the face mask. The contribution matrix and feature map have the same width and height. The structure of the spatial contribution estimator

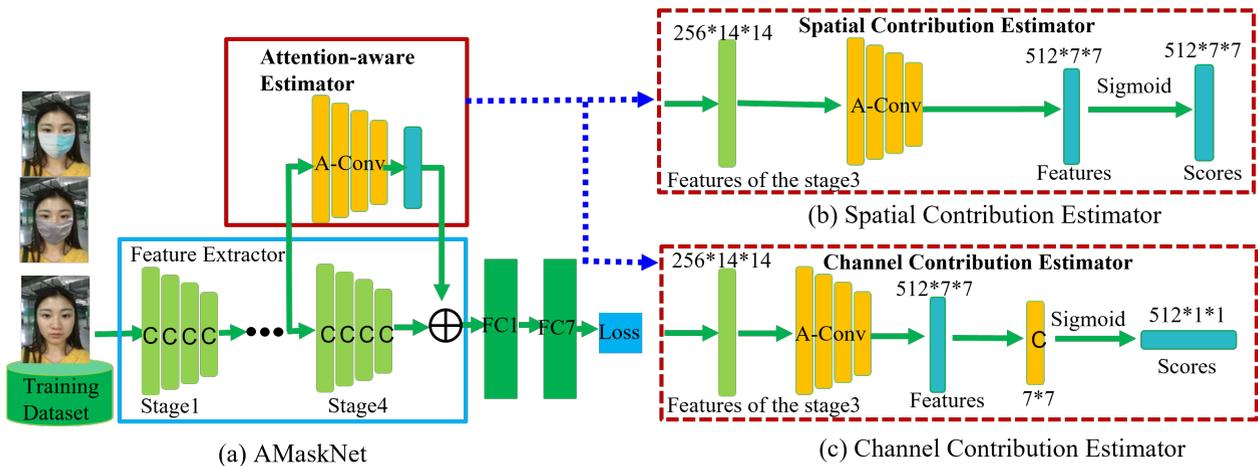


FIGURE 3. The architecture of AMaskNet. (b) is the details of the spatial contribution estimator. (c) is the details of the channel contribution estimator. The inputs of (a) and (b) are the outputs of a layer in the middle of the backbone network. The backbone is divided into four stages. Here we take the output of the third stage as an example. The network structure of A-Conv is the same as that of the stage4 of backbone of the AMaskNet. The sigmoid function is utilized to map the output onto the interval $[0, 1]$, which is used as a contribution coefficient for weighting features. (Best viewed in color)

may have different complexities, ranging from one to several convolution layers. A more complex network may result in a better ability to learn, albeit at the cost of the extra computational effort and the risk of overfitting.

Channel contribution estimator. Akin to the spatial contribution estimator, a branch architecture is further added as the channel contribution estimator, expecting to estimate the contribution for each channel of the feature map, as shown in Fig. 3(c). In this module, a contribution matrix will be learnt to put more attention on useful channels. Similarly, different structures of various complexity may be adopted.

3) Feature aggregator

Let the feature extracted by feature extractor be $F_I \in R^{C \times H \times W}$, the spatial and channel contribution matrixes be $C_S \in R^{C \times H \times W}$ and $C_C \in R^C$ respectively, the final feature is derived as:

$$F_C = C_C \otimes (C_S \oplus F_I) \quad (1)$$

where, \otimes denotes matrix multiplication, \oplus represents element-wise multiplication.

4) Training strategy

In model training, the ArcFace loss [5] is used to penalize identification errors. The bias is fixed as $b_j = 0$, logit is transformed as $W_j^T F_R = \|W_j^T\| \|F_R\| \cos \theta_j$, where θ_j is the angle between the weight W_j and the refined feature F_R . The individual weight is fixed as $\|W_j\| = 1$ by L_2 normalization. The refined feature is fixed as $\|F_R\|$ by L_2 normalization and is re-scaled to s . This normalization makes the prediction probability only depend on the angle between

the feature and the weight. So, the loss can be formulated as:

$$L_{id} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(\theta_{y_i} + m))}}{e^{s(\cos(\theta_{y_i} + m))} + \sum_{j=1, j \neq y_i}^N e^{s(\cos \theta_j)}} \quad (2)$$

Here, m is the additive angular margin penalty between F_R and W_{y_i} to simultaneously enhancing the intra-class compactness and inter-class discrepancy.

C. MASK-AWARE SIMILARITY MATCHING STRATEGY(MS)

In the real application, the two face photos for matching are usually in different style, that is, the ID or mugshot photos coming from gallery are front portrait images without a mask, while the probe images captured on site sometimes are with a mask. Obtaining effective features by focusing more on the non-masked facial region is helpful in such situation, however, it still cannot eliminate the loss of accuracy caused by the presence of a mask. One straightforward method is to extract features only from upper facial regions when comparing two images with and without masks, but some important information will be neglected, e.g., shape information contained in the mask region. To solve this problem, a mask-aware similarity matching strategy(MS) is proposed, as illustrated in Fig. 1(c). This involves transfer the mask from the masked face image to a non-masked image, thus mitigating the difference caused by the mask without loss of spatial information. This method can be applied to any face recognition scene in which one image with a mask and the other without a mask is presented, especially useful for the 1:1 face verification scene.

IV. EXPERIMENT AND RESULTS

In this section, several benchmark datasets and several baseline models are firstly introduced. Then, some qualitative and

quantitative analyses are presented to attain more insight into how the mask affects the performance of face recognition. Finally, the proposed mitigating models are compared with top-performing public models to confirm the effectiveness of our method.

A. DATASETS AND PROTOCOL

1) TRAINING DATASET

DeepGlint [26] is used as the training corpus. It includes cleaned MS-Celeb-1M [6] and the celebrity Asia dataset, achieving totally 6.6 million celebrity images of 172,000 celebrities therein. Due to the lack of a large volume of masked face photos to train the model, data augmentation is used for synthesizing masked face images by the proposed mask-transfer technique. For each training image in DeepGlint, one mask image is randomly selected from the mask gallery and the mask is transferred to this training image. With this manner, the training data is doubled, resulting in a quantity of about 13M photos.

2) TESTING DATASET

Several commonly used benchmark datasets, such as RMFRD [1], COX [2], and Public-IvS [3], are used for testing.

COX [2] comprises 1,000 still images and 3,000 videos of 1,000 subjects. The video footage is captured using three cameras at different locations while the subjects are walking in a large gymnasium to simulate a surveillance scenario. The video-to-still (V2S) protocol proposed by the author is adopted for performance evaluation, where the true acceptance rate ($TAR@FAR = 10^{-4}$) is used for the 1:1 verification.

Public-IvS [3] designed for ID vs. spot recognition, contains 1,262 identities and 5,503 images. The true acceptance rate ($TAR @ FAR = 10^{-5}$) for the 1:1 verification protocol is used to evaluate its performance.

To analyze the effectiveness of the proposed method on masked face recognition, the following four test conditions are designed to add masks to face photos given the test image pair in COX and Public-IvS, as shown in Fig. 4 and Fig. 5:

- (1) **Wo/Wo**: no mask, original test images are used.
- (2) **W/W**: different masks are added to both images in a pair to simulate a scene where both gallery and probe images are of masked faces.
- (3) **Wo/W**: the recognition scene where gallery images do not contain masked faces, but probe images do. To simulate this condition, a mask is added to COX video frames (Public-IvS spot image) while keeping the COX still images (Public-IvS ID image) unchanged.
- (4) **Wo/W+MS**: to improve the accuracy of **Wo/W**, the mask of the COX video frames (Public-IvS spot image) is transferred to COX still images (Public-IvS ID image), i.e., the same masks are guaranteed to appear in each image pair.

RMFRD [1] is crawled from the Internet, including 5,000 pictures of 525 people wearing masks, and 90,000 images of the same 525 subjects without masks, which is mainly devoted to evaluate the existing face recognition system

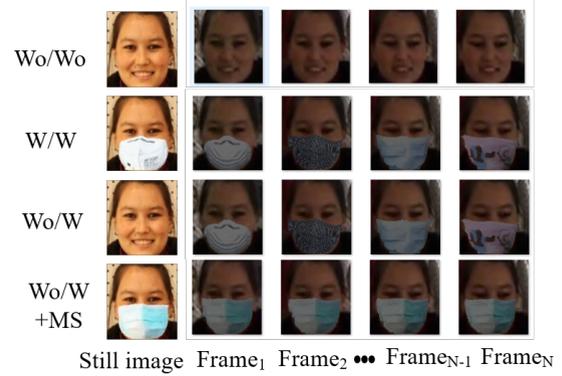


FIGURE 4. Some pairs of face image from the COX and COX -mask. Here, **Wo/Wo**: no mask, original test images are used. **W/W**: different masks are added to both images in a pair. **Wo/W**: the recognition scene where gallery images do not contain masked faces, but probe images do. **Wo/W+MS**: to improve the accuracy of **Wo/W**, the mask of the COX video frames (Public-IvS spot image) is transferred to COX still images (Public-IvS ID image), i.e., the same masks are guaranteed to appear in each image pair.



FIGURE 5. Some pairs of face image from the Public-IvS and Public-IvS-mask (ID image vs. Spot image).



FIGURE 6. Some pairs of face image from the RMFRD: face images without a mask(up) and with a mask(down).

on masked images during the COVID-19 pandemic. Some sample images are shown in Fig. 6. The COX and private-IvS datasets are masked by the proposed mask transfer method, which is convenient to analyze the impact of masks on the performance of the model. However, in order to evaluate the performance of our model on the real-world masked face recognition, we also conducted experiments on RMFRD dataset and compared with other state-of-the-art methods.

B. FACE RECOGNITION MODEL AND IMPLEMENTATION

1) Face recognition model

The ArcFace model [5], which achieves state-of-the-art performance against several face recognition benchmarks such as LFW and YTF, is selected for comparison. Ar-

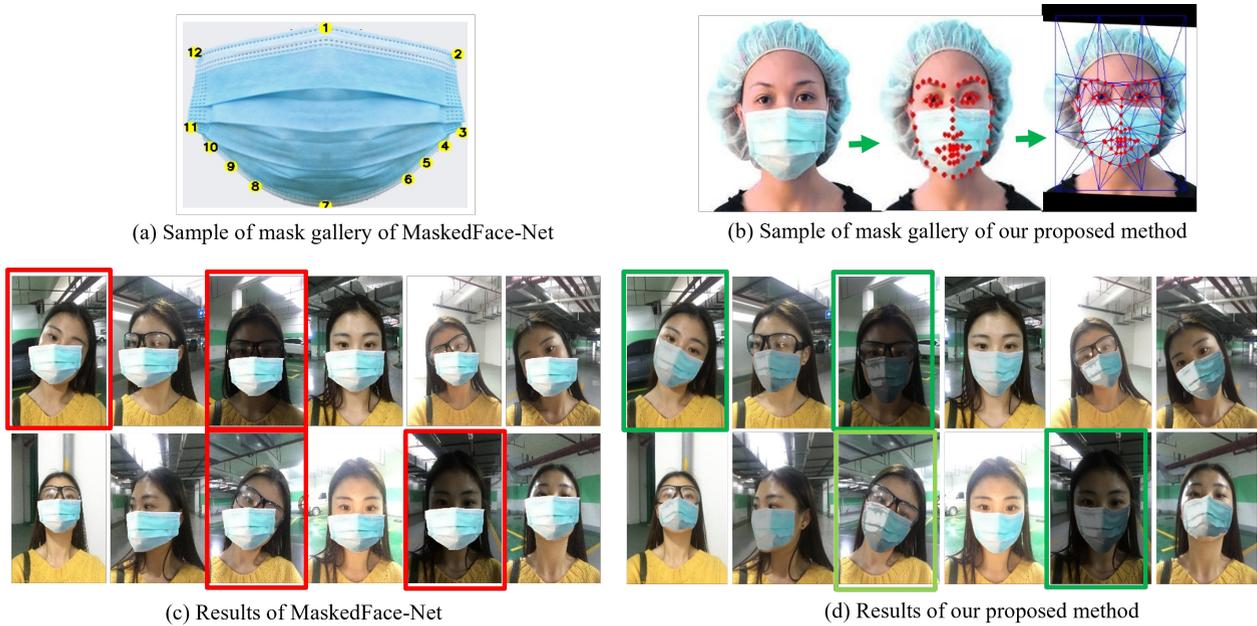


FIGURE 7. Comparison of MaskedFace-Net [7] and our method. MaskedFace-Net requires manually labeling several key points on the mask boundary (a), while our proposed method automatically extract the mask region from masked face image (b). Exemplar results of adding mask are shown in (c) and (d) respectively.

cFace introduces additive angular margin loss to enhance the discriminative power of the face recognition model, therefore, it is robust to the condition of wearing a mask. Four publicly available models, MobileFaceNet (ArcFaceM), LResNet34E-IR (ArcFace34), LResNet50E-IR (ArcFace50), and LResNet100E-IR (ArcFace100) are adopted. These models were downloaded from <https://github.com/deepinsight/insightface/wiki/Model-Zoo>. As introduced previously, the ResNet34 is selected in our method as the backbone for feature extraction, with which the DeepGlint without mask data augmentation is used to build a baseline model to study the effect of face masks. To mitigate the negative effects of mask, two variants based on ResNet34 are built including data augmentation and the proposed attention scheme, as follows:

(1) R34-Baseline. A ResNet-34 model is trained with the original DeepGlint dataset, to quantify the loss of accuracy the mask may induce.

(2) R34-Mask. A ResNet-34 model is trained by only the mask augmented DeepGlint dataset. This model can significantly improve the performance of masked-face recognition, but it will degrade the performance of non-masked face recognition.

(3) R34-AMaskNet. This is the proposed AMaskNet model trained with combination of the mask augmented DeepGlint and the original DeepGlint, which is expected to improve the performance of masked-face recognition while reducing the loss of accuracy inherent to non-masked face recognition.

2) Implementation

The ArcFace loss [5] is used to train the model, where the feature scale s is set to 64 and the $arccos$ margin m is set to 0.5. In training, stochastic gradient descent is adopted with momentum and weight decay values of 0.9 and 0.0005, respectively. The training begins with a learning rate 0.1 for seven epochs, which is then decreased every five epochs by a factor of 10. Totally 25 epochs are taken for the training with the augmented DeepGlint dataset, and 760 images are used in each mini-batch.

C. EFFECTIVENESS OF THE PROPOSED METHOD

1) Effectiveness of the proposed mask transfer for masked face synthesis

Fig. 7 presents some mask transfer examples. Figs 7(a) and 7(b) show the mask gallery of the traditional method and proposed method. The proposed method is simply a face image with a mask, which is easy to obtain and does not need manual annotation. It is low-cost, rapid, and convenient for model development. Figs 7(c) and 7(d) compare respectively the synthesized mask face images of the traditional method and the proposed method. From the results, it can be observed that the proposed mask transfer method is effective and can maintain consistency of illumination.

2) Effectiveness of the proposed AMaskNet

Results on COX. As shown in Fig. 8, by re-training the R34-Baseline using a masked augmented dataset, the R34-Mask model gain a significant improvement on images containing mask, e.g. a 28.9 percentage points improvement in Wo/W on Cam1 of COX; however, this method is likely to cause per-

Model	Results on Cam1 of COX			Results on Cam2 of COX			Results on Cam3 of COX		
	Wo/Wo	W/W	Wo/W	Wo/Wo	W/W	Wo/W	Wo/Wo	W/W	Wo/W
ArcFaceM	85.50	42.50	33.30	74.90	38.70	29.60	93.70	51.90	41.10
ArcFace34	94.20	58.60	51.80	86.10	52.10	47.50	86.10	52.10	47.50
ArcFace50	95.20	60.40	55.80	87.60	50.80	47.20	98.20	65.60	62.50
ArcFace100	97.60	72.00	66.60	94.80	62.20	57.50	99.20	75.50	72.70
R34-Baseline	98.00	70.60	64.90	95.40	65.70	59.20	99.30	77.30	74.40
R34-Mask	96.40	93.40	93.80	93.20	90.70	90.70	99.00	94.50	94.30
R34-AMaskNet	98.60	93.90	93.50	96.40	91.80	92.00	99.50	94.30	94.10

FIGURE 8. Results on COX dataset with 1:1 verification protocol at $TAR@FAR=10^{-4}$. Here, Wo means without wearing a mask, W means with wearing a mask.

Model	Wo/Wo	W/W	Wo/W
	ArcFaceM	91.40	27.80
ArcFace34	95.50	47.80	23.40
ArcFace50	95.70	57.90	53.80
ArcFace100	96.20	70.00	67.50
R34-Baseline	96.00	63.80	59.70
R34-Mask	93.60	93.90	93.90
R34-AMaskNet	96.00	94.60	95.40

FIGURE 9. Results on Public-IvS dataset with 1:1 verification protocol at $TAR@FAR=10^{-5}$. Here, Wo means without wearing a mask, W means with wearing a mask.

formance degradation in terms of general face recognition, e.g. a 1.6 percentage points decline in the case of Wo/Wo for the Cam1 of COX. The comparison between R34-AMaskNet and R34-Mask shows that AMaskNet is able to improve the performance, especially for masked face recognition, e.g. a 1.1 percentage points improvement in the case of W/W on the Cam2 of COX, which indicates that the proposed contribution estimator can learn an effective contribution matrix and automatically assign higher weights to the feature map activated by the non-masked facial parts and lower weights to those that are activated by masked facial parts. Meanwhile, the performance of R34-AMaskNet undergoes no significant decline and may even be slightly improved in the case of Wo/Wo. This is because COX is a low-quality video face recognition dataset with dramatic illumination and motion blur. However, AMaskNet can localize the salient facial areas and puts more weight to discriminative features, thus improving the performance of wearing masks while minimizing the effect of general face recognition on the existing face systems.

Results on Public-IvS. As shown in Fig. 9, from the results on Public-IvS, a similar conclusion to the COX can be obtained. Although R34-Mask can improve the performance of masks in the case of Wo/W or W/W, it will degrade the performance in the case of Wo/Wo. On the contrary, the R34-AMaskNet improves masked face recognition performance with a little cost of performance decrease on general face recognition.

3) Effectiveness of the proposed mask-aware similarity matching strategy (MS)

Results on COX. Table 1 shows the recognition result of without and with mask transfer on the Wo/W conditions (Wo/W vs. Wo/W+MS), where one contains a mask, while the other does not. In addition to R34-Mask and R34-AMaskNet models, the performance of the models are greatly improved after using the proposed mask-aware similarity matching strategy, e.g., a 11.3 percentage points improvement for ArcFaceM on the Cam2 of COX between images treated without mask-transfer to those with mask-transfer (because there were no mask strategy). Meanwhile, for R34-Mask and R34-AMaskNet, although the data augmentation strategy has been adopted, it is still improved in most cases, e.g., a 0.7 percentage points improvement for the R34-AMaskNet on the Cam2 of COX between images treated without mask-transfer to those with mask-transfer.

Results on Public-IvS. The comparison between without mask transfer (Wo/W) and with mask transfer (Wo/W+MS) on the Wo/W (one is with mask, while other is without mask) in Table 2 shows that the recognition performance is improved by the use of the mask-aware similarity matching strategy (MS), especially for the general face recognition model, e.g. a 0.4 percentage points improvement for the R34-Mask, compared to a 14 percentage points improvement for ArcFaceM on the Public-IvS dataset from without mask transfer to with mask transfer. The results shows that transfer the mask from masked image to non-masked image in the face pairs can mitigate the difference caused by the mask without loss of

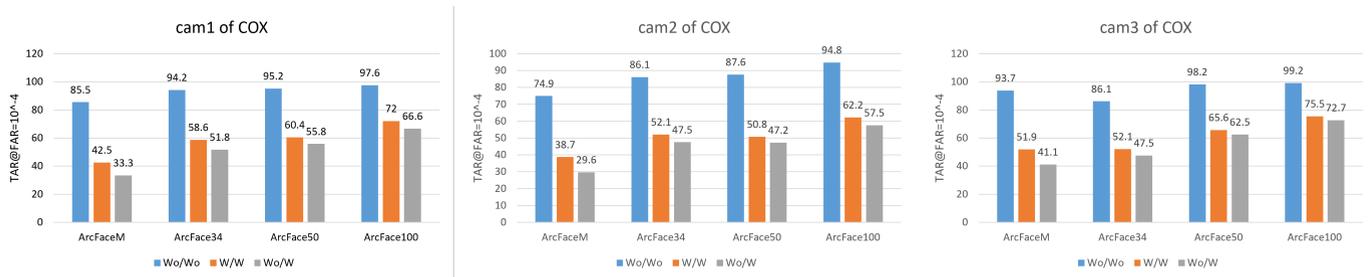


FIGURE 10. Results on COX dataset with 1:1 verification protocol at $TAR@FAR=10^{-4}$. Here, Wo means without wearing a mask, W means with wearing a mask.

TABLE 1. Results on COX dataset with 1:1 verification protocol at $TAR@FAR=10^{-4}$.

Model	Cam1		Cam2		Cam3	
	×	✓	×	✓	×	✓
ArcFaceM	33.3	41.4	29.6	40.9	41.1	55.8
ArcFace34	51.8	56.9	47.2	49.8	47.5	49.8
ArcFace50	55.8	59.8	47.5	50.7	62.5	68.2
ArcFace100	66.6	71.0	57.5	65.4	72.7	81.0
R34-Baseline	64.9	73.1	59.2	68.4	74.4	83.2
R34-Mask	93.8	92.9	90.7	92.2	94.3	98.6
R34-AMaskNet	93.5	93.0	92.0	92.7	94.1	98.5

✓ means using the proposed mask-aware similarity matching strategy (Wo/W+MS), while × means not applicable (Wo/W).

TABLE 2. Results on Public-IvS dataset with 1:1 verification protocol at $TAR@FAR=10^{-5}$.

Model	Public-IvS	
	×	✓
ArcFaceM	23.4	60.3
ArcFace34	23.4	60.3
ArcFace50	53.8	65.3
ArcFace100	67.5	75.7
R34-Baseline	59.7	70.8
R34-Mask	93.9	94.3
R34-AMaskNet	95.4	95.2

✓ means using the proposed mask-aware similarity matching strategy (Wo/W+MS), while × means not applicable (Wo/W).

TABLE 3. Results on RWMFD dataset.

Method	Accuracy
J. Luttrell et al. [27]	85.7
Hariri et al. [28]	84.6
Almabdy et al. [10]	87.0
Walid Hariri. [15]	91.3
ArcFaceM	34.6
ArcFace34	43.2
ArcFace50	49.3
ArcFace100	61.7
R34-Baseline	81.9
R34-Mask	92.5
R34-AMaskNet	94.3

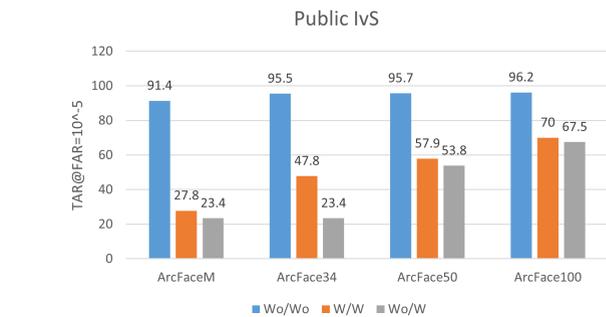


FIGURE 11. Results on Public-IvS dataset with 1:1 verification protocol at $TAR@FAR=10^{-5}$. Here, Wo means without wearing a mask, W means with wearing a mask.

spatial information in the inference stage.

D. COMPARISON OF STATE-OF-THE-ART METHODS ON RWMFD DATASET

To further verify the effectiveness of the proposed method on real mask data, we compare our method, public models and other literature models on RWMFD dataset, with the result shown in the Table 3, where the results of literature methods are taken from the corresponding papers. The R34-AMaskNet outperforms R34-Baseline with an improvement of up to 12.5 percentage points. Meanwhile, R34-AMaskNet outperforms the competitive methods by a significant margin, which indicates the proposed method is efficient when applied to a real masked face recognition scenario.

E. EFFECT OF MASK ON PERFORMANCE

The four publicly available models are evaluated under three test conditions, i.e., Wo/Wo, W/W, and Wo/W, to study the effect of wearing a mask on recognition accuracy.

Effect on COX. As shown in Fig.10, compared to the scene without a mask, the recognition accuracy is largely decreased when a mask is present. For example, on Cam1 of COX data, ArcFace50 decreases from 95.2% on Wo/Wo to 55.8% on Wo/W, resulting in a 39.4 percentage points loss of accuracy. This kind of accuracy decrease holds on all camera scenes of COX dataset. Moreover, the poorer the model performance, the greater the loss of accuracy, e.g. 52.2% and 31% losses in ArcFaceM and ArcFace100, respectively, from

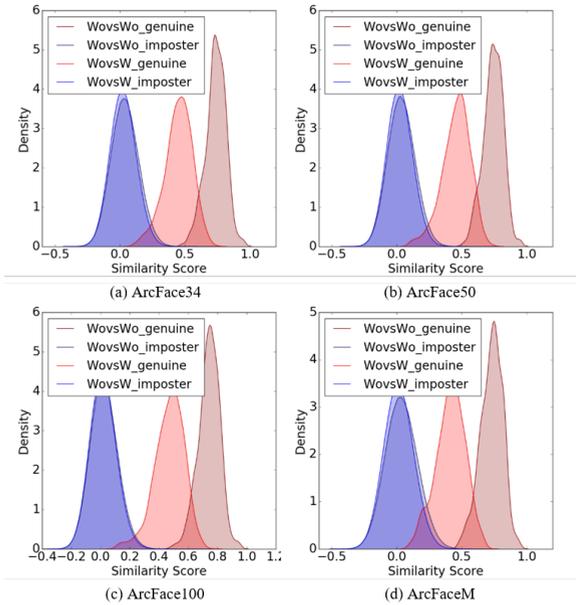


FIGURE 12. Distribution comparison of similarity scores on public model. Here, Wo means without wearing a mask, W means with wearing a mask.

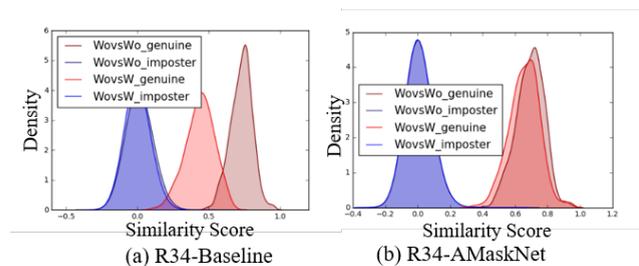


FIGURE 13. Distribution comparison of similarity scores. (a) In baseline model, the genuine scores significantly shift towards the imposter scores when the image is with a mask. (b) In R34-AMaskNet, the shift of similarity scores is largely mitigated. Here, Wo means without wearing a mask, W means with wearing a mask

Wo/Wo to Wo/W. It is worth mentioning that having one image with a mask while the other image remains without has a greater adverse effect on performance than both having masks, e.g. the loss of accuracy of ArcFaceM from Wo/Wo to Wo/W is 52.2 percentage points, but only 40 percentage points from Wo/Wo to W/W.

Effect on Public-IvS. The recognition results on Public-IvS are shown in Fig.11. Again, we get similar finding to CoX, that is, the recognition accuracy is largely decreased when a mask is present. For example, ArcFace50 decreases from 95.7% on Wo/Wo to 53.8% on Wo/W, resulting in a 41.9 percentage points loss of accuracy.

Effect of the similarity distributions. To understand the reason behind, we analyze the similarity score distributions of genuine and imposter pairs in Wo/Wo and Wo/W test conditions, with the result on Public-IvS dataset shown in Fig. 12. The choice of FAR determines the score threshold,

and then affects the results of TAR. In comparison with the Wo/Wo condition, the scores of genuine pairs strongly transfer towards the imposters when one image is with a mask (Wo/Wo vs. Wo/W). That means, the scores of genuine pairs become smaller and are nearer to imposter pairs because of the influence of masks, which will cause the TAR to become smaller at the same FAR, making them less recognizable and performance degradation.

F. QUALITATIVE ANALYSIS

The distributions of similarity score. The matching score distributions of genuine and imposter pairs in Wo/Wo and Wo/W test conditions are presented in Fig. 13. In comparison with the Wo/Wo condition, the scores of genuine pairs of the baseline model shift towards the imposter counterparts when one image is of a masked face (Wo/Wo vs. Wo/W), implying that the scores of genuine pairs decrease and are closer to those of imposter pairs due to the influence of masks, which makes them less recognizable. With the R34-AMaskNet system, however, the score distribution of the genuine pairs shifts only slightly toward imposter ones, which clearly indicates that a stronger recognition capability is obtained in R34-AMaskNet.

Contribution estimation. We randomly selected some samples from the testing dataset for visual analysis. Fig. 14 qualitatively shows the contribution estimation result using CAM [4] for the purpose of intuitive understanding. It can be found from this result that R34-AMaskNet can focus on the non-masked regions and exclude the background for the most samples, suggesting that discriminative regions for face recognition are obtained. Even in images without a mask, this attention scheme can also localize the facial area and eliminate background interference, that is why the model performs slightly better under Wo/Wo conditions.

V. CONCLUSION

An effective method is proposed with which to mitigate the effect of mask defects in face recognition. Firstly, a low-cost, accurate method of masked face synthesis is proposed for use in data augmentation and a mask-aware similarity matching strategy is developed, which is low-cost, rapid, and convenient for model development. Secondly, an attention-aware masked face recognition method (AMaskNet) is developed to improve the performance of masked face recognition, which includes two modules: a feature extractor and a contribution estimator, and the latter is used to learn the contribution matrix, thus outputting refined features by successive matrix multiplication. This method can learn an effective contribution matrix and automatically assign higher weights to the feature map activated by the non-masked facial parts and lower weights to those that are activated by masked facial parts. Finally, a mask-aware similarity method is proposed for use in the inference stage, thus mitigating the difference caused by the mask without loss of spatial information. Both qualitative and quantitative results show that the proposed

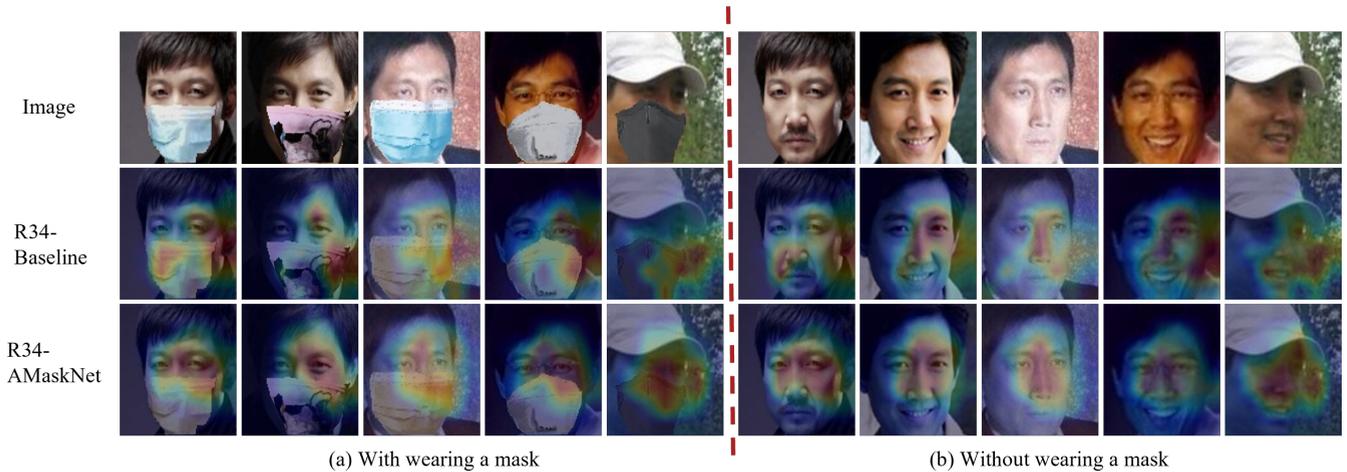


FIGURE 14. Visualization of attention result. (a) are the images with wearing a mask and its visualization results on R34-Baseline and R34-AMaskNet respectively. (b) are the images without wearing a mask and its visualization results on R34-Baseline and R34-AMaskNet respectively. From the first line to the third line are the original image, the attention results of R34-Baseline and R34-AMaskNet respectively. The maps highlight the discriminative image regions used for face recognition. The model with contribution module can be successfully able to localize the discriminative regions for face recognition.

model can mitigate the effects of mask defects in face recognition.

While the method is designed for masked face recognition, it can also be applied in other computer vision tasks, especially for other face-related applications such as facial attribute recognition.

REFERENCES

- [1] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, H. Chen, Y. Miao, Z. Huang, and J. Liang, "Masked face recognition dataset and application," *CoRR*, vol. abs/2003.09093, 2020. [Online]. Available: <https://arxiv.org/abs/2003.09093>
- [2] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, and X. Chen, "A benchmark and comparative study of video-based face recognition on COX face database," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5967–5981, 2015. [Online]. Available: <https://doi.org/10.1109/TIP.2015.2493448>
- [3] X. Zhu, H. Liu, Z. Lei, H. Shi, F. Yang, D. Yi, G. Qi, and S. Z. Li, "Large-scale bisample learning on ID versus spot face recognition," *Int. J. Comput. Vis.*, vol. 127, no. 6-7, pp. 684–700, 2019. [Online]. Available: <https://doi.org/10.1007/s11263-019-01162-8>
- [4] How to protect yourself & others from coronavirus disease. [Online]. Available: <https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html>
- [5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 4690–4699. [Online]. Available: http://openaccess.thecvf.com/content/CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html
- [6] Y. Guo, L. Z. and Yuxiao Hu and Xiao dong He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9907. Springer, 2016, pp. 87–102. [Online]. Available: https://doi.org/10.1007/978-3-319-46487-9_6
- [7] Y. Kim, W. Park, M. Roh, and J. Shin, "Groupface: Learning latent groups and constructing group-based representations for face recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 5620–5629. [Online]. Available: <https://doi.org/10.1109/CVPR42600.2020.00566>
- [8] A. Cabani, K. Hammoudi, H. Benhabiles, and M. Melkemi, "Maskedface-net - A dataset of correctly/incorrectly masked face images in the context of COVID-19," *CoRR*, vol. abs/2008.08016, 2020. [Online]. Available: <https://arxiv.org/abs/2008.08016>
- [9] learnopencv. [Online]. Available: <https://github.com/spmallick/learnopencv>
- [10] A. Anwar and A. Raychowdhury, "Masked face recognition for secure authentication," *CoRR*, vol. abs/2008.11104, 2020.
- [11] [Online]. Available: https://github.com/ash368/face_mask
- [12] [Online]. Available: <https://github.com/davisking/dlib>
- [13] Y. Shen, P. Luo, J. Yan, X. Wang, and X. Tang, "Faceid-gan: Learning a symmetry three-player GAN for identity-preserving face synthesis," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 821–830. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Shen_FaceID-GAN_Learning_a_CVPR_2018_paper.html
- [14] D. Saez-Trigueros, L. Meng, and M. Hartnett, "Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss," *Image Vis. Comput.*, vol. 79, pp. 99–108, 2018. [Online]. Available: <https://doi.org/10.1016/j.imavis.2018.09.011>
- [15] W. Hariri, "Efficient masked face recognition method during the COVID-19 pandemic," *CoRR*, vol. abs/2105.03026, 2021. [Online]. Available: <https://arxiv.org/abs/2105.03026>
- [16] Mundial, H. I. Q., M. S. U., Tiwana, M. I., Qureshi, W. S., and E. Alanazi, "Towards facial recognition problem in covid-19 pandemic," in *ELTICOM 2020 4rd International Conference on Electrical, Telecommunication and Computer Engineering*. IEEE, 2020, pp. 210–214.
- [17] M. Emambakhsh and A. N. Evans, "Nasal patches and curves for expression-robust 3d face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 995–1007, 2017. [Online]. Available: <https://doi.org/10.1109/TPAMI.2016.2565473>
- [18] Y. Rao, J. Lu, and J. Zhou, "Attention-aware deep reinforcement learning for video face recognition," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 3951–3960. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.424>
- [19] N. Sankaran, S. Tulyakov, S. Setlur, and V. Govindaraju, "Metadata-based feature aggregation network for face recognition," in *2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018*. IEEE, 2018, pp. 118–123. [Online]. Available: <https://doi.org/10.1109/ICB2018.2018.00028>
- [20] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV,*

USA, June 27-30, 2016. IEEE Computer Society, 2016, pp. 2921–2929. [Online]. Available: <https://doi.org/10.1109/CVPR.2016.319>

[21] Y. Sun, X. Wang, and X. Tang, “Deeply learned face representations are sparse, selective, and robust,” in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 2892–2900. [Online]. Available: <https://doi.org/10.1109/CVPR.2015.7298907>

[22] S. Woo, J. Park, J. Lee, and I. S. Kweon, “CBAM: convolutional block attention module,” in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11211. Springer, 2018, pp. 3–19. [Online]. Available: https://doi.org/10.1007/978-3-030-01234-2_1

[23] H. Ling, J. Wu, L. Wu, J. Huang, J. Chen, and P. Li, “Self residual attention network for deep face recognition,” *IEEE Access*, vol. 7, pp. 55 159–55 168, 2019. [Online]. Available: <https://doi.org/10.1109/ACCESS.2019.2913205>

[24] C. Rother, V. Kolmogorov, and A. Blake, ““grabcut”: interactive foreground extraction using iterated graph cuts,” *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004. [Online]. Available: <https://doi.org/10.1145/1015706.1015720>

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1106–1114. [Online]. Available: <https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html>

[26] [Online]. Available: <http://trillionpairs.deepglint.com/overview>

[27] Y. Z. C. Z. P. G. B. Y. J. Luttrell, Z. Zhou and R. Li, “A deep transfer learning approach to fine-tuning facial recognition models,” *IEEE Conference on Industrial Electronics and Applications (ICIEA)*, p. 2671–2676, 2018.

[28] W. Hariri, H. Tabia, N. Farah, A. Benouareth, and D. Declercq, “3d face recognition using covariance based descriptors,” *Pattern Recognit. Lett.*, vol. 78, pp. 1–7, 2016. [Online]. Available: <https://doi.org/10.1016/j.patrec.2016.03.028>



MENG ZHANG received his MSc degree in Beijing University Of Technology, China, in 2016. Since then, he worked as a researcher in Fujitsu Research and Development Center Co. Ltd., Beijing, China. He is currently pursuing the Ph.D. degree in intelligent systems at graduate school of informatics, Nagoya University, Japan. His research interests include image processing, pattern recognition, face recognition and deep learning.



RUJIE LIU received his B.s., M.s., and Ph.D. degrees in electronic engineering from Beijing Jiaotong University in 1995, 1998, and 2001, respectively. Since then, he worked as a researcher in Fujitsu Research and Development Center Co. Ltd., Beijing, China. His research interests are in the areas of AI, pattern recognition, and image processing. He has published more than 40 papers and tens of inventions.



DAISUKE DEGUCHI (Member, IEEE) received his B.Eng and M.Eng degrees in Engineering and Ph.D degree in Information Science from Nagoya University, Japan, in 2001, 2003, and 2006, respectively. He became a Post Doctoral Fellow at Nagoya University, Japan in 2006. From 2008 to 2012, he had been an Assistant Professor at the Graduate School of Information Science. From 2012 to 2019, he was an Associate Professor in Information Strategy Office, and Since 2019, he

has been an Associate Professor at the Graduate School of Informatics. He is working on the object detection, segmentation, recognition from videos, and their applications to ITS technologies, such as detection and recognition of traffic signs. He is a member of IEICE, IPS Japan and IEEE.



HIROSHI MURASE (Life Fellow, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees in electrical engineering from Nagoya University, Japan. In 1980, he joined Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993, he was a Visiting Research Scientist with Columbia University, New York. He has been a Professor with Nagoya University, since 2003. His research interests include computer vision, pattern recognition, and multimedia information processing. He is a fellow of the IPSJ and the IEICE. He was awarded the IEEE CVPR Best Paper Award, in 1994, the IEEE ICRA Best Video Award, in 1996, the IEICE Achievement Award, in 2002, the IEEE Multimedia Paper Award, in 2004, and the IEICE Distinguished Achievement and Contributions Award, in 2018. He received the Medal with Purple Ribbon from the Government of Japan, in 2012.

...