IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 4, 2022, accepted February 24, 2022, date of publication March 21, 2022, date of current version April 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3160828

Masked Face Recognition From Synthesis to Reality

GEE-SERN JISON HSU[1], (Senior Member, IEEE), HUNG-YI WU[1], CHUN-HUNG TSAI[1], SVETLANA YANUSHKEVICH[2], (Senior Member, IEEE), AND MARINA L. GAVRILOVA[3], (Senior Member, IEEE)

[1]Artificial Vision Laboratory, Department of Mechanical Engineering, National Taiwan University of Science and Technology, Taipei 10607, Taiwan
[2]Department of Electrical and Software Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada
[3]Department of Computer Science, University of Calgary, Calgary, AB T2N 1N4, Canada

Corresponding author: Gee-Sern Jison Hsu (jison@mail.ntust.edu.tw)

This work was supported in part by the Ministry of Science and Technology (MOST), Taiwan, under Grant 108-2923-E-011-003-MY3, Grant 109-2221-E-011-124-MY3, and Grant 110-2634-F-002-050; in part by the Center for Cyber-Physical System Innovation through the Higher Education Sprout Project, Ministry of Education (MOE), Taiwan; and in part by the collaboration with the University of Calgary, Canada.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT As we have been seriously hit by the COVID-19 pandemic, wearing a facial mask is a crucial action that we can take for our protection. This paper reports a comprehensive study on the recognition of masked faces. By using facial landmarks, we synthesize the facial mask for each face in several benchmark databases with different challenging factors. The IJB-B and IJB-C databases are selected for evaluating the performance against the variation across pose, illumination and expression (PIE). The FG-Net database is selected for evaluating the performance across age. The SCface is chosen for evaluating the performance on low-resolution images. The MS-1MV2 is exploited as the base training set. We use the ResNet-100 as the feature embedding network connected to state-of-the-art loss functions designed for tackling face recognition. The loss functions considered include the Center Loss, the Marginal Loss, the Angular Softmax Loss, the Large Margin Cosine Loss and the Additive Angular Margin Loss. Both verification and identification are conducted in our evaluation. The performances for recognizing faces with and without the synthetic masks are all evaluated for a complete comparison. The network with the best loss function for recognizing synthetic masked faces is then assessed on a real masked face database, the cleaned RMFRD (c-RMFRD) dataset. Compared with a human user test on the c-RMFRD, the network trained on the synthetic masked faces outperforms human vision for a large gap. Our contributions are fourfold. The first is a comprehensive study for tackling masked face recognition by using state-of-the-art loss functions against various compounding factors. For comparison purpose, the second is another comprehensive study on the recognition of faces without masks by using the same loss functions against the same challenging factors. The third is the verification of the network trained on synthetic masked faces for tackling the real masked face recognition with performance better than human inspectors. The fourth is the highlight on the challenges of masked face recognition and the directions for future research. Our code, trained models and dataset are available via the project GitHub site.

INDEX TERMS Face recognition, face database, facial mask, COVID-19.

I. INTRODUCTION

One of the core topics in the fields of computer vision is face recognition. Due to the success of deep learning approaches, face recognition has been substantially improved in recent years, and the improvements will continue as more challenging tasks and advanced approaches are emerging. Recognition of masked faces is a challenging task. The importance of this problem cannot be overemphasized, especially at the current time as the COVID-19 pandemic has changed our

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

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

daily lives. The global and local economies have been seriously hit by the COVID-19 pandemic. Many countries implement social distancing, travel restrictions and strict lockdown. Wearing a facial mask in the locations where social distancing is hard to keep has become a required or mandatory action in many countries. The Centers for Disease Control (CDC) issues Your Guide to Masks [1] to recommend wearing masks. The research reported in this paper transforms the latest face recognition solutions to the approaches that can handle masked face recognition. We extensively compare the state-of-the-art loss functions designed for common face recognition, i.e., without masks, and evaluate their performance for the recognition of synthetic and real masked faces.

A typical deep convolutional neural network (CNN) for face recognition is composed of feature embedding layers, fully-connected layers and an output layer with a loss function. The loss function is a core part that determines how well the target task is solved. The most common loss function is the softmax function that computes the cross-entropy loss for solving a classification problem. The improvements or modifications to the softmax function is an active research topic with the goal of finding more effective loss functions. As the loss function aims to decrease the intra-class variance and increase the extra-class variance, the effectiveness of the loss function is generally evaluated by comparing the similarity scores of the image pairs formed by the intra-class and extra-class sets.

In recent years, researchers have developed several loss functions which are mostly advanced from the conventional softmax loss function. Five state-of-the-art loss functions are selected in this study, including the Center Loss [2], the Marginal Loss [3], the Angular Softmax Loss [4], the Large Margin Cosine Loss [5] and the Additive Angular Margin Loss [6]. All these loss functions are designed with special aspects and features. For example, the Angular Softmax Loss [4] is defined in an angular feature space instead of the common Euclidean space so that the angular margin for measuring the inter-class variance can be computed, leading to an improvement to the recognition performance. The Large Margin Cosine Loss [5] considers a cosine margin penalty to the target logit, resulting in a better performance than the Angular Softmax Loss. The Additive Angular Margin Loss [6] further introduces the additive angular margin penalty between the normalized features and weights, achieving a better performance than the Large Margin Cosine Loss.

The performance of the selected loss functions has been reported on a few databases, including the Labeled Faces in the Wild (LFW) [7], the IARPA Janus Benchmark–A (IJB-A) [8], the YouTube Faces Database (YTF) [9], and the MegaFace [9]. Although these databases offer a range of variation in pose, illumination, expression (PIE) good for performance assessment, the following issues require our attention:

- The databases used in the previous evaluations are mostly with PIE variation and inappropriate for evaluating the performance against other factors, for example, age and resolution. The databases with these specific factors must be evaluated to better identify more challenging factors. Moreover, a few latest benchmarks, such as IJB-B and IJB-C, which are more challenging than previous databases must be tested to update the performance against the PIE variation.
- Recognition of masked faces is a required study as facial masks are accepted as a common safety protection means during and after the COVID-19 pandemic. To better evaluate the loss functions for recognizing masked faces, the aforementioned factors must be considered as well so that the performance on the compounding effects, for example, low-resolution masked faces, can be studied.
- The collection of a large number of masked faces can be difficult, not to mention the masked faces with the aforementioned compounding effects. It is necessary to develop an approach to transform the (no-mask) faces in the existing face databases to masked faces to facilitate the study on masked face recognition. The solution developed based on the transformed masked faces needs to be validated on real-life masked faces.

To address the above issues, we choose several benchmark databases with specific challenging factors, propose a landmark-based technique for making a synthetic facial mask to each face in the databases, and evaluate the performance of the state-of-the-art loss functions that were originally designed for generic face recognition, but are now trained and tested on the databases with synthetic masked faces added in. This study does not just focus on the masked face recognition, but also on the recognition against three compounding factors: age, resolution and PIE variation. For comparison purpose, we also conduct the same experiments on the original databases without the synthetic masked faces.

To study the effects made by the compounding factors, the IJB-B [10] and IJB-C [11] are selected for evaluating the performance against the generic PIE variation (pose, illumination and expression). The FG-Net Aging Database (FG-Net) [12] is selected for evaluating the cross-age recognition. The Surveillance Cameras Face Database (SCface) [13] is chosen for evaluating the performance for low resolution images. We select the MS-1MV2 dataset [6], which is a cleaned version of the MS-Celeb-1M [14] dataset, as the base training set used throughout this study.

We exploit the ResNet-100 [15] as the feature embedding network, replace the default softmax loss function by the five latest loss functions, and evaluate the performance on the aforementioned four databases with and without the synthetic masks added on. Figure 1 shows the system configuration. The best loss function determined from the study on the synthetic masked faces is then experimented on the Real-world Masked Face Recognition Dataset (RMFRD) [16]. As the RMFRD has many mislabeled data, we manually cleaned the dataset and redefine a cleaned RMFRD (c-RMFRD), on which we conduct our experiments.

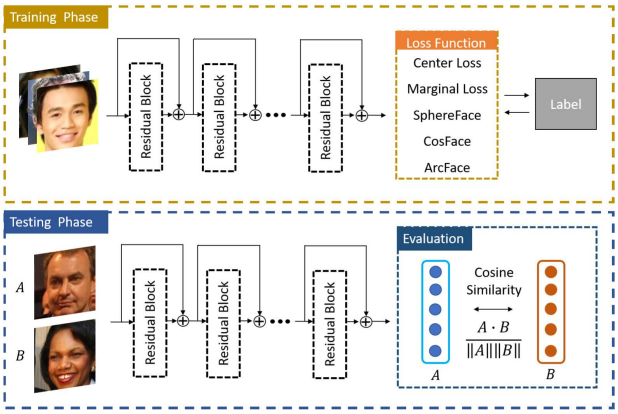


FIGURE 1. The network is composed of the ResNet-100 feature embedding layers and one of the five selected loss functions. For training, we use the MS-1MV2 dataset with and without synthetic masks added on. For testing, we use the SCFace for low-resolution images, the FG-Net for cross -age performance, the IJB-B and IJB-C for PIE variation and the RMFRD for the real masked faces. Both scenarios with and without masks are experimented.

The contributions of this study are summarized as follows:

- A comprehensive study on the masked face recognition by evaluating 5 state-of-the-art loss functions against 4 challenging factors is offered. This can be one of the most in-depth studies for masked face recognition.
- A comprehensive study on the recognition of faces without masks by evaluating the same 5 loss functions against the same 4 factors is also offered for comparison purpose. Differences from previous work are the databases and protocols that specify the performance against the 4 factors: 1) General PIE variation, 2) Facial age, 3) Partial occlusion and 4) Low resolution.
- It is verified that our model trained on synthetic masked faces offers an effective solution for the recognition of real masked faces.
- The experiments on the c-RMFRD dataset reveal the challenges when recognizing masked faces, highlighting the directions for future research.

This work is a substantial extension of our preliminary study reported in a recent CVPR workshop paper [17], where we only presented some part of the evaluation against the challenging factors on normal no-mask faces. The synthesis of facial masks, the recognition of masked faces and the performance against the amalgam of the facial masks and the three challenging factors are reported only in this paper, with extended experiments. Our code, trained models and dataset are available via https://github.com/AvLab-CV/Face_Mask_Generator. The rest of the paper is organized as follows: We first review recent works in Sec. II, followed by another review on the selected loss functions in Sec. III. The making of synthetic masks given is presented in Sec. IV. The experimental setup and results are given in Sec. V, with a conclusion to this study in Sec. VI.

II. RELATED WORK

Several approaches have been proposed recently. Zheng *et al.* propose a weakly supervised meta-learning approach to learn from the images collected from the web without manual annotation along with limited fully annotated datasets [18]. They capitalize on readily-available web images with noisy annotations to learn a robust representation for masked faces. Both the spatial and frequency domain features extracted from the unoccluded facial parts are considered. Li *et al.* propose a framework composed of a de-occlusion network and a distillation network [19]. The former uses a generative adversarial network to recover the facial region under the mask. The latter takes a pretrained face recognition model as a teacher to train the former as a student for improving the performance of de-occlusion. The knowledge to train the student is represented in structural relations and serves as a posterior regularization to enable the adaptation. The FocusFace [20] consists of two components, one for mask detection and the other for the contrastive learning of masked and unmasked faces. The MS-1MV2 with synthetic masks was used for training, and the performance was evaluated on the real masked faces in the IJCB-MFR-2021 competition [21]. However, the IJCB-MFR-2021 evaluation dataset is not released to the public. The approach proposed by Li *et al.* integrates a cropping-based approach with a convolutional block attention module that focuses on the region around the eyes [22]. Two training and testing scenarios are considered and mutually improve the performance of each other. Using facial landmarks, Anwar and Raychowdhury make an open-source tool, MaskTheFace, to synthesize masks for transforming general faces to masked faces [23]. The performance of training on synthesized masked faces is verified on a small-sized customized dataset, the MFR2. We also consider the MFR2 in our experiments. See Sec.V for details.

III. SELECTED LOSS FUNCTIONS

The loss functions selected for our study include the Center Loss [2], the Marginal Loss [3], the Angular Softmax Loss [4], the Large Margin Cosine Loss [5] and the Additive Angular Margin Loss [6]. As these loss functions consider the Softmax Loss as a core reference, we introduce the Softmax Loss first and the others follow.

The Softmax Loss function can be written as follows:

$$L_s = -\frac{1}{N_b}\sum_{i=1}^{N_b}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{1}$$

where $x_i \in \mathbb{R}^d$ denotes the $d-$dim deep feature of the i-th sample, belonging to the y_i-th class, $y_i \in [1, 2, \ldots, n]$. $W_j \in \mathbb{R}^d$ denotes the j-th column of the weight $W \in \mathbb{R}^{d\times n}$ and $b_j \in \mathbb{R}^n$ is the bias term. N_b and n are the batch size and the class number, respectively. The softmax loss is widely used in deep face recognition [24]. However, the softmax loss function does not optimize the feature embedding to enhance the similarity between intra-class samples and the diversity

between inter-class samples. This motivates the development of other loss functions.

A. CENTER LOSS

The Center Loss [2] was proposed to improve the softmax loss for face verification. It learns a center for the features of each class while trying to pull the deep features of the same class close to the corresponding center. Given the deep feature x_i in a batch, the center loss can be computed as:

$$L_{ce} = \frac{1}{2}\sum_{i=1}^{N_b} \left\| x_i - c_{y_i} \right\|_2^2 \tag{2}$$

where $c_{y_i} \in \mathbb{R}^d$ is the center of class y_i. During training, the center loss encourages the instances of the same classes to be closer to a learnable class center. However, since the class centers are updated at each iteration based on a mini-batch instead of the whole dataset, the learning process can be unstable. It has to be under the joint supervision of the softmax loss during training. Therefore, the following combined loss is considered when applying center loss:

$$\begin{aligned} L_c &= L_s + \lambda L_{ce} \\ &= -\sum_{i=1}^{N_b} \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} + \frac{\lambda}{2}\sum_{i=1}^{N_b} \left\| x_i - c_{y_i} \right\|_2^2 \end{aligned} \tag{3}$$

where L_s is the softmax loss (1) and λ is a hyper-parameter that balances the two losses.

B. MARGINAL LOSS

The Marginal Loss function [3] was proposed to simultaneously maximize the inter-class distances and minimize the intra-class variations. The Margin Loss function focuses on the marginal samples and is computed as follows:

$$L_{ma} = \frac{1}{N_b^2 - N_b} \sum_{i,j,i\neq j}^{N_b} \left(\xi - y_{ij} \left(\theta - \left\| \frac{x_i}{\|x_i\|} - \frac{x_j}{\|x_j\|} \right\|_2^2 \right) \right) \tag{4}$$

The term $y_{ij} \in \{\pm 1\}$ indicates whether the faces x_i and x_j are from the same class or not, θ is the distance threshold to distinguish whether the faces are from the same person/class, and ξ is the error margin besides the classification hyperplane [3]. Similar to the center loss prone to be unstable in training, the Marginal Loss will also be unstable at training because of the batch normalization. It is thus computed with the joint supervision with the Softmax loss L_s, as given below:

$$L_m = L_s + \lambda L_{ma} \tag{5}$$

The hyper-parameter λ balances the two losses. The coupling with the cross-entropy loss provides separable features and prevents the loss from degrading to zero [3].

C. ANGULAR SOFTMAX LOSS

The Angular Softmax Loss function [4] was proposed by Liu *et al.* to improve the issues with the bias $b_j = 0$ and the layer weight vector $\|W_j\| = 1$. The issue of the bias $b_j = 0$ is handled by transforming the logit as $W_j^T x_i = \|W_j\| \|x_i\| \cos\theta_j$, where θ_j is the angle between the layer weight W_j and the feature x_i [25]. The issue with the individual weight $\|W_j\| = 1$ is handled by taking the l_2 normalization to make the prediction only depend on the angle between the feature vector and the weight vector.

To make it discriminative, Liu *et al.* generalize it to the following Angular Softmax (called A-Softmax in short) Loss L_{AS}, and name their solution "SphereFace."

$$L_{as} = -\frac{1}{N_b}\sum_{i=1}^{N_b} \log \frac{e^{\|x_i\| \cos(m\theta_{y_i})}}{e^{\|x_i\| \cos(m\theta_{y_i})} + \sum_{j=1, j\neq y_i}^{n} e^{\|x_i\| \cos\theta_j}}. \tag{6}$$

where $\theta_{y_i} \in [0, \frac{\pi}{m}]$ and m is a hyperparameter. A-Softmax loss has stronger requirements for a correct classification when $m \geq 2$, which generates an angular classification margin between the learned features of different classes. A-Softmax loss imposes a discriminative power to the learned features via angular margin, equivalent to learning features that are discriminative on a hypersphere manifold, while Euclidean margin losses learn features in Euclidean space.

D. LARGE MARGIN COSINE LOSS

The Large Margin Cosine Loss function [5] was proposed by Wang *et al.* to solve the issues with the above A-Softmax loss. The decision boundary of the A-Softmax loss is defined over the angular space by $\cos(m\theta_1) = \cos(\theta_2)$, which can be difficult to optimize due to the non-monotonicity of the cosine function. To overcome this difficulty, the Large Margin Cosine Loss takes the normalized features as input to learn the highly discriminative features by maximizing the inter-class cosine margin. Wang *et al.* define a hyper-parameter m such that the decision boundary is given by $cos(\theta_1) - m = cos(\theta_2)$, where θ_i is the angle between the feature vector and weight vector of the class i. They reformulate the softmax loss as a cosine loss by applying the l_2 normalization on both the feature and weight vectors to remove radial variations, based on which a cosine margin term m is introduced to further maximize the decision margin in the angular space. Wang *et al.* call their solution "CosFace" [5]. The large margin Cosine loss L_{co} is computed as follows:

$$L_{co} = -\frac{1}{N_b}\sum_{i=1}^{N_b} \log \frac{e^{s(\cos\theta_{y_i} - m)}}{e^{s(\cos\theta_{y_i} - m)} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}} \tag{7}$$

where $s = \|x_i\|$.

E. ADDITIVE ANGULAR MARGIN LOSS

The Additive Angular Margin Loss [6] was proposed by Deng *et al.* to further improve the discriminative power of the loss function considered in a classification model and to

FIGURE 2. Mislabeled samples in RMFRD [16]. According to the official labels, each row denotes the same subject and the left one is the corresponding masked face. Many unmasked faces are distorted in the aspect ratio to some extent.

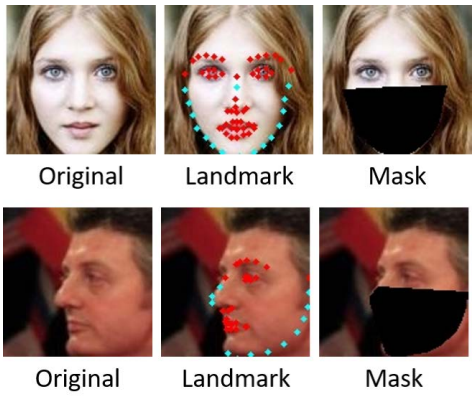


FIGURE 3. The mask for a frontal face is made by the region enclosed by the landmarks indexed the 2^{nd} to 16^{th} and 29^{th}, and that for a profile face is made by those indexed 8^{th} to 16^{th} and 29^{th} to 31^{th}, shown in blue. The red landmarks are not used for making the masks. The indices follow those defined by the FAN [26].

stabilize the training process. Following the work in [4], [5], the authors further normalize the feature and weight vectors, and coin their solution "ArcFace." The difference is that they add an additive angular margin penalty m between x_i and W_{y_i} to simultaneously enhance the intra-class compactness and inter-class discrepancy. The Additive Angular Margin Loss L_{aa} is computed as follows:

$$L_{aa} = -\frac{1}{N_b}\sum_{i=1}^{N_b}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}}. \tag{8}$$

Despite the formulation similarity between the ArcFace and previous works, the proposed additive angular margin has a better geometric attribute as the angular margin has the exact correspondence to the geodesic distance. It is shown in [6] that the ArcFace has a constant linear angular margin throughout the decision boundaries when handling binary classification; however, the SphereFace and CosFace yield nonlinear angular margins.

IV. REAL AND SYNTHETIC MASKED FACE DATASET

Facial mask is a special form of occlusion. We need a database made of masked faces for our study. The Real-world Masked Face Recognition Dataset (RMFRD) [16] is one of the very few databases available to date. However, the RMFRD suffers from the following issues:

- Mislabeling and distortion: A significant portion of the data is mislabeled, and many images are distorted in the aspect ratio to some extent. A few cases are shown in Figure 2.
- Difficult for verification: Some data can hardly be verified by human inspectors, and the causes include extreme pose, insufficient observable facial region, poor image quality and the above combined.
- Mixed factors: The data vary in pose, illumination, resolution, image quality and level of distortion. It can be difficult to know which factors affect more on the performance than others from the experimental results.

To settle the first two issues, we manually clean the dataset by removing the images which are difficult to verify, and wind up with a cleaned version of the RMFRD, or the c-RMFRD in short. The c-RMFRD has 831 masked faces of 248 subjects and 3,627 images of the same 248 subjects without masks. The c-RMFRD suffers from data insufficiency and mixed factors. In order to have a better understanding toward the masked face recognition with different compounding factors, we conduct an extensive study by using the synthetic masked faces. We find the best solution for the recognition of the synthetic masked faces, and then verify the performance on the c-RMFRD. In summary, we only use the c-RMFRD for performance validation and not for training because of the aforementioned data quantity and quality issues.

The synthetic masked faces are made by exploiting the facial landmarks. We exploit the Face Alignment Network (FAN) proposed by Bulat and Tzimiropoulos [26] for locating the facial landmarks. The FAN is constructed on a stack of four HourGlass (HG) networks [27] for successive pooling and upsampling to improve the landmark localization. Instead of using the bottleneck block as the building block for the HG (as in [27]), the FAN employs the residual block [15] for better accuracy. The FAN is trained on the 300W-LP-2D and 300W-LP-3D [28] for localizing the 2D and 3D facial landmarks, respectively. We only exploit the 2D facial landmarks for making the synthetic facial masks.

Given a face, the FAN detects 68 facial landmarks for the pose within $[-45^\circ \sim 45^\circ]$ in yaw (i.e., the so-called frontal pose range) and 39 for the pose beyond this range (the profile pose range). For each frontal face, we select the landmarks along the outside of the cheek to the chin area, and one on the nose, and make a patch enclosed by the selected

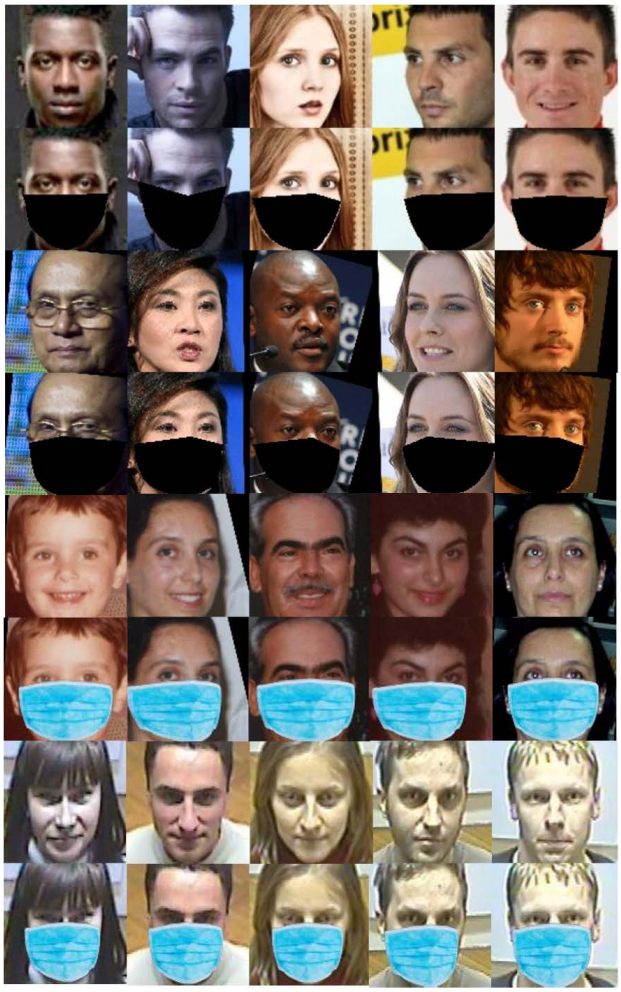


FIGURE 4. Samples with synthetic masks added on. Rows 1 and 2 are from the MS-1MV2 [6], Rows 3 and 4 are from the IJB-C [11], Rows 5 and 6 are from the FG-Net [12], Rows 7 and 8 are from the SCFace [13].

landmarks to form the synthetic mask. In terms of the FAN landmark indices, those selected for a frontal face are from the 2^{nd} to 16^{th} and the 29^{th}, as shown in Figure 3. For each profile face, we select the landmarks along the outside of the cheek to the chin area (indexed 8^{th} to 16^{th}), and three on the nose (indexed 29^{th} to 31^{st}) to form the synthetic mask. Figure 3 shows the landmarks used for making the facial marks in blue, and the remaining landmarks in red.

Figure 4 shows some sample faces from the databases used in our experiments, with the synthetic masks on. Rows 1 and 2 are taken from the masked MS-1MV2 [6], Rows 3 and 4 are taken from the masked IJB-C [11], Rows 5 and 6 are taken from the masked FG-Net [12], Rows 7 and 8 are taken from the masked SCFace [13]. It can be seen that the nose and mouth regions of each face are well covered by the synthetic mask regardless of the pose, illumination and expression. For low resolution images in the SCFace, we applied a blurring filter on the synthetic mask so that the entire image can reveal low-resolution quality.

In additions, We refer to [16] to obtain the width and height of the mask, and then use the dlib [29] and affine transform method to synthesize real facial mask. Figure 4 Rows 5, 6, 7 and 8 show some sample from the FG-Net [12] and SCFace [13] with and without real mask.

V. EXPERIMENTS

The ResNet-100 [15] is exploited as the feature embedding network. The ResNet, which is the winning architecture in the ILSVRC 2015 classification competition [15], introduces the identity shortcut connection for tackling the vanishing gradient issue when increasing the network depth. The default loss function in the ResNet-100 is a softmax function. In our experiments, the softmax function is replaced by the selected loss functions, then the network is re-trained, and then the facial feature vector is extracted. When matching two faces, the similarity score is computed by the cosine distance between the pair of the associated facial features. The system configuration is shown in Figure 1, and the experiments are designed for the following inspections:

1) The performance of the state-of-the-art loss functions on the recognition of synthetic masked faces under three challenging factors.
2) Same settings as in 1), but on the faces without masks, for comparison purpose.
3) The performance of the best loss function determined from 1) for recognizing the real masked faces in the c-RMFRD dataset.

The MS-1MV2 database [6] is selected for training. We also explore an augmented MS-1MV2 with all faces in the database are duplicated with the masks added on, i.e., the training set is double the size of the original MS-1MV2 and each face has a mask-on counterpart. A few samples are illustrated in Figure 4. We compare the performance of using only the original MS-1MV2 without the synthetic masks for training and the double-sized version with the synthetic masks added on. Additionally, we also consider the training set further augmented with low-resolution copies of each face, with and without the synthetic mask. The details of the training and testing databases are given in Sec. V-A. Due to the different characteristics of each testing database, our experiments are carried out as follows:

- The IJB-B and IJB-C are tested with and without the synthetic masks added on.
- The FG-Net is tested with and without the synthetic masks added on.
- The SCFace is also tested with and without the synthetic masks added on. However, for comparison purpose, we also augment the training set with low-resolution images. Each 112×112 face in the MS-1MV2 is first processed by a Gaussian filter and downsized to 56×56 and 28×28 pixels to handle the 70×70 and 40×40 faces in the SCFace.
- The best loss function determined from the above studies is validated on the c-RMFRD dataset.

In the following, we first give the details of the aforementioned databases and the system settings in Sec.V-A, and the experimental results in Sec.V-B with a discussion.

A. DATABASES AND HW/SW SETTINGS

1) DATASET FOR TRAINING

MS-1MV2 The MS-1MV2 dataset, which is a cleaned version of the MS-Celeb-1M database, offers 5.8M images of 85K celebrities [6]. The MS-Celeb-1M is one of the largest face datasets to date, and contains about 10M images for 100K celebrities [14]. The data in the MS-Celeb-1M cover a very broad scope of factors/variables, including pose, illumination, expression (PIE), occlusion, image resolution and others. However, it suffers from mislabeling noises. Deng *et al.* [6] hired ethnicity-specific annotators for cleaning the MS-Celeb-1M as the celebrities in the database are with multi-ethnic backgrounds. The ethnicity-specific annotators can better verify the faces of the same ethnic groups. The cleaned version is coined the MS-1MV2 dataset.

Depending on whether the synthetic masks are added on and whether the low-resolution copies are included, we design four different training sets:

1) The original MS-1MV2 without the synthetic masks (5.8M images);
2) The mask-augmented MS-1MV2 (11.6M images);
3) The low-resolution-augmented MS-1MV2 (11.6M images);
4) The mask- and low-resolution-augmented MS-1MV2 (23.2M images).

The experimental results are reported with different training sets. The above training sets 1) and 2) are used for all testing databases, and 3) and 4) only used for testing on the SCface for the investigation on the low-resolution images with or without masks.

2) DATASETS FOR PERFORMANCE EVALUATION

IJB-B and IJB-C: The IARPA Janus Benchmark–B (IJB-B) [10] contains 76.8K face images of 1,845 individuals, offering 12,115 templates with 10,270 genuine matches (intra pairs) and 8M impostor matches (extra pairs). The IARPA Janus Benchmark–C (IJB-C) [11] has 148.8K face images of 3,531 individuals, offering 23,124 templates with 19,557 genuine matches and 15,639K impostor matches. Both datasets contain still images and image frames taken off videos, with different image conditions regardless of subject conditions (pose, expression, occlusion) or acquisition conditions (illumination, standoff, etc.). We evaluated the performance on the mixed-media (frames and stills) 1:1 verification protocol and open-set 1:N identification protocol using the mixed media (frames, stills) as the probe set. When experimenting on masks, one face in each verification pair must wear a mask; and all faces in the probe set were with masks on and those in the gallery without masks.

FG-Net: The FG-Net Aging Database contains 1002 images of 82 subjects with ages ranging from newborns to 69 years [12]. Each subject has 6-18 face images at different ages. To conduct the cross-age 1:1 verification, we randomly formed 490,545 pairs, including 5,693 genuine pairs and 484,852 impostor pairs. The genuine pairs were formed by faces of the same subjects with different ages. The impostor pairs were formed by faces of different subjects with the same or different ages. To conduct the cross-age 1:N identification, we selected a face image that was closest to the age of 20 from each subject to form the gallery set. The rest of images were all used as the probe set. For the experiments with facial masks, the similar settings as implemented in the tests on the IJB-B and IJB-C were undertaken. For verification, one face in each test pair must wear a mask. For identification, the faces in the probe set wore masks, and those in the gallery set without masks.

SCface: The Surveillance Cameras Face Database offers 4,160 face images of 130 subjects captured in an uncontrolled indoor environment by using five video surveillance cameras of various qualities [13]. We follow the protocols defined by the authors with the testing dataset formed by 688 images of 43 subjects taken at three distances, 4.20 m (d1), 2.60 m (d2), and 1.00 m (d3), with average face size 40 × 40, 70 × 70 and 110 × 110, respectively. There are 688 genuine pairs and 27,090 impostor pairs for conducting 1:1 verification. For conducting face identification, we run a 5-fold cross validation with 26 subjects enrolled to the gallery set by their mugshot images at each fold, and all images taken by the surveillance cameras form the probe set. For the cross-resolution face identification, we follow the testing protocol conducted in the previous works [30]–[32]. We randomly select 80 subjects to test, frontal mugshot images are employed as gallery images and images taken by surveillance

TABLE 1. Verification rates (in TAR%, AUC%) for the loss functions tested on the IJB-B original (top 5 rows) and the masked IJB-B (bottom 10 rows). Top 10 rows are trained on 1) the original MS-1MV2. The bottom 5 rows with subscript $_{mask}$ are trained on 2) the mask-augmented MS-1MV2.

Model	TAR(%)@FAR			AUC (%)
	0.01	0.001	0.0001	
Center Loss [2]	88.9	80.2	68.3	98.7
Marginal Loss [3]	90.1	82.5	72.6	98.9
SphereFace [4]	94.3	91.4	81.3	99.6
CosFace [5]	95.9	92.6	89.1	99.4
ArcFace [6]	**97.4**	**94.9**	**92.6**	**99.5**
Center Loss [2]	69.5	4.8	0	97.3
Marginal Loss [3]	73.5	31.4	0	97.6
SphereFace [4]	82.7	45.9	7.3	98.7
CosFace [5]	85.4	66.5	40.4	98.5
ArcFace [6]	92.5	82.1	63.9	99.1
Center Loss$_{mask}$	90.2	77.6	52.4	99.2
Marginal Loss$_{mask}$	94.7	84.6	66.8	98.9
SphereFace$_{mask}$	96.3	90.6	78.1	99.4
CosFace$_{mask}$	94.7	88.5	79.9	99.3
ArcFace$_{mask}$	**96.3**	**92.9**	**88.5**	**99.4**

cameras at three distances are used as the probe images. For the experiments with facial masks, the similar settings as conducted in the experiments on the IJB-B, IJB-C and FG-Net were implemented.

c-RMFRD: In [16], the authors claim that the original RMFRD (Real-world Masked Face Recognition Dataset) offers 5,000 images of 525 subjects wearing masks, and 90,000 images of the same 525 subjects without masks. However, the downloaded version only contains 2,203 images with masks and a large number of mislabeled images and subjects, as several samples shown in Figure 2. We manually cleaned the dataset by removing the mislabeled images, and ended up with a cleaned version, the c-RMFRD. The c-RMFRD offers 831 masked face images of 248 subjects, and 3627 images of the same 248 subjects without masks. We form 9263 genuine pairs and 158508 impostor pairs for conducting the 1:1 verification. For conducting the face identification, the images without masks form the gallery set, and the images with masks form the probe set.

3) DATA PROCESSING AND EXECUTION

For processing the face images, we followed the procedure of making the MS-1MV2 dataset reported in [6]. We used the MTCNN [33] to detect the facial regions and the associated five landmarks. Given the landmarks, each face in all datasets was cropped and normalized to 112×112 pixels. In the testing phase, we followed the best settings for the loss functions reported in [2]–[5], and computed the cosine distance of the two feature vectors extracted from the last fully-connected layer to obtain the similarity score. Our programs were written in Python with the MXNet deep learning framework [34]. We used the batch size 64 and trained the networks on a Ubuntu 18.04 with Titan X GPU, and CUDA 9.0 with cuDNN 7.6. The learning rate started from 0.1 and was divided by 10 at the 8^{th}, 12^{th} and 16^{th} epochs.

B. RESULTS AND DISCUSSION

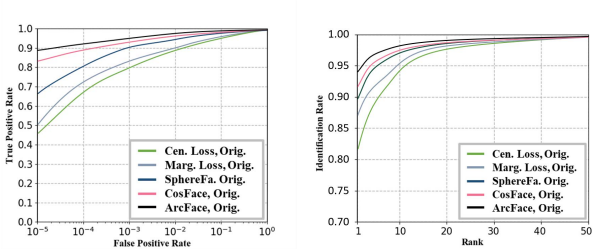
1) PERFORMANCE AGAINST PIE

Table 1 shows the verification rates of using the loss functions for handling the original IJB-B (the top 5 rows, without masks) and the masked IJB-B (the bottom 10 rows, with masks on). The bottom 10 rows are further divided into the top five trained on 1) the original MS-1MV2, and the bottom five trained on 2) the mask-augmented MS-1MV2. Table 2 shows the identification rates at FPIR = 0.01, 0.1 and Rank-1, 5, and the rows are arranged in the same way as those in Table 1.

With the model trained on 1) the original MS-1MV2, the ROC and CMC curves for testing on the original IJB-B are shown in Figures 5(a) and 5(b), respectively. Using the model trained on 2) the mask-augmented MS-1MV2 compared with the same model but trained on 1) the original MS-1MV2, the ROC and CMC curves for testing on the masked IJB-B are shown in Figures 6(a) and 6(b). The same training settings but tested on the original IJB-C and the masked IJB-C are shown in Tables 3 and 4 and Figures 7(a), 7(b), 8(a) and 8(b).

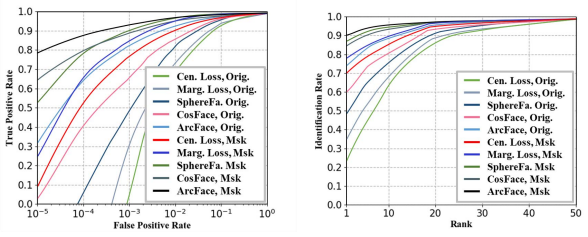
TABLE 2. Identification rates at FPIR = 0.01, 0.1 and Rank-1, 5 for the loss functions tested on IJB-B original (top 5 rows) and the masked IJB-B (bottom 10 rows). Top 10 rows trained on 1) the original MS-1MV2. The bottom 5 rows with $_{mask}$ are trained on 2) the mask-augmented MS-1MV2.

Model	FPIR=0.01	FPIR=0.1	Rank-1	Rank-5
Center Loss [2]	45.3	77.8	82.4	89.5
Marginal Loss [3]	48.5	79.4	87.6	92.9
SphereFace [4]	49.7	81.8	89.9	94.0
CosFace [5]	66.4	86.1	90.9	94.5
ArcFace [6]	**76.9**	**91.7**	**93.8**	**96.4**
Center Loss [2]	0	0	24.5	46.5
Marginal Loss [3]	0	0	36.2	52.6
SphereFace [4]	0.2	2.0	49.1	66.8
CosFace [5]	12.8	33.4	60.4	74.3
ArcFace [6]	16.7	55.4	74.4	84.8
Center Loss$_{mask}$	37.8	66.4	70.2	78.8
Marginal Loss$_{mask}$	40.3	67.3	77.8	85.2
SphereFace$_{mask}$	47.4	72.9	87.8	93.4
CosFace$_{mask}$	62.3	77.2	85.6	90.9
ArcFace$_{mask}$	**72.6**	**86.2**	**90.4**	**94.2**



(a) ROC on orig. IJB-B. (b) CMC on orig. IJB-B.

FIGURE 5. The ROC and CMC of using 1) the original MS-1MV2 for training and testing on original (no-mask) IJB-B.



(a) ROC on IJB-B$_{mask}$. (b) CMC on IJB-B$_{mask}$.

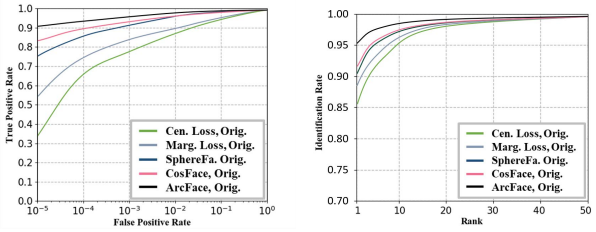
FIGURE 6. The ROC and CMC of using 1) the original MS-1MV2 and 2) the mask-augmented MS-1MV2 for training and testing on the synthetic masked IJB-B$_{mask}$.

The performance shown in the above tables and figures can be summarized as follows:

- Masks substantially degrade the performance of the network trained on the mask-free faces, i.e., the original MS-1MV2. The best performer ArcFace shows verification rates 92.6% on IJB-B and 93.6% on IJB-C @FAR

TABLE 3. Verification rates (in TAR%, AUC%) for the loss functions tested on the IJB-C original (top 5 rows) and the masked IJB-C (bottom 10 rows). Top 10 rows are trained on 1) the original MS-1MV2. The bottom 5 rows with subscript $_{mask}$ are trained on 2) the mask-augmented MS-1MV2.

Model	TAR(%)@FAR			AUC (%)
	0.01	0.001	0.0001	
Center Loss [2]	87.5	78.5	65.1	98.9
Marginal Loss [3]	90.4	83.6	74.9	99.1
SphereFace [4]	96.8	91.7	86.1	99.5
CosFace [5]	96.8	93.3	90.2	99.5
ArcFace [6]	**97.9**	**96.1**	**93.6**	**99.6**
Center Loss [2]	74.3	1.3	0	97.6
Marginal Loss [3]	78.4	23.7	0	98.1
SphereFace [4]	83.9	41.2	4.9	98.9
CosFace [5]	86.6	67.1	39.5	98.7
ArcFace [6]	93.4	84.0	63.7	99.2
Center Loss$_{mask}$	92.8	74.5	51.2	98.9
Marginal Loss$_{mask}$	92.5	81.9	63.7	99.0
SphereFace$_{mask}$	97.1	92.6	82.1	99.5
CosFace$_{mask}$	95.7	90.7	82.9	99.4
ArcFace$_{mask}$	**97.1**	**94.4**	**90.7**	**99.6**



(a) ROC on orig.IJB-C. (b) CMC on orig. IJB-C.

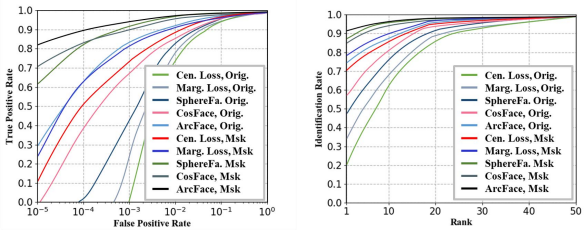
FIGURE 7. The ROC and CMC of using 1) the original MS-1MV2 for training and testing on original (no-mask) IJB-C.

10^{-4}, but drops to 63.9% and 63.7% on the masked IJB-B and IJB-C, respectively.

- The degraded performance is worsened for identification. At FPIR = 0.01, the ArcFace, trained on the original MS-1MV2, shows identification rates 76.9% on IJB-B and 88.9% on IJB-C, but drops to 16.7% and 5.4% on the masked IJB-B and IJB-C, respectively.
- The training on the mask-augmented MS-1MV2 can effectively improve the above issues. The ArcFace$_{mask}$ shows verification rates 88.5% and 90.7% on the masked IJB-B and IJB-C, respectively; and identification rates 72.6% and 79.1%, respectively.
- In most cases, the ArcFace outperforms all, followed by the CosFace, then the SphereFace, then the Marginal loss and then the Center loss. However, when recognizing the masked faces using the mask-augmented training, the SphereFace may outperform the CosFace, as shown by several cases in Tables 1, 2, 3 and 4.

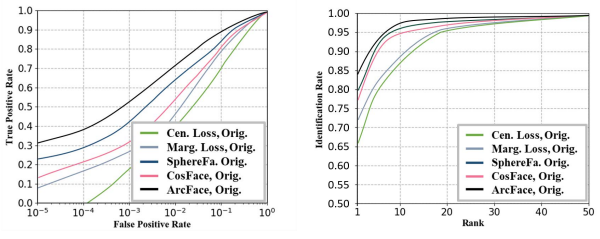
TABLE 4. Identification rates at FPIR = 0.01, 0.1 and Rank-1, 5 for the loss functions tested on IJB-C original (top 5 rows) and the masked IJB-C (bottom 10 rows). Top 10 rows trained on 1) the original MS-1MV2. The bottom 5 rows with $_{mask}$ are trained on 2) the mask-augmented MS-1MV2.

Model	FPIR=0.01	FPIR=0.1	Rank-1	Rank-5
Center Loss [2]	40.7	72.6	85.4	92.4
Marginal Loss [3]	44.8	76.3	88.6	94.1
SphereFace [4]	49.7	79.5	91.2	94.2
CosFace [5]	79.0	88.0	92.2	94.8
ArcFace [6]	**88.9**	**93.2**	**95.2**	**96.8**
Center Loss [2]	0	0	20.0	42.7
Marginal Loss [3]	0	0	35.4	53.4
SphereFace [4]	0	0.3	48.1	63.4
CosFace [5]	6.0	23.9	58.6	71.1
ArcFace [6]	5.4	42.5	73.8	83.5
Center Loss$_{mask}$	3.9	38.6	70.4	89.9
Marginal Loss$_{mask}$	24.7	56.5	79.7	85.1
SphereFace$_{mask}$	49.0	71.8	88.8	93.2
CosFace$_{mask}$	64.5	76.6	86.4	91.2
ArcFace$_{mask}$	**79.1**	**87.6**	**91.7**	**94.6**



(a) ROC on IJB-C$_{mask}$. (b) CMC on IJB-C$_{mask}$.

FIGURE 8. The ROC and CMC of using 1) the original MS-1MV2 and 2) the mask-augmented MS-1MV2 for training and testing on the synthetic masked IJB-C$_{mask}$.



(a) ROC on orig. FG-Net. (b) CMC on orig. FG-Net.

FIGURE 9. The ROC and CMC of using 1) the original MS-1MV2 for training and testing on original (no-mask) FG-Net.

2) PERFORMANCE AGAINST AGE

The tables and figures for the evaluations on the FG-Net are organized in the same way as those for the above IJB-B and IJB-C. Table 5 shows the verification rates on the original FG-Net (the top 5 rows, without masks) and the masked FG-Net (the bottom 10 rows). The bottom 10 rows

TABLE 5. Verification rates (in TAR%, AUC%) for the loss functions tested on the FG-Net original (top 5 rows) and the masked FG-Net (bottom 10 rows). Top 10 rows are trained on 1) the original MS-1MV2. The bottom 5 rows with subscript $_{mask}$ are trained on 2) the mask-augmented MS-1MV2.

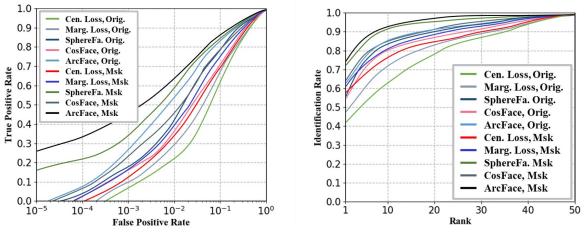
Model	TAR(%)@FAR			AUC (%)
	0.1	0.01	0.001	
Center Loss [2]	71.4	40.2	18.8	93.7
Marginal Loss [3]	89.2	46.6	27.5	94.8
SphereFace [4]	86.1	65.6	43.6	95.1
CosFace [5]	84.1	56.6	33.7	94.2
ArcFace [6]	**89.7**	**71.3**	**52.3**	**96.3**
Center Loss [2]	62.3	23.5	7.5	88.6
Marginal Loss [3]	67.9	30.1	10.2	89.7
SphereFace [4]	79.1	43.7	18.4	92.9
CosFace [5]	71.9	36.9	17.0	89.8
ArcFace [6]	83.8	54.3	28.9	94.4
Center Loss$_{mask}$	69.7	34.6	13.4	89.9
Marginal Loss$_{mask}$	75.7	40.5	17.2	90.2
SphereFace$_{mask}$	84.0	59.0	35.1	94.3
CosFace$_{mask}$	79.3	47.2	25.3	92.5
ArcFace$_{mask}$	**86.4**	**65.6**	**45.7**	**95.2**

TABLE 6. Identification rates at FPIR = 0.01, 0.1 and Rank-1, 5 for the loss functions tested on FG-Net original (top 5 rows) and the masked FG-Net (bottom 10 rows). Top 10 rows trained on 1) the original MS-1MV2. The bottom 5 rows with $_{mask}$ are trained on 2) the mask-augmented MS-1MV2.

Model	FPIR=0.01	FPIR=0.1	Rank-1	Rank-5
Center Loss [2]	46.4	50.3	65.8	80.3
Marginal Loss [3]	48.3	55.3	72.3	81.9
SphereFace [4]	57.9	68.1	79.9	90.3
CosFace [5]	46.4	61.7	77.11	89.2
ArcFace [6]	**60.2**	**72.6**	**84.5**	**92.4**
Center Loss [2]	7.8	25.9	43.1	54.1
Marginal Loss [3]	10.7	31.5	47.2	62.3
SphereFace [4]	13.8	35.0	55.9	76.3
CosFace [5]	21.1	35.2	54.9	71.8
ArcFace [6]	29.8	46.1	63.3	78.4
Center Loss$_{mask}$	21.1	39.2	58.5	67.4
Marginal Loss$_{mask}$	28.7	45.6	62.4	73.3
SphereFace$_{mask}$	41.8	58.5	72.1	85.3
CosFace$_{mask}$	33.8	51.6	65.5	81.1
ArcFace$_{mask}$	**48.1**	**61.8**	**74.5**	**88.1**

are further divided into the top five trained on the original MS-1MV2, and the bottom five trained on the mask-augmented MS-1MV2. Table 6 shows the identification rates at FPIR = 0.01, 0.1 and Rank-1, 5, and the rows are arranged in the same way as those in Table 5.

Considering the model trained on the original MS-1MV2, the ROC and CMC curves for testing on the original FG-Net are shown in Figures 9(a) and 9(b), respectively. Considering



(a) ROC on FG-Net$_{mask}$. (b) CMC on FG-Net$_{mask}$.

FIGURE 10. The ROC and CMC of using 1) the original MS-1MV2 and 2) the mask-augmented MS-1MV2 for training and testing on the synthetic masked FG-Net$_{mask}$.

the model trained on the mask-augmented MS-1MV2 compared with the same model but trained on the original Ms-1MV2, the ROC and CMC curves for testing on the masked FG-Net are shown in Figures 10(a) and 10(b), respectively. The performance shown in these tables and figures can be summarized as follows:

- The performance degradation of the networks trained on the original MS-1MV2 (without masks) and tested on the masked FG-Net is worse than that observed on the masked IJB-B/IJB-C. The best performer ArcFace, trained on the original MS-1MV2, shows verification rate 52.3% on the original FG-Net @FAR 10^{-3}, drops to 28.9% on the masked FG-Net.
- The degraded performance is worsened for identification. At FPIR = 0.01, the ArcFace trained on the original MS-1MV2 shows the identification rate 60.2% on the FG-Net, drops to 29.8% on the masked FG-Net.
- Similar to the above experiments on the IJB-B/IJB-C, the training on the mask-augmented MS-1MV2 can substantially improve the degraded performance. The ArcFace$_{mask}$ shows the verification rate 45.7% (versus 28.9% without augmented training) and identification rates 48.1% (versus 29.8% without augmented training) on the masked FG-Net.
- As shown in Tables 5 and 6, the ArcFace performs the best, followed by the SphereFace, then the CosFace, then the Marginal loss and then the Center loss. Again, this observation that the SphereFace outperforms the CosFace contradicts the experimental validation reported in [5] where different face benchmarks were considered. Together with our results obtained on the IJB-B/IJB-C, the CosFace can be better than the SphereFace when dealing with PIE variation. However, when handling other factors, for example, occlusion and age, the SphereFace appears a better option.
- A special observation of this study is that the cross-age performance is the lowest among the three factors with a clear margin. This highlights a potential direction for the future research in this field.

TABLE 7. Verification rates (in TAR%, AUC%) for the loss functions tested on the original SCface (no mask). Top 5 rows are trained on 1) the original MS-1MV2, the bottom 5 rows with $_{low}$ are trained on 3) the low-resolution -augmented MS-1MV2.

Model	TAR(%)@FAR			AUC (%)
	0.1	0.01	0.001	
Center Loss [2]	90.8	71.9	49.2	96.7
Marginal Loss [3]	90.6	74.0	58.4	96.8
SphereFace [4]	91.2	75.8	65.4	97.2
CosFace [5]	93.2	79.2	69.2	97.6
ArcFace [6]	95.2	84.3	74.3	98.2
Center Loss$_{low}$	95.4	84.4	69.6	98.4
Marginal Loss$_{low}$	95.5	85.2	72.1	98.2
SphereFace$_{low}$	**97.9**	**89.2**	74.7	**99.3**
CosFace$_{low}$	97.7	88.4	78.5	99.1
ArcFace$_{low}$	97.5	88.2	**79.1**	99.1

TABLE 8. Verification rates (in TAR%, AUC%) for the loss functions tested on the masked SCface. Top 5 rows are trained on 1) the original MS-1MV2, middle 5 rows with $_{low}$ are trained on 2) the mask-augmented MS-1MV2, and the bottom 5 rows with $_{low+mask}$ are trained on 4) the mask-low-resolution augmented MS-1MV2.

Model	TAR(%)@FAR			AUC (%)
	0.1	0.01	0.001	
Center Loss [2]	68.6	43.4	24.1	90.5
Marginal Loss [3]	74.1	45.1	29.3	91.0
SphereFace [4]	74.4	50.5	33.4	91.1
CosFace [5]	80.3	62.9	46.6	93.4
ArcFace [6]	80.4	62.0	48.4	92.2
Center Loss$_{mask}$	85.4	66.4	46.9	95.8
Marginal Loss$_{mask}$	87.7	68.0	52.2	96.0
SphereFace$_{mask}$	89.9	69.7	56.1	96.3
CosFace$_{mask}$	88.1	73.0	60.7	96.7
ArcFace$_{mask}$	88.8	73.8	59.7	96.4
Center Loss$_{low+mask}$	87.6	75.5	58.4	96.5
Marginal Loss$_{low+mask}$	88.5	78.3	61.2	96.8
SphereFace$_{low+mask}$	**94.1**	81.2	63.1	97.9
CosFace$_{low+mask}$	93.5	82.8	67.3	98.3
ArcFace$_{low+mask}$	93.8	**83.5**	**67.7**	**98.5**

3) PERFORMANCE AGAINST LOW RESOLUTION

Tables 7 and 8 show the verification rates of using the loss functions on the original SCface and the masked SCface, respectively. The top 5 rows in Table 7 are trained on the original MS-1MV2 (Training Set 1), and the bottom five trained on the low-resolution-augmented MS-1MV2 (Training Set 3). In Table 8, the top 5 rows are trained on the mask-augmented MS-1MV2 (Training Set 2), and the bottom five trained on the low-resolution-mask-augmented MS-1MV2 (Training Set 4).

Tables 9 and 10 show the identification rates on the original SCface and the masked SCface, respectively, at three distances, 4.20 m (d1), 2.60 m (d2), and 1.00 m (d3). The training sets are arranged the same way as for Tables 7 and 8. As the identification rates on the SCface are commonly reported for the Rank-1 at three distances, the CMC curves are skipped and only the ROC curves for verification are shown. The ROC curves for training on the original MS-1MV2 and testing on the original SCface are shown in Figure 11(a). The ROC curves for training on the mask-augmented MS-1MV2 and on the mask-low-resolution augmented MS-1MV2, and testing on the masked SCface are shown in Figure 11(b). The performance shown in these tables and figures can be summarized as follows:

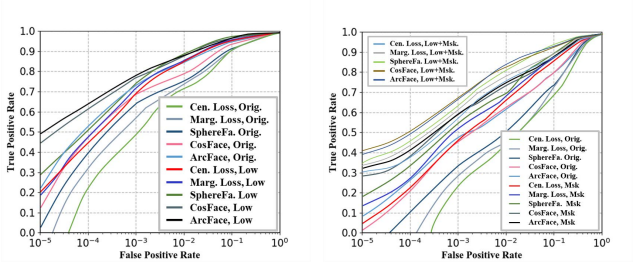
TABLE 9. Rank-1 identification rates for the loss functions tested on the original SCface (no mask) subsets of three distances (d_1 the farthest). Top 5 rows trained on 1) the original MS-1MV2, the bottom 5 rows with $_{low}$ trained on 3) the low-resolution-augmented MS-1MV2.

Model	d_1	d_2	d_3	avg.
Center Loss [2]	22.7	66.4	85.2	58.1
Marginal Loss [3]	27.6	68.7	89.8	62.0
SphereFace [4]	32.5	76.1	93.8	67.4
CosFace [5]	41.9	83.9	97.1	74.3
ArcFace [6]	51.1	87.5	97.7	78.7
Center Loss$_{low}$	53.9	79.4	88.6	73.9
Marginal Loss$_{low}$	58.5	82.2	90.1	76.9
SphereFace$_{low}$	60.2	86.1	94.3	80.2
CosFace$_{low}$	64.8	**92.5**	**98.4**	85.2
ArcFace$_{low}$	**69.1**	92.3	98.2	**86.5**

- Again, masks substantially degrade the performance of the networks trained on faces without masks. The best network, ArcFace, trained on the original MS-1MV2, gives verification rate 74.3% on the original SCface @FAR 10^{-3}, but drops to 48.4% on the masked SCface.
- The degraded performance is worsened for identification. At the farthest d1, the ArcFace trained on the original MS-1MV2 gives identification rate 51.1% on SCface, but drops to 43.3% on the masked SCface.
- The training on the mask-low-resolution augmented MS-1MV2 can effectively improve the performance. The ArcFace$_{low+mask}$ shows verification rate 67.7% and identification rate 60.7% on the masked SCface.
- For the cases without masks, the training on the low-resolution-augmented MS-1MV2 improves the performance. The ArcFace$_{low}$ shows verification rate 79.1% and identification rate 69.1% on SCface, as shown in Tables 7 and 9.

4) PERFORMANCE ON REAL MASKED FACE DATASET

To better evaluate the performance on real masked faces, we select the best three loss functions from the previous comparisons, and test them on the c-RMFRD dataset. As many faces in the c-RMFRD are with large poses, we define a subset of the c-RMFRD with facial orientation larger than 60° in yaw removed, and label it as c-RMFRD$_o$. The only difference between the c-RMFRD and c-RMFRD$_o$ is the 165 large-pose faces removed in the latter. The top 6 rows in Table 11 and Table 12 show the verification and identification rates on the



(a) ROC on orig. SCface. (b) ROC on masked SCface$_{mask}$.

FIGURE 11. Left: ROC of using 1) the original MS-1MV2 for training and testing on the original SCface. Right: ROC of using 2) the mask-augmented MS-1MV2 for training and testing on the synthetic masked SCface$_{mask}$.

TABLE 10. Rank-1 identification rates for the loss functions tested on the masked SCface. Top 5 rows are trained on 1) the original MS-1MV2, middle 5 rows with $_{low}$ are trained on 2) the mask-augmented MS-1MV2, and the bottom 5 rows with $_{low+mask}$ are trained on 4) the mask-low-resolution augmented MS-1MV2

Model	d$_1$	d$_2$	d$_3$	avg.
Center Loss [2]	17.6	50.4	76.4	48.1
Marginal Loss [3]	22.6	58.7	80.1	53.8
SphereFace [4]	28.1	66.4	88.3	60.2
CosFace [5]	35.7	75.1	92.6	67.8
ArcFace [6]	43.3	79.2	94.8	72.4
Center Loss$_{mask}$	29.7	67.3	88.2	61.7
Marginal Loss$_{mask}$	34.3	71.2	90.9	65.4
SphereFace$_{mask}$	38.3	76.9	93.7	69.6
CosFace$_{mask}$	44.1	82.3	96.1	74.1
ArcFace$_{mask}$	44.6	84.8	96.7	75.3
Center Loss$_{low+mask}$	42.6	73.5	89.7	68.6
Marginal Loss$_{low+mask}$	46.5	76.1	92.1	71.5
SphereFace$_{low+mask}$	52.3	80.4	94.9	75.8
CosFace$_{low+mask}$	54.7	87.3	96.8	79.6
ArcFace$_{low+mask}$	**60.7**	**89.1**	**97.5**	**82.4**

TABLE 11. Verification rates (in TAR%, AUC%) for the loss functions tested on the c-RMFRD. Top 6 rows tested on 1) the original c-RMFRD. The bottom 6 rows tested on 2) the c-RMFR$_o$ without large pose.

Model	TAR(%)@FAR			AUC (%)
	0.1	0.01	0.001	
SphereFace [4]	61.7	25.4	7.5	85.3
CosFace [5]	52.7	18.4	5.1	82.3
ArcFace [6]	76.4	45.1	27.1	90.8
SphereFace$_{mask}$	77.6	41.8	20.0	92.3
CosFace$_{mask}$	61.9	28.2	10.6	85.3
ArcFace$_{mask}$	**82.8**	**52.8**	**26.9**	**93.6**
SphereFace [4]	64.3	26.0	8.7	87.2
CosFace [5]	59.0	24.7	8.6	84.4
ArcFace [6]	83.2	52.6	28.9	93.9
SphereFace$_{mask}$	84.1	46.0	18.1	94.6
CosFace$_{mask}$	71.5	38.7	17.0	89.5
ArcFace$_{mask}$	**89.9**	**66.6**	**41.4**	**96.4**

c-RMFRD, respectively; the bottom 6 rows in Table 11 and Table 12 show the verification and identification rates on the c-RMFRD$_o$. The results can be described as follows:

TABLE 12. Identification rates at FPIR = 0.01, 0.1 and Rank-1, 5 for the loss functions tested on c-RMFRD. Top 6 rows tested on 1) the original c-RMFRD. The bottom 6 rows tested on 2) the c-RMFR$_o$ without large pose.

Model	FPIR=0.01	FPIR=0.1	Rank-1	Rank-5
SphereFace [4]	2.2	6.6	20.4	42.7
CosFace [5]	1.7	7.1	21.8	41.5
ArcFace [6]	9.5	22.9	43.2	64.7
SphereFace$_{mask}$	4.4	14.6	42.7	66.5
CosFace$_{mask}$	4.5	14.1	33.1	53.6
ArcFace$_{mask}$	**19.2**	**35.3**	**53.8**	**75.3**
SphereFace [4]	3.2	11.6	26.2	50.5
CosFace [5]	7.3	14.3	27.2	49.1
ArcFace [6]	16.7	28.1	51.6	72.4
SphereFace$_{mask}$	13.5	26.4	60.2	82.7
CosFace$_{mask}$	12.4	21.3	42.1	65.4
ArcFace$_{mask}$	**23.5**	**46.2**	**67.8**	**83.5**

- The mask-augmented training, labeled by the subscript $_{mask}$, yields a better performance than the training without the synthetic masks. This indicates that the learning based on synthetic facial masks can assist the masked face recognition.
- Facial pose is a major challenging factor for masked face recognition. Table 11 shows that the verification rate at FAR 0.1% drops from 89.9% on the c-RMFR$_o$ to 82.8% on the c-RMFR. The Rank-1 identification rate drops from 67.8% (c-RMFRD$_o$) to 53.8% (c-RMFRD). The portion of large-pose data in the c-RMFRD is larger than those in the IJB-B and IJB-C, indicating that the MS-1MV2 training set does not have a sufficient portion of large-pose faces. A few intra-pair samples considered in the verification test from the c-RMFRD are displayed in Figure 12. Many masked faces do not just appear in large pose, but also reveal limited visible facial regions because of the partial occlusion made by the caps or hair.
- Although the ArcFace still outperforms all, followed by the SphereFace and then the CosFace, their performances are apparently worse than those in the previous experiments on the synthetic masked faces. Figure 13 and Figure 14 are samples of the intra (genuine) pairs and extra (imposter) pair, respectively, and all failed to be verified by using the ArcFace. Visual inspection of these and other cases shows that the causes for the failures can be the aforementioned large poses and limited visible facial regions. Many faces in the c-RMFRD are wearing caps, but the synthetic masked faces in the MS-1MV2 training set are clear in the forehead and hair regions. Moreover, the different races in the training and testing datasets can also be a cause. The majority of MS-1MV2, IJB-B and IJB-C are Caucasians, but the faces in c-RMFRD are all Asians.

Since we found that we had difficulties to manually verify many masked faces in the c-RMFRD$_o$, we conducted a human

FIGURE 12. Intra (genuine) pair samples from the c-RMFRD with masked faces in large pose, and visible regions affected by caps and hair.

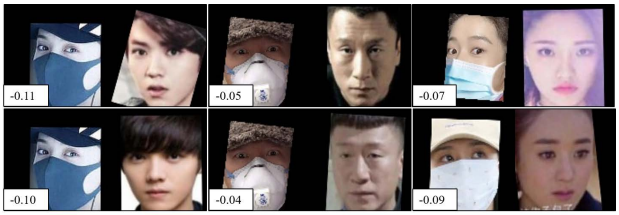


FIGURE 13. Samples of intra (genuine) pairs failed to be verified by ArcFace, the number is the cosine similarity and threshold is 0.31 for FAR = 0.1%.

FIGURE 14. Samples of extra (imposter) pairs failed to be verified by ArcFace, the number is the cosine similarity and threshold is 0.31 for FAR = 0.1%.

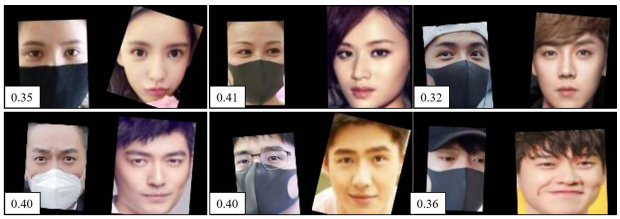


FIGURE 15. Samples of errors made by human inspectors. Each pair denotes the same subject in label. The number denotes the cosine similarity

TABLE 13. Verification and identification rates of ArcFace versus human test on RMFRD.

Verification Rate (FAR∼0.1%)	89.9 % (ArcFace)	69.2 % (Human)
Identification Rate	67.8 % (ArcFace)	26.2 % (Human)

TABLE 14. Performance on the AR Face database [35], MFR2 database [23].

Verification			
AR Face		MFR2	
Model	AUC (%)	Model	AUC (%)
Li et al. [22]	98.4	Anwar et al. [23]	95.99
FocusFace [20]	99.2	FocusFace [20]	96.56
Ours	**99.8**	Ours	**97.48**
Identification			
AR Face		MFR2	
Model	Rank-1	Model	Rank-1
Li et al. [19]	98.0	Anwar et al. [23]	94.81
FocusFace [20]	99.1	FocusFace [20]	95.99
Ours	**99.6**	Ours	**96.78**

user test to compare human performance with the ArcFace. For this test five inspectors manually selected their individual rank-1 faces for the identification test, and verified whether each test pair were the same subject for the verification test. The majority of the 5 votes was taken as the legitimate outcome, and the comparison is shown in Table 13. As the human FAR for the verification test is close to 0.1%, we take the corresponding TAR from Table 11 for the comparison. To our surprise, there is a clear performance gap between the ArcFace and human; especially for the Rank-1 identification, the gap is more than 40%. Figure 15 shows a few intra-pair samples that the inspectors failed but the ArcFace succeeded to verify. The reason for the large performance gap can be that the ArcFace model has been trained by the synthetic masked faces so that it performs well recognizing real masked faces, but the human inspectors are not used to recognizing masked faces. It is therefore experimentally verified that the training on synthetic masked faces helps the real masked face recognition.

5) COMPARISON WITH RECENT APPROACHES

We compare our best model with the ArcFace loss with several recent methods [19], [20], [22], [23] which are reviewed in Sec.II. Note that the training and testing protocol considered in [18] is different from ours, it is excluded from this comparison. The comparison is shown in Table 14. Our model outperforms all in both the verification and identification tasks on the AR and MFR2 datasets.

We offer our code for synthesizing the masks, two trained models and the c-RMFRD dataset in our project GitHub site, https://github.com/AvLab-CV/Face_Mask_Generator.

VI. CONCLUSION

As the COVID-19 virus may stay with us for a while, wearing a mask can be a common or mandatory action to take for our protection. It is challenging to recognize masked faces as most facial features are covered, substantially degrading the human vision which performs exceptionally well recognizing uncovered faces. This study shows that a deep learning solution trained on synthetic masked faces can outperform human for masked face recognition. We also show that masked face recognition can be made more challenging when considering other compounding factors, including age, image resolution and facial pose. How these factors affect the generic face recognition, i.e., recognizing faces without masks, is also studied with extended experiments. Therefore, how to improve the masked face recognition against the compounding factors will be a potential topic for future research.

The inspection of the errors in the synthesized masked faces when tackling each compounding factor can be an important step. Additionally, the making of more databases cannot be overemphasized. The RMFRD is composed of primarily Asian faces, the datasets of other ethnicities are needed for the understanding of cross-ethnicity issues. We have been working on some of these issues, and will report new results when available.

REFERENCES

[1] *Use and Case of Masks*. Accessed: Sep. 20, 2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/about-face-coverings.html

[2] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. ECCV*. Amsterdam, The Netherlands: Springer, 2016, pp. 499–515.

[3] J. Deng, Y. Zhou, and S. Zafeiriou, "Marginal loss for deep face recognition," in *Proc. CVPRW*, 2017, pp. 60–68.

[4] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. CVPR*, 2017, pp. 212–220.

[5] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. CVPR*, 2018, pp. 5265–5274.

[6] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4690–4699.

[7] B. Gary Huang, M. Ramesh, T. Berg, and E. Learned-Miller. (2007). *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. [Online]. Available: http://vis-www.cs.umass.edu/lfw/

[8] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A," in *Proc. CVPR*, 2015, pp. 1931–1939.

[9] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *Proc. CVPR*, 2011, pp. 529–534.

[10] C. Whitelam, "IARPA Janus benchmark-b face dataset," in *Proc. CVPRW*, 2017, pp. 90–98.

[11] B. Maze, "IARPA Janus benchmark-C: Face dataset and protocol," in *Proc. ICB*, 2018, pp. 158–165.

[12] A. Lanitis, J. C. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 442–455, 2002.

[13] M. Grgic, K. Delac, and S. Grgic, "Scface–surveillance cameras face database," *Multimedia Tools. Appl.*, vol. 51, no. 3, pp. 863–879, 2011.

[14] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Proc. ECCV*. Amsterdam, The Netherlands: Springer, 2016, pp. 87–102.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

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

[17] G.-S.-J. Hsu, H.-Y. Wu, and M. H. Yap, "A comprehensive study on loss functions for cross-factor face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 826–827.

[18] Wenbo Zheng, Yan, Lan Fei-Yue Wang, and Chao Gou, "Learning from the web: Webly supervised meta-learning for masked face recognition," in *Proc. CVPRW*, 2021, pp. 4304–4313.

[19] C. Li, S. Ge, D. Zhang, and J. Li, "Look through masks: Towards masked face recognition with de-occlusion distillation," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3016–3024.

[20] P. C. Neto, F. Boutros, J. R. Pinto, N. Damer, A. F. Sequeira, and J. S. Cardoso, "FocusFace: Multi-task contrastive learning for masked face recognition," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Dec. 2021, pp. 1–8.

[21] F. Boutros, "MFR 2021: Masked face recognition competition," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Aug. 2021, pp. 1–10.

[22] Y. Li, K. Guo, Y. Lu, and L. Liu, "Cropping and attention based approach for masked face recognition," *Appl. Intell.*, vol. 51, no. 5, pp. 3012–3025, 2021.

[23] A. Anwar and A. Raychowdhury, "Masked face recognition for secure authentication," 2020, *arXiv:2008.11104*.

[24] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. FG*, 2018, pp. 67–74.

[25] G. Pereyra, G. Tucker, J. Chorowski, Å. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," 2017, *arXiv:1701.06548*.

[26] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem (and a dataset of 230,000 3D facial landmarks)," in *Proc. ICCV*, 2017, pp. 1021–1030.

[27] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*, 2016, pp. 483–499.

[28] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. CVPR*, 2016, pp. 146–155.

[29] D. E. King, "DLIB-ML: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jul. 2009.

[30] F. Yang, W. Yang, R. Gao, and Q. Liao, "Discriminative multidimensional scaling for low-resolution face recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 388–392, Mar. 2018.

[31] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2051–2062, Apr. 2019.

[32] P. Li, L. Prieto, D. Mery, and P. J. Flynn, "On low-resolution face recognition in the wild: Comparisons and new techniques," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 8, pp. 2000–2012, Aug. 2019.

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

[34] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems," 2015, *arXiv:1512.01274*.

[35] A. M. Martinez, "The ar face database," CVC, New Delhi, India, Tech. Rep. 24, 1998.

GEE-SERN JISON HSU (Senior Member, IEEE) received the dual M.S. degree in electrical and mechanical engineering and the Ph.D. degree in mechanical engineering from the University of Michigan, Ann Arbor, MI, USA, in 1993 and 1995, respectively. From 1995 to 1996, he was a Postdoctoral Fellow with the University of Michigan. From 1997 to 2000, he was a Senior Research Staff with the National University of Singapore. In 2001, he joined Penpower Technology, where he led research on face recognition and intelligent video surveillance. His team at Penpower Technology was a recipient of the Best Innovation and Best Product Award from the SecuTech Expo for three consecutive years from 2005 to 2007. In 2007, he joined the Department of Mechanical Engineering, National Taiwan University of Science and Technology (NTUST), where he is currently a Professor. His research interests include computer vision and deep learning. He is a member of IAPR. He received the Best Paper Awards in ICMT 2011, CVGIP 2013, CVPRW2014, ARIS 2017, and CVGIP 2018.

HUNG-YI WU received the B.S. degree in mechanical engineering from the National Changhua University of Education, Changhua, Taiwan, in 2019. He is currently pursuing the M.S. degree in mechanical engineering with the National Taiwan University of Science and Technology, Taipei, Taiwan. His research interests include face recognition and deep learning.

CHUN-HUNG TSAI received the B.S. degree in mechanical engineering from Yuan Ze University, Taoyuan, Taiwan, in 2020. He is currently pursuing the M.S. degree in mechanical engineering with the National Taiwan University of Science and Technology, Taipei, Taiwan. His research interests include face reenactment and face recognition.

SVETLANA YANUSHKEVICH (Senior Member, IEEE) received the Dr.Sci. (Habilitation) degree from the Technical University of Warsaw, in 1999. She is currently a Professor with the Department of Electrical and Software Engineering (ESE), Schulich School of Engineering, University of Calgary. She directs the Biometric Technologies Laboratory, University of Calgary, the only research facility dedicated to biometric systems design in Canada. She was with the West-Pomeranian University of Technology, Szczecin, Poland, prior to joining the ESE Department, University of Calgary, in 2001. She contributed to the area of artificial intelligence for digital design and biometrics, since 1996. Most recently, she and her team have developed novel risk, trust and bias assessment strategies based on machine reasoning, with applications in biometric-enabled border control, forensics, and healthcare.

MARINA L. GAVRILOVA (Senior Member, IEEE) is currently a Full Professor with the Department of Computer Science, University of Calgary, the Head of the Biometric Technologies Laboratory, and a Board Member of ISPIA. Her publications include over 200 journals and conference papers, edited special issues, books, and book chapters in the areas of image processing, pattern recognition, machine learning, biometric and online security. She has founded ICCSA–an international conference series with LNCS/IEEE, co-chaired a number of top international conferences. She is the Founding Editor-in-Chief of *LNCS Transactions on Computational Science* journal. She has given over 50 keynotes, invited lectures, and tutorials at major scientific gatherings and industry research centers, including at Stanford University; SERIES Center, Purdue; Microsoft Research USA; Oxford University, U.K.; Samsung Research, South Korea; and others. She currently serves as an Associate Editor for IEEE Access, IEEE Transactions on Computational Social Systems, *The Visual Computer*, and the *International Journal of Biometrics*. She was appointed by the IEEE Biometric Council to serve on IEEE Transactions on Biometrics, Behavior, and Identity Science Committee.