Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
PLFace: Progressive Learning for Face Recognition with Mask Bias
Baojin Huang\textsuperscript{a}, Zhongyuan Wang\textsuperscript{a,*}, Guangcheng Wang\textsuperscript{a}, Kui Jiang\textsuperscript{a}, Zhen Han\textsuperscript{a}, Tao Liu\textsuperscript{b}, Chao Liang\textsuperscript{a}
\textsuperscript{a}NERCMS, School of Computer Science, Wuhan University, Wuhan 430072, China
\textsuperscript{b}School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China

A R T I C L E   I N F O
Article history:
Received 5 November 2021
Revised 3 October 2022
Accepted 2 November 2022
Available online 9 November 2022

Keywords:
Face recognition
Progressive learning
Mask bias

A B S T R A C T
The outbreak of the COVID-19 coronavirus epidemic has promoted the development of masked face recognition (MFR). Nevertheless, the performance of regular face recognition is severely compromised when the MFR accuracy is blindly pursued. More facts indicate that MFR should be regarded as a mask bias of face recognition rather than an independent task. To mitigate mask bias, we propose a novel Progressive Learning Loss (PLFace) that achieves a progressive training strategy for deep face recognition to learn balanced performance for masked/mask-free faces recognition based on margin losses. Particularly, our PLFace adaptively adjusts the relative importance of masked and mask-free samples during different training stages. In the early stage of training, PLFace mainly learns the feature representations of mask-free samples. At this time, the regular sample embeddings shrink to the corresponding prototype, which represents the center of each class while being stored in the last linear layer. In the later stage of training, PLFace converges on mask-free samples and further focuses on masked samples until the masked sample embeddings are also gathered in the center of the class. The entire training process emphasizes the paradigm that normal samples shrink first and masked samples gather afterward. Extensive experimental results on popular regular and masked face benchmarks demonstrate that our proposed PLFace can effectively eliminate mask bias in face recognition. Compared to state-of-the-art competitors, PLFace significantly improves the accuracy of MFR while maintaining the performance of normal face recognition.

© 2022 Elsevier Ltd. All rights reserved.

1. Introduction

In recent years, Convolutional Neural Networks (CNNs) are widely used in the field of computer vision owing to their powerful feature representation capabilities. Great progress has been made in face recognition system [1–5]. For face recognition based on deep learning, existing CNNs [6,7] supervised by margin-based loss functions [8–10] show the ability to fit large-scale face recognition datasets [11,12], but they have been almost carried out on fully exposed faces. This results in great difficulties for masked face recognition (MFR) since people usually wear masks to prevent infection during COVID-19.

At present, current popular face recognition training sets, such as WebFace [11], MS1MV3 [9], Glint360k [12], etc., contain tens of millions of normal face images with tens of thousands of identities. Whereas, it is challenging to collect large-scale masked images with identity information. To cope with this issue, previous researchers [15–17] used GAN-based and landmark-based methods to synthesize masked face images for data augmentation. Simultaneously, the validity of the synthesized data for MFR was also proved by quantitative experiments. On this basis, some dedicated MFR methods [15,16,18] have been proposed and made positive progress. However, they mainly pursue the performance of MFR and ignore the accuracy of normal face recognition, which is not friendly in practical applications. As shown in Fig. 1, we report the performance of masked and normal face recognition with different ratios of masked face augmentation. It is intuitive to explain that the improvement of the accuracy of MFR will inevitably compromise the performance of normal face recognition. Therefore, MFR should be regarded as a special case of face recognition rather than an independent task. In other words, it is expected to improve the performance of masked face recognition while maintaining the accuracy of normal face recognition. We call it to mask bias mitigation.

In this work, we propose a novel progressive learning loss, termed PLFace, to accomplish a reasonable training strategy for deep face recognition with mask bias. As shown in Fig. 2, previous margin-based SOTA methods such as ArcFace [9] and Curric-
ularFace [10] adopt a fixed margin for all samples to expand the inter-class distance and reduce the intra-class distance. There exists the fact that the masked samples are subjectively so different from the mask-free samples. The model that treats them equally will be biased towards the party that fits the masked samples or the mask-free samples. This is perhaps the main reason why fixed-margin-based methods are susceptible to mask bias. Even if CurricularFace [10] proposes the additional margin of negative samples to emphasize easy samples first and hard samples later, the improvement as shown in Fig. 1 is limited since the definition of hard samples in CurricularFace is ambiguous for masked samples. Based on the above analysis, we consider designing an exclusive learning approach for the masked face samples. Particularly, we make the model learn mask-free and masked samples progressively. In the early stage of training, the model first learns the feature representations of mask-free samples. At this time, the mask-free sample features shrink to the prototype. In the later stage of training, the model focuses on masked samples based on the learned normal face recognition knowledge. To achieve the goal of progressive learning in the entire training, we adjust the margin for mask-free samples and masked samples by the two procedures: 1) the margin for mask-free samples remains fixed, and 2) the margin for masked samples increases linearly from 0 with the training process, but does not exceed the margin of mask-free samples.

The advantages of the proposed PLFace can be summarised as follows:

**Effective.** PLFace significantly improves the accuracy of MFR while maintaining the performance of normal face recognition on several face recognition benchmarks, including mask-free and masked datasets.

**Easy.** PLFace can be easily migrated to the existing loss functions of face recognition, e.g., CosFace [8], ArcFace [9], CurricularFace [10].

**Efficient.** PLFace only adds negligible computational complexity during the training process, and has the same cost as the backbone model during the inference process.

2. Related Works

2.1. Margin-based Loss Functions

Face recognition aims to map the raw face image into a feature with small intra-class and large inter-class distance. The loss function is crucial for large-scale face recognition. Most state-of-the-art face recognition methods utilize softmax-based classification loss. The original softmax classification loss [19] can only learn features with large inter-class distances, which is suitable for the closed-set classification problem. To enhance the intra-class compactness, several margin-based [8,9,20] losses employ constant margin between classes to maximize angular margins among feature embeddings of different identities, which result in more discriminative features. Recently, some adaptive margin-based loss functions have been further proposed, such as AdaptiveFace [21], DBM [22], MV-Softmax [23] CurricularFace [10] and MagFace [24], in which the margin can be adjusted automatically. However, these methods adjust the margin from the perspective of difficult and easy samples. They cannot well adapt to the bias caused by masked face augmentation because the concept of difficult and easy samples is ambiguous for masks.

2.2. Masked Face Recognition

As a special case of face recognition, masked face recognition technology has received great attention in the early stage of the COVID-19 coronavirus epidemic. As the deep learning based face recognition approaches, the primary challenge of MFR is the lack of large-scale labeled masked face datasets. To cope with this issue, Wang et al. [25] released three types of masked face datasets freely available for the first time, including Masked Face Detection Dataset (MFDD), Real-world Masked Face Recognition Dataset (RMFRD) and Synthetic Masked Face Recognition Dataset (SMFRD). Since then, many MFR methods have emerged, such as LPD [15], DCR [16], CAB [18] and EUM [26]. Since the lower half of the masked face image is occluded, LPD and CAB focus on the half of the face, enabling the model to embed the identity information of the visible part in a targeted manner. DCR and EUM explore the representation differences between masked and normal samples, thereby producing embeddings similar to these of unmasked faces of the same identities. To promote the development of MFR, many international conferences hold related competitions, such as ICCV-MFR-2021 [13], IJCB-MFR-2021 [27] and FG-COVID19 [28]. The winner usually employs practical tricks (data augmentation and knowledge distillation [29]) to improve masked face recognition performance. Nevertheless, many studies [30-32] have been proposed to promote MFR from the perspective of learning strategies and network structures. FocusFace [30] is a multi-task architecture that uses contrastive learning to perform masked face recognition accurately. DCF [31] is proposed to extract the features from unobstructed regions of the faces (i.e., eyes and forehead). These feature maps are employed to obtain covariance-based features. Zhang et al. [32] propose a dual-branch training strategy to guide the model to focus on the upper half of the face for MFR. We can observe that these methods mainly pursue the performance of MFR by utilizing prior information that mask-occlusion is located in the lower half of the face. However, they ignore the accuracy of
normal face recognition, which is not friendly in practical applications. In other words, MFR should be regarded as a special case of face recognition rather than an independent task. It is expected to improve the performance of masked face recognition while maintaining the accuracy of normal face recognition.

2.3. Debiasing Methods in Face Recognition

Due to potential differences in the distribution of datasets (race, gender, etc.), face recognition algorithms are prone to local bias, resulting in unsatisfactory generalization performance. The current methods to eliminate this bias can be summarized into two categories: balancing the dataset and designing a more discriminative loss function. Since bias is an abstract concept, we take race bias as an example. Wang et al. [33] collected a balanced dataset and proved that the performance of the model trained on the balanced dataset is superior to that of the original dataset. From the perspective of model optimization, Gong et al. [34] proposed a novel debiasing adversarial network (DebFace) that learns to extract disentangled feature representations for biased face recognition. Yang et al. [35] designed the race adaptive margin based face recognition (RamFace) model to automatically derive different optimal margins for different races in training the model, which further mitigates the racial bias. The objects of these debiasing methods are specific identities, such as race and gender attributes, which will not change for the same identity. However, the target of mask bias is a specific sample of a certain identity, which means that the samples of a certain identity still cannot be treated equally during training.

3. The Proposed PLFace

In this section, we first briefly introduce the preliminary knowledge of the face recognition loss function, and then specifically detail our proposed PLFace.

3.1. Preliminaries on Loss Functions

As the most widely used classification loss, softmax loss is formulated as follows:

$$L = - \frac{1}{N} \sum_{i=1}^{N} \log \sum_{j=1}^{C} e^{w_j^T x_i + b_j},$$  \hspace{2cm} (1)

where $x_i \in \mathbb{R}^d$ ($d$ is the embedding feature dimension) represents the $i$-th image feature vector in the batch input, which belongs to the $y_i$-th class. $w_j, b_j \in \mathbb{R}^d$ denotes the $j$-th and $y_i$-th column of the weight $W \in \mathbb{R}^{d \times C}$, respectively. $b_j \in \mathbb{R}^d$ is the bias term, $N$ is the batch size, and $C$ is the class number. As illustrated in SphereFace [20], fixing the $b_j = 0$, normalizing the $\|w_j\| = 1$ and re-scaling the embedding $\|x_i\| = s$ are more reasonable for face recognition. Thus the logit is transformed as $w_j^T x_i = \|w_j\| \|x_i\| \cos \theta_j = \cos \theta_j$. The softmax loss can be modified as follows:

$$L = - \frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{c_j \cos \theta_i}}{\sum_{j=1}^{C} e^{c_j \cos \theta_i}},$$  \hspace{2cm} (2)

Where $\theta_j$ is the angle between the embedding feature $x_i$ and the weight $w_j$. Note that in this way $w_j$ can be regarded as the center of the $j$-th class, also known as the prototype of the $j$-th class. It can be seen that the prediction only depends on the angle $\theta_j$. However, Eq. (2) still cannot enhance the intra-identity compactness. To make the loss be discriminative enough for practical face recognition problem, several margin-based variants are proposed and can be formulated in a universal form:

$$L = - \frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{f(\theta_i)}}{\sum_{j=1}^{C} e^{f(\theta_j)}} + \frac{m}{2} \sum_{j=1}^{C} e^{f(\theta_j)}.$$  \hspace{2cm} (3)

ArcFace [9] is the most representative constant margin-based loss function, in which $f(\theta_i) = \cos(\theta_i) + m$, $g(\theta_j) = \cos \theta_j$. Similarly, $f(\theta_i) = \cos(\theta_i) - m$, $g(\theta_j) = \cos \theta_j$ in CosFace [8]. The margin force the $\theta_i$ more diminutive to enhance the intra-class compactness. However, these constant margin loss ignore the distribution of different identities. Recently, Liu et al. proposes the AdaptiveFace [21] to adaptively set the margin for different identity prototypes, i.e., $f(\theta_i) = \cos(\theta_i) - m \eta_i$, $g(\theta_j) = \cos \theta_j$. Except from the perspective of modulating positive cosine similarities, Huang et al. proposes CurricularFace [10] to adaptively adjusts the relative importance of easy and hard samples by modulating the negative cosine similarities. Specifically, the positive and negative cosine similarity functions in Eq. (3) are modified as

$$f(\theta_i) = \cos(\theta_i) + m,$$  \hspace{2cm} (4)

$$g(\theta_j) = \begin{cases} \cos \theta_j, & f(\theta_i) - \cos \theta_j \geq 0 \\ \cos \theta_j (t + \cos \theta_i), & f(\theta_i) - \cos \theta_j < 0. \end{cases}$$  \hspace{2cm} (5)

As shown in Figure 2, the symbol $d$ in (c) and (d) denotes the same margin of negative samples proposed by CurricularFace, which is formulated as

$$d = (t + \cos \theta_j - 1) \cos \theta_j,$$  \hspace{2cm} (6)

where $t$ is set as an adaptive parameter by exponential moving average, see CurricularFace [10] for more details.

Nevertheless, the above loss functions take each identity class as a whole and do not consider the intra-class difference. In other words, the face images of the same identity after masked face augmentation hold apparent discrepancies, which may cause intra-class masked bias. In this paper, we design a novel progressive learning loss function to delay the masked face image approaching the center of the class.

3.2. Analysis of the Mask Bias

Compared with the normal face image, the masked face image lacks a part of the available information (mouth and nose), and thus the masked face recognition suffers from an inevitable performance drop. To enable deep networks to explore more discriminative feature representations of masked samples, researchers empirically mix synthetic masked samples in normal face recognition training datasets. This raises a problem, that the larger the proportion of synthetic mask samples in the training dataset, the more the model trained on this dataset will lose accuracy on normal samples. The model trained with a small proportion of mask samples cannot enjoy the ideal mask face recognition performance. In other words, masked samples in the training set prevent the model from learning discriminative representations of normal faces. To get an insight on mask bias, we further analyze the differences between several bias. As shown in Fig. 3, we enumerate the data distributions with racial and mask bias. Specifically, race is a constant attribute of each subject. The reason for racial bias is the uneven identity of races in the dataset (usually more Caucasian). Therefore, racial bias can be alleviated by non-uniform attention to different racial identities. In contrast, mask bias only appears in the image distribution of the same subject. Thus, mitigating mask bias from the perspective of identity is not enough. Ideally, the model is expected to learn normal face samples without being disturbed by mask samples. Subsequently, mask samples can be further fitted based on the learned knowledge of judging normal samples. Under this analysis, we designed a progressive learning strategy from the sample dimension to alleviate mask bias, which is discussed in the following subsection.
3.3. Progressive Learning Loss

This section introduces the details of our proposed progressive learning loss, which aims to reduce the mask bias for robust face recognition. As aforementioned, our proposed loss function is also derived from Eq. (3), the positive cosine similarity function is defined as

\[ f(\theta_y) = \cos(\theta_y + \alpha \vartheta(\alpha)), \]
\[ \vartheta(\alpha) = \begin{cases} m, & z_i = 0 \\ \alpha \cdot m, & z_i = 1. \end{cases} \]  

(7)

Where the \( z_i \) indicates whether the \( i \)-th face sample wears a mask (1 means yes, 0 means no) and the \( m \) is the same fixed margin as in Eq. (4). \( \{\cdot\} \) denotes the block gradient operator. The \( \alpha \) denotes the training progress coefficient, which will be discussed later. It should be noted that the negative cosine similarity can employ any margin-based loss functions, and here we adopt CurricularFace as an example. Therefore, \( g(\theta) \) is the same as in Eq. (5). Note that the parameter \( t \) in Eq. (5) for our PLFace is set in the same adaptive way as CurricularFace. We have formulated an additional Eq. (6), where the parameter \( d \) for our PLFace is included. The advantage of our proposed PLFace lies in the progressive learning of masked samples during the training process by adjusting the positive cosine similarity of the masked samples. In other words, we expect that in the early stage of training, the model first learns the feature representations of mask-free samples. At this time, the mask-free sample features shrink to the prototype. In the later stage of training, the model focuses on masked samples based on the learned normal face recognition knowledge. We call the entire training process progressive learning.

It is critical to determine appropriate \( \alpha \) for progressive training. Ideally, the value of \( \alpha \) in the early stage of training is small, and \( \alpha \) in the later stage of training is relatively larger. Inspired by CurricularFace \cite{Huang2019}, we adopt the average of positive cosine similarities of the mask-free sample to indicate the training stages, which is formulated as

\[ \alpha = \frac{\sum_{i=1}^{N} (1 - z_i) \cos \theta_y}{N - \sum_{i=1}^{N} z_i}. \]  

(8)

It can be clearly seen that we choose ArcFace as the basis for \( f(\theta_y) \) and dynamically adjust the margin to control the training process. Therefore, at the early stage of training, the value of \( \theta(\alpha) \) when \( z_i = 1 \) is close to 0, so that the model tends to distinguish mask-free samples. As the training progresses, mask samples gradually become the main focus of the model. Alternatively, the cosine similarities of the mask-free sample increase and the value of \( \theta(\alpha) \) when \( z_i = 1 \) increases accordingly. Thus, the masked samples are emphasized with larger weights. At the later stage of training, the model converges on mask-free samples, and the features of mask-free samples approach the prototype. At this time, the model fully focuses on the learning of masked samples until the features of masked sample are also gathered in the center of the class. Moreover, our PLFace does not introduce any hyperparameters, which eliminates the complexity of manual model tuning.

3.4. Optimization

In this section, we show how our PLFace can be easily optimized by the commonly used stochastic gradient descent (SGD) and analyse the entire progressive training process from an optimization perspective. Here we take \( N \) as 1 in Eq. (3) to simplify the representation. In the forward process, when the input is mask-free, it is the same as the CurricularFace. Moreover, the parameter \( t \) of CurricularFace is set to be adaptive. Specifically, parameter \( t \) reflects the training stage of the model, and the later the training stage is, the larger the value of \( t \) is. Our proposed PLFace aims to adjust positive cosine similarities to achieve progressive learning for masked samples. Thus the value of parameter \( t \) remains the same as in CurricularFace (adaptive). When the input is masked sample, the target logit \( f(\theta_y) = \cos(\theta_y + \alpha \cdot m) \). In the backward propagation process, the gradient with respect to the face embedding is computed as

\[ \frac{\partial L}{\partial \theta_y} = \frac{\partial L}{\partial f(\theta_y)} \frac{\partial f(\theta_y)}{\partial \cos(\theta_y)} \frac{\partial \cos(\theta_y)}{\partial \theta_y} \frac{\partial \theta_y}{\partial \theta_y} = s(P_j - 1) \frac{\sin(\theta_y + \alpha \vartheta(\alpha))}{\sin(\theta_y)} w_{yj}, \]  

(9)

where \( P_j = \frac{e^{s(\theta_y)}}{e^{s(\theta_y)} + \sum_{j=1 \neq y} e^{s(\theta_j)}} \) and \( w_{yj} \) denotes the \( y_j \)-th column of the weight \( W \). We can find that the gradient of the face sample is determined by \( H(x) = \frac{\sin(\theta_y + \alpha \vartheta(\alpha))}{\sin(\theta_y)} \), which consists of two parts, the positive cosine similarity \( \cos \theta_y \) and the value of \( \vartheta(\alpha) \). The
first derivative of \(H(\cdot)\) is
\[
\frac{\partial H}{\partial \theta_{yi}} = -\sin(\theta(\alpha)), \\
\frac{\partial H}{\partial \theta} = \frac{\sin^2(\theta_{yi})}{\cos(\theta_{yi} + \theta(\alpha))}, \\
\frac{\partial \theta(\alpha)}{\partial \alpha} = \frac{\sin(\theta_{yi})}{\sin(\theta(\alpha))},
\]
(10)

Here we take \(m\) as 0.5. Based on Eq. (8), we have \(\alpha \in [0, 1]\), thus \(\theta(\alpha) \in [0, 0.5]\). It is obvious that \(\frac{\partial H}{\partial \theta} < 0\), which proves that the function \(H(\cdot)\) is monotonically decreasing with respect to \(\theta_{yi}\). Therefore, the gradient of the face embedding keeps the same changing law with \(H(\cdot)\) and \(\theta(\alpha)\). Specifically, when the input is masked sample, the \(\alpha\) gradually increases as the training progresses. Therefore, the masked face embedding gradient becomes larger at the later stage, emphasizing the masked samples. As we excepted, masked samples are suppressed in early training stage but emphasized later, which is consistent with our envisioned progressive learning philosophy. To sum up, the entire training process is summarized in Algorithm 1.

### Algorithm 1: PLFace

**Input:** The deep feature \(x_i\) of \(i\)-th sample with identity label \(y_i\) and masked label \(z_i\), last fully-connected layer parameters \(W\), cosine similarity \(\cos\theta\) of two vectors, embedding network parameters \(\Theta\) and margin \(m\).

**Initialization:** \(m \leftarrow 0.5\); Randomly initialize the parameter \(W\) and \(\Theta\).

**while not converged do**

Compute the training coefficients \(\alpha\) by Eq. 8:
- if \(z_i = 0\) then
  \[
  f(\theta_{yi}) = \cos(\theta_{yi} + m); \\
  \]
- else
  \[
  f(\theta_{yi}) = \cos(\theta_{yi} + \alpha \cdot m); \\
  \]
end

Compute the negative cosine similarity \(g(\theta)\) by Eq. (5);
Compute the PLFace loss \(L\) by Eq. (3);
Update the parameters \(W\) and \(\Theta\) by Stochastic Gradient Descent (SGD);

**end**

**Output:** Parameters \(W\) and \(\Theta\).

3.5 Discussions with SOTA Loss Functions

3.5.1 Comparison with AdaptiveFace and MagFace

We first discuss the difference between our PLFace and the two margin-based competitors, AdaptiveFace [21] and MagFace [24], from the perspective of the margin value in Table 1. We should point out that the negative cosine similarity of AdaptiveFace and MagFace is \(\cos\theta\), thus the performance of the model is only determined by the target logit. AdaptiveFace introduces an adaptive learning strategy to adjust the margins for different classes adaptively. But in the masked face dataset, the mask data augmentation is oriented to all identities, i.e., each class contains masked samples. Therefore, it is not enough to adjust the margin between classes for face recognition with mask bias. MagFace introduces an adaptive mechanism to learn a well-structured within-class feature distribution by pulling easy samples to class centers while pushing hard samples away. Nevertheless, as the number of masked samples in the dataset increases, the quality assessment of MagFace for normal faces will be overwhelmed by masked faces. In other words, too many “noisy” samples can mislead the model. Conversely, we divide the training of the model into two stages, emphasizing the normal and masked face samples, respectively. This progressive learning approach makes the model more prioritized and thus avoids mask bias which reduces the robustness of the model.

| Table 1 |
|-----------------|
| Loss | Target logit | Margin value |
| Softmax | \(\cos \theta_{yi}\) | None |
| SphereFace | \(\cos (m \theta_{yi})\) | Constant |
| CosFace | \(\cos \theta_{yi} - m\) | Constant |
| ArcFace | \(\cos (\theta_{yi} + m)\) | Constant |
| CurricularFace | \(\cos (\theta_{yi} + m)\) | Adaptive learning for each class |
| MagFace | \(\cos (\theta_{yi} + m(\alpha))\) | Increase with the magnitude of i-th sample |
| PLFace (Ours) | \(\cos (\theta_{yi} + \theta(\alpha))\) | Constant (mask-free) \(\times\) the training process (masked) |

3.5.2 Comparison with CurricularFace

The essence of CurricularFace [10] is to differentiate between hard and easy samples based on the positive and negative aspects of \(f(\theta_{yi}) - \cos\theta_{yi}\) in Eq. (5). Thus it emphasizes easy samples first and hard samples later during the training process. However, the definition of hard samples in CurricularFace is ambiguous for masked samples, since additional masked samples are actually deviates from normal samples and cannot be simply distinguished by difficulty. In contrast, our PLFace is based on the characteristics of masked data augmented with known mask labels, and adopts a progressive learning approach to learn normal samples first and then masked samples. The whole training process is clearer and more reasonable while not introducing additional parameters.

4. Experiments

4.1 Implementation Details

4.1.1 Datasets

We employ MS1MV3 [9] and its masked version MS1MV3+MA-X augmented by existing method [14] as our training dataset. The mask templates for masked augmentation are shown in Fig. 4. MS1MV3 contains about 5.1M face images of 9.3K subjects and "MA-X" means masked face augmentation with a specific probability of X. For evaluation, we adopt RMFRD [17] and ICCV2021-MFR-MASK [13] to report the performance of masked faces. In order to test the regular face recognition performance, we also evaluate our model on several popular testing datasets, including LFW [36], CFP-FF [37] and AgeDB-30 [38], IJB-C [39] and ICCV2021-MR-ALL [13]. All images are aligned to \(112 \times 112\) according to the settings in ArcFace [9].

4.1.2 Baselines

We reimplement state-of-the-art baselines, including CosFace [8], ArcFace [9], CurricularFace [10], MagFace [24], LPD [15], FocusFace [30] and EUM [26]. Particularly, LPD, FocusFace and EUM are the representative algorithms for masked face recognition. For each model, we separately adopt ResNet50 and ResNet100 in ArcFace as the embedding network and use the recommended hyperparameters (e.g., \(s = 64\), \(m = 0.5\) for CurricularFace).

4.1.3 Evaluation Metrics

For all the face recognition models in this paper, we evaluate their 1:1 face verification performance. Specifically, the accuracy of
face verification refers to the true positive rate (TPR) or false rejection rate (FRR) under a certain false acceptance rate (FAR). For the RMFRD and ICCV2021-MFR-MASK datasets, the true acceptance rate (TAR) is measured with mask-to-unmask 1:1 protocol under FAR less than $10^{-4}$. For other datasets, TAR is measured on all-to-all 1:1 protocol under FAR less than $10^{-6}$. In addition, we also report the receiver operating characteristic (ROC) curve to compare the performance of the models intuitively.

4.1.4. Training Settings

Our PLFace is implemented by Pytorch framework. We train models on two NVIDIA 3090 (24GB) GPUs by stochastic gradient descent (SGD) with momentum 0.9 and weight decay 5e-4. Particularly, we set the embedding dimension to 512 and the batch size to 512 during the training process. The learning rate starts from 0.1 and is divided by 10 at 10, 18, 24 epochs, and we stop the training at the 28 epoch. Following ArcFace [9], the scale $s$ and margin $m$ are set to 64, 0.5 in our PLFace, respectively.

4.2. Ablation Study

In this subsection, we only discuss the margin acting on the positive sample, and all the rest settings are kept consistent with CurricularFace, including the negative cosine similarity function, etc.

4.2.1. Effects on Same vs. Different Margins between Masked and Mask-free Samples

We first analyse the effect of different margins between masked and mask-free samples. Specifically, we choose three fixed values to replace the $\vartheta(\alpha)$ when $z_{i} = 1$ in Eq. (7) for comparison. 0.5 (i.e., CurricularFace) means the masked and mask-free samples own the same margin during the training. In contrast, 0.45 and 0.4 imply that the masked samples have a smaller margin, making the model easier to be optimized. Table 2 reports the performance of the corresponding model on normal face datasets (LFW, CFP-FF, AgeDB-30 and MR-ALL) and masked face dataset (RMFRD and MR-MASK), which shows that it is more effective to learn masked and mask-free samples with different margins.

4.2.2. Effects on Fixed vs. Progressive Margins for Masked Samples

To show the effectiveness of our proposed progressive margin, we conduct the same experiment with progressive margin and a fixed margin. From the results of the last row in Table 2, our proposed PLFace achieves extraordinary performance on the MR-ALL while maintaining competitive accuracy on the MFR-MASK. This demonstrates that our proposed progressive learning strategy adaptively adjusts the relative importance of masked and mask-free samples during different training stages, thus effectively alleviating the mask bias. Simply using smaller margins for masked samples still cannot eliminate the accuracy drop on normal samples caused by masked data augmentation. To visualize the effect of our progressive margin on model training, we plot the average positive cosine similarities of PLFace and CurricularFace during training, respectively. As shown in Fig. 5, we can see that three jumps occur on both models in terms of average positive cosine similarity, which is due to the fact that the learning rate of the model training undergoes three decays. Specifically for CurricularFace, the average positive cosine similarities are flat for the masked and mask-free samples during the entire training process. This implies that even if CurricularFace uses a strategy of learning the easy samples first and then the difficult samples, it would not highlight the differences between the mask-free samples and masked samples. In contrast, the positive cosine similarities of our PLFace for the mask-free samples are basically above that of the masked samples at the first stage of training (before the third leap). As expected, the positive cosine similarity of mask-free samples is unchanged, while the masked samples quickly rise to the same level at the second stage (after the third leap). This shows that the masked samples do not interfere with the learning of the mask-free samples at the early training stage, and the later model mainly focuses on the masked samples based on the learned mask-free face discrimination ability.

4.3. Comparisons with SOTA Methods

In this section, we compare our PLFace with the SOTA competitors on various benchmarks. We should point that our progressive learning strategy can be adapted to any existing loss functions by replacing the positive cosine similarity function $f(\theta_{p})$ in Eq. (3). Particularly, PLFace-Arc and PLFace-Cur denote ArcFace and CurricularFace after adopting progressive learning, respectively.

4.3.1. Results on RMFRD and ICCV2021-MFR

To validate the performance of our PLFace for face recognition with mask bias, we evaluate our PLFace and other STOA face recognition methods on RMFRD and ICCV2021-MFR. RMFRD contains 7178 masked and unmasked sample pairs. The ICCV2021-MFR is divided into two test benchmarks: MFR-MASK and MFR-MR. MFR-MASK refers to the masked face testing set, covering 6,964 masked images and 13,928 mask-free images of 6,964 identities. There are a total of 13,928 positive pairs and 96,983,824 negative pairs. MR-ALL denotes the multi-racial testing set, containing 242,143 identities and 1,624,305 images.

From the results in Table 3, we can see several important observations. Firstly, the mainstream baseline models (e.g., CosFace, ArcFace, CurricularFace, MagFace) trained on MS1MV3 (X=0) show impressive performance on normal samples. However, their accuracy on masked samples decreases quarterly because of the domain gap and the lack of masked data augmentation. Secondly, the results of ArcFace with ResNet50 as the backbone shows

| $\vartheta(\alpha)$ | LFW | CFP-FF | AgeDB | RMFRD | MFR-MASK | MR-ALL |
|--------------------|-----|--------|-------|-------|----------|--------|
| Masked             |     |        |       |       |          |        |
| 0.5                | 99.72 | 97.63 | 97.53 | 86.84 | 80.24    | 71.18  |
| 0.45               | 99.80 | 97.90 | 98.01 | 87.10 | 80.34    | 72.27  |
| 0.4                | 99.75 | 98.06 | 98.10 | 87.29 | 80.42    | 73.87  |
| Progressive        | 99.73 | 98.23 | 98.27 | 87.33 | 80.00    | 77.51  |
that mask bias indeed exists in the existing method. As the masked face augmentation data increases, the accuracy of ArcFace at MFR-MASK (63.85% → 79.27%) increases while the accuracy at MFR-MR (80.53% → 69.01%) decreases sharply. In particular, the mask bias is particularly prominent in the proprietary masked face recognition method LPD, i.e., outstanding performance on MFR-MASK and inferior performance on MFR-MR. Thirdly, the comparable results of PLFace quantitatively verify the effectiveness of our progressive learning strategy on face recognition with mask bias. Particularly, after adopting progressive learning loss on MS1MV3+MA-1.0 dataset, the PLFace-Arc (76.34% vs. 74.18%) and PLFace-Cur (78.76% vs. 75.71%) render more balanced mean performance than original ArcFace and CurricularFace with backbone ResNet50. Compared to the recent MagFace with backbone ResNet50, our PLFace-Cur also achieves better mean performance (78.76% vs. 73.63%). It is surprising that CosFace performs better than CurricularFace in terms of mean accuracy. We conjecture that masked face samples interfere with the hard-mining strategy of CurricularFace. Specifically, CurricularFace emphasizes learning easy samples first and then learning hard ones. However, masked samples can easily be directly considered as hard samples, which is not conducive for the model to learn non-occlusion samples that are inherently hard. On the contrary, CosFace is not affected by masked face samples, and simple fitting data can get good results.

Furthermore, we conduct the experiments where SOTA competitors are trained with a smaller margin for masked samples. Among them, the margin in CosFace is set to 0.4 for normal samples and 0.3 for masked samples. In CurricularFace, the margin is set to 0.5 for normal samples and 0.4 for masked samples. In terms of mean accuracy, the model using a smaller margin for the masked samples outperforms one using the same margin for the normal and masked samples. Therefore, using a smaller margin for the masked samples can eliminate mask bias to some extent but cannot realize the desired effect. Our proposed PLFace-Cur still achieves superior performance. Besides, the experiments on MS1MV3+MA-1.0 dataset with backbone ResNet100 show that our PLFace-Cur performs the best mean accuracy against SOTA methods. This demonstrates that our proposed progressive learning strategy adaptively adjusts the relative importance of masked and mask-free samples during different training stages, thus effectively alleviating the mask bias.

**Fig. 5.** The average positive cosine similarities of our PLFace and CurricularFace during training (Backbone: ResNet100, Dataset: MS1MV3+MA-1.0).

**Table 3** Comparisons on face verification (%) on RMRFD, ICCV2021-MFR-MASK and ICCV2021-MR-ALL. "MA-X" means masked face augmentation with a specific probability of X. "Mean" denotes the average of MFR-MASK and MR-ALL. Mean represents the margin for masked samples.

| Dataset | Backbone | Method       | Size / MB | RMRFD | MFR-MASK | MR-ALL | Mean |
|---------|----------|--------------|-----------|--------|-----------|--------|------|
| MS1MV3  | R50      | ArcFace      | 166       | 68.68  | 63.85     | 80.53  | 72.19|
| MS1MV3-MA-1.0 | R50 | ArcFace      | 166       | 68.40  | 79.27     | 60.91  | 74.14|
| MS1MV3  | R50      | CosFace      | 166       | 71.93  | 67.92     | 82.10  | 75.01|
| MS1MV3-MA-1.0 | R50 | CosFace      | 166       | 86.49  | 80.78     | 74.10  | 77.44|
| MS1MV3  | R50      | CurricularFace| 166       | 72.11  | 67.20     | 81.59  | 74.40|
| MS1MV3-MA-1.0 | R50 | CurricularFace| 166       | 84.84  | 80.24     | 71.18  | 75.71|
| MS1MV3  | R50      | CurricularFace| 166       | 85.23  | 80.42     | 73.87  | 77.15|
| MS1MV3-MA-1.0 | R50 | CurricularFace| 166       | 87.22  | 80.42     | 73.87  | 77.15|
| MS1MV3  | R50      | PLFace-Cur   | 166       | 87.27  | 78.57     | 74.10  | 76.34|
| MS1MV3-MA-1.0 | R50 | PLFace-Cur   | 166       | 87.33  | 80.00     | 77.51  | 78.76|
| MS1MV3  | R100     | ArcFace      | 248       | 87.92  | 83.90     | 72.78  | 78.34|
| MS1MV3-MA-1.0 | R100 | ArcFace      | 248       | 88.26  | 85.24     | 75.41  | 80.32|
| MS1MV3  | R100     | CosFace      | 248       | 88.02  | 84.21     | 77.67  | 80.94|
| MS1MV3-MA-1.0 | R100 | CosFace      | 248       | 88.12  | 84.88     | 75.02  | 79.95|
| MS1MV3  | R100     | CurricularFace| 248       | 88.88  | 83.90     | 78.08  | 80.99|
| MS1MV3-MA-1.0 | R100 | CurricularFace| 248       | 88.33  | 85.61     | 64.78  | 75.20|
| MS1MV3  | R100     | PLFace-Arc   | 248       | 88.88  | 84.49     | 72.06  | 78.28|
| MS1MV3-MA-1.0 | R100 | PLFace-Arc   | 248       | 88.35  | 85.70     | 73.64  | 79.67|
| MS1MV3  | R100     | PLFace-Cur   | 248       | 88.02  | 83.51     | 72.11  | 77.81|
| MS1MV3-MA-1.0 | R100 | PLFace-Cur   | 248       | 87.86  | 82.71     | 77.58  | 80.15|
| MS1MV3  | R100     | PLFace-Cur   | 248       | 88.24  | 84.41     | 80.99  | 82.70|
Table 4
Verification performance (%) on LFW, CFP-FP, AgeDB, RMFRD and IJB-C (Backbone: ResNet50, Dataset: MS1MV3+MA-1.0).

| Method      | Dataset            | LFW   | CFP-FP | AgeDB | RMFRD | IJB-C (FAR = 1e-5) | IJB-C (FAR = 1e-6) |
|-------------|--------------------|-------|--------|-------|-------|---------------------|---------------------|
| CosFace     | MS1MV3+MA-1.0      | 99.70 | 97.69  | 97.68 | 87.41 | 94.07              | 89.24              |
| ArcFace     | MS1MV3+MA-1.0      | 99.74 | 97.67  | 97.68 | 86.40 | 93.24              | 89.53              |
| CurricularFace | MS1MV3+MA-1.0    | 99.72 | 97.63  | 97.53 | 86.84 | 93.39              | 86.99              |
| LPD         | MS1MV3+MA-1.0      | 99.60 | 96.82  | 96.98 | 87.56 | 90.26              | 82.64              |
| FocusFace   | MS1MV3+MA-1.0      | 99.65 | 96.65  | 96.57 | 87.28 | 89.73              | 83.09              |
| EUM         | MS1MV3+MA-1.0      | 99.67 | 96.93  | 97.09 | 87.76 | 91.20              | 83.38              |
| MagFace     | MS1MV3+MA-1.0      | 99.76 | 97.74  | 97.75 | 86.60 | 93.74              | 89.90              |
| PLFace-Arc (Ours) | MS1MV3+MA-1.0    | 99.77 | 98.20  | 97.83 | 87.27 | 94.05              | 89.95              |
| PLFace-Cur (Ours) | MS1MV3+MA-1.0  | 99.73 | 98.23  | 98.27 | 87.33 | 94.30              | 89.74              |

Fig. 6. ROC of 1:1 verification protocol on IJB-C.

4.3.2. Results on LFW, CFP-FP, AgeDB and IJB-C
In this section, we evaluate the performance on several popular benchmarks, including LFW for unconstrained face verification, CFP-FP for large pose variations, AgeDB for age variations, RMFRD for mask variations and IJB-C for large-scale 1:1 verification. The IJB-C dataset contains about 3,500 identities with 31,334 images and 117,542 unconstrained video frames, which is further organized into 19,557 positive pairs and 15,638,932 negative pairs for 1:1 verification. As shown in Table 4, the experiments on these five datasets prove that our PLFace can offer more discriminative features than the baselines like CurricularFace and MagFace, which are also generic methods for face recognition. Fig. 6 shows the ROC curves of PLFace, ArcFace and CurricularFace on IJB-C with the backbone ResNet50, where our method achieves better performance.

4.3.3. Comparison with Non-uniform Strategy
Inspired by racial debiasing methods [33–35], where racial bias can be alleviated by assigning non-uniform attention to different racial identities, we design specific experiments to verify the idea of giving non-uniform attention to masked and mask-free samples. For the realization of non-uniform attention, we adopt two different ways: 1) giving different margins to masked and mask-free samples; 2) assigning different weights to the training loss of masked and mask-free samples. Taking CurricularFace as an example, on the one hand, we set the margin to 0.4 and 0.45 for mask samples, respectively (the margin is 0.5 for normal samples). On the other hand, we give the weight of 0.8 and 1.2 for the loss of mask samples (weight is 1 for normal samples), respectively. As shown in Table 5, we report the performance of four models trained on the MS1MV3+MA-1.0 dataset. Specifically, using a smaller margin for masked samples can only slightly eliminate mask bias, and the accuracy on normal samples still drops a lot. When the weight for the loss value of the masked samples is greater than 1, the accuracy on the masked samples is improved, but the accuracy on the normal samples is decreased. When the weight of the mask sample loss value is less than 1, the accuracy change is just the opposite. This proves that simply giving non-uniform attention to masked and mask-free samples cannot achieve the ideal mask debiasing effect.

4.3.4. Discussion on Mask Augmentation Ratio
In this section, we discuss the effect of mask augmentation ratio in training set on model performance. Specifically, we report the experiments on ArcFace and PLFace-Cur with different augmentation ratios. As tabulated in Table 6, our proposed PLFace and ArcFace achieve the best balance when the mask augmentation ratio is 0.4 and 0.2, respectively, implying that the most appropriate mask augmentation ratio is not exactly the same for different models. Therefore, setting the mask augmentation ratio to fixed is not beneficial for other models. For fairness, we conduct comprehensive experiments and individually picked out the most suitable mask augmentation ratio for each comparison method. Particularly, for margin-based methods, we also select the best margin for masked face samples. The evaluation results of SOTA methods using the best hyper-parameters are reported in Table 7. We can observe that comparison methods with the best hyper-parameters slightly improve the average accuracy. Because the mask augmentation ratio and margin both play a role in balancing the accuracy of the model for masked and normal samples, but the effect is relatively limited. PLFace-Cur still outperforms other normal face recognition methods and marked face recognition methods in terms of average accuracy.

4.3.5. Visualization
To demonstrate the de-mask-bias of the PLFace learned features, we conduct the visualization comparisons at the feature level. Specifically, we randomly extract the deep features of 1K masked and mask-free images with CurricularFace and PLFace-Cur, respectively. The features are visualized using t-SNE [40], as shown in Fig. 7. After adopting progressive learning, more masked and mask-free ones are mixed in feature space, thus there is no obvious mask bias between them.
Table 5
Comparisons with non-uniform strategy. “MA-X” means masked face augmentation with a specific probability of X. “Mean” denotes the average of MFR-MASK and MR-ALL. $M_{\text{mask}}$ represents the margin for masked samples, and $W_{\text{mask}}$ represents the weight for the loss value of the masked sample.

| Dataset | Backbone | Method       | Size / MB | RMFRD | MFR-MASK | MR-ALL | Mean   |
|---------|----------|--------------|-----------|-------|----------|--------|--------|
| MS1MV3+MA-1.0 | R50 | CurricularFace | 166 | 86.84 | 80.24 | 71.18 | 75.71 |
| MS1MV3+MA-1.0 | R50 | CurricularFace ($M_{\text{mask}} = 0.4$) | 166 | 87.29 | 80.42 | 73.87 | 77.15 |
| MS1MV3+MA-1.0 | R50 | CurricularFace ($M_{\text{mask}} = 0.45$) | 166 | 87.10 | 80.34 | 72.27 | 76.31 |
| MS1MV3+MA-1.0 | R50 | CurricularFace ($M_{\text{mask}} = 0.8$) | 166 | 87.47 | 79.87 | 71.50 | 75.69 |
| MS1MV3+MA-1.0 | R50 | CurricularFace ($M_{\text{mask}} = 1.2$) | 166 | 87.03 | 80.62 | 70.99 | 75.80 |
| MS1MV3+MA-1.0 | R50 | PLFace-Cur (Ours) | 166 | 87.33 | 80.00 | 77.51 | 78.76 |

Table 6
Comparisons of ArcFace and PLFace with different augmentation ratios. “MA-X” means masked face augmentation with a specific probability of X. “Mean” denotes the average of MFR-MASK and MR-ALL.

| Dataset | Backbone | Method       | Size / MB | RMFRD | MFR-MASK | MR-ALL | Mean   |
|---------|----------|--------------|-----------|-------|----------|--------|--------|
| MS1MV3+MA-0.2 | R50 | Arcface | 166 | 86.24 | 74.89 | 77.39 | 76.14 |
| MS1MV3+MA-0.4 | R50 | Arcface | 166 | 86.57 | 77.18 | 74.11 | 75.65 |
| MS1MV3+MA-0.6 | R50 | Arcface | 166 | 86.83 | 78.17 | 71.26 | 74.72 |
| MS1MV3+MA-0.8 | R50 | Arcface | 166 | 87.05 | 79.18 | 69.35 | 74.15 |
| MS1MV3+MA-1.0 | R50 | Arcface | 166 | 86.40 | 79.27 | 69.01 | 74.14 |
| MS1MV3+MA-0.2 | R50 | PLFace-Cur (Ours) | 166 | 86.74 | 76.02 | 81.47 | 78.75 |
| MS1MV3+MA-0.4 | R50 | PLFace-Cur (Ours) | 166 | 86.80 | 75.82 | 80.29 | 79.40 |
| MS1MV3+MA-0.6 | R50 | PLFace-Cur (Ours) | 166 | 87.19 | 75.36 | 78.34 | 78.85 |
| MS1MV3+MA-0.8 | R50 | PLFace-Cur (Ours) | 166 | 86.98 | 79.27 | 77.87 | 78.57 |
| MS1MV3+MA-1.0 | R50 | PLFace-Cur (Ours) | 166 | 87.33 | 80.00 | 77.51 | 78.76 |

Table 7
Comparisons on face verification (%) on RMFRD, ICCV2021-MR-FASK and ICCV2021-MR-ALL. “MA-X” means masked face augmentation with a specific probability of X. “Mean” denotes the average of MFR-MASK and MR-ALL. $M_{\text{mask}}$ represents the margin for masked samples. Particularly, these methods with the mask augmentation ratio and margin in this table lead to the best average accuracy for them.

| Dataset | Backbone | Method       | Size / MB | RMFRD | MFR-MASK | MR-ALL | Mean   |
|---------|----------|--------------|-----------|-------|----------|--------|--------|
| MS1MV3+MA-0.2 | R50 | Arcface ($M_{\text{mask}} = 0.4$) | 166 | 86.42 | 75.91 | 76.53 | 76.22 |
| MS1MV3+MA-0.4 | R50 | CosFace ($M_{\text{mask}} = 0.3$) | 166 | 87.13 | 80.35 | 76.30 | 78.33 |
| MS1MV3+MA-0.6 | R50 | CurricularFace ($M_{\text{mask}} = 0.4$) | 166 | 86.95 | 79.82 | 76.50 | 78.16 |
| MS1MV3+MA-0.4 | R50 | LPD | 166 | 86.45 | 81.79 | 66.88 | 74.34 |
| MS1MV3+MA-0.4 | R50 | FocusFace | 166 | 86.98 | 80.30 | 71.06 | 75.68 |
| MS1MV3+MA-0.4 | R50 | EU | 166 | 87.34 | 81.65 | 72.92 | 77.29 |
| MS1MV3+MA-0.4 | R50 | MagFace | 166 | 86.43 | 79.06 | 70.51 | 74.79 |
| MS1MV3+MA-0.4 | R50 | PLFace-Cur (Ours) | 166 | 86.80 | 78.52 | 80.29 | 79.40 |

5. Conclusion

This paper proposes a novel Progressive Learning Loss (PLFace) that achieves a progressive training strategy for deep face recognition to learn balanced performance for masked/mask-free faces based on margin losses. Our key idea is to address mask-free samples in the early training stage and masked ones in the later stage. Moreover, PLFace can be easily migrated to the existing face recognition loss functions, only adding negligible computational complexity during training. Extensive experimental results on popular regular and masked face benchmarks demonstrate that our proposed PLFace can effectively eliminate mask bias in face recognition. Compared to state-of-the-art competitors, PLFace significantly improves the accuracy of MFR while maintaining the performance of normal face recognition.
Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

Acknowledgement

This work was supported by National Natural Science Foundation of China (U1903214, 62071339, 62072347, 62072350, 61876135, 62171328), Natural Science Foundation of Hubei Province (2021CFB464) and Guangdong-Macau Joint Laboratory for Advanced and Intelligent Computing (2020B1212030003). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] J. Wan, Y. Chen, B. Bai, Joint feature extraction and classification in a unified framework for cost-sensitive face recognition, Pattern Recognition 115 (2021) 107618.
[2] R. He, Y. Li, X. Wu, L. Song, Z. Chai, X. Wei, Coupled adversarial learning for semi-supervised heterogeneous face recognition, Pattern Recognition 110 (2021) 107618.
[3] M. He, J. Zhang, S. Shan, M. Kan, X. Chen, Deformable face net for pose invariant face recognition, Pattern Recognition 100 (2020) 107133.
[4] X. Luo, Y. Xu, J. Yang, Multi-resolution dictionary learning for face recognition, Pattern Recognition 93 (2019) 283–292.
[5] B. Huang, Z. Wang, K. Jiang, Q. Zou, X. Tian, T. Lu, Z. Han, Joint segmentation and identification feature learning for occlusion face recognition, IEEE Transactions on Neural Networks and Learning Systems (2022), Doi:10.1109/TNNLS.2022.3176104.
[6] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[7] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[8] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, Cosface: Large margin cosine loss for deep face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5265–5274.
[9] J. Deng, J. Guo, N. Xue, S. Zaferiou, Arcface: Additive angular margin loss for deep face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.
[10] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, F. Huang, CurricularFace: adaptive curriculum learning loss for deep face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 5901–5910.
[11] D. Yi, Z. Lei, S. Liao, S.Z. Li, Learning face representation from scratch, arXiv preprint arXiv:1411.7923 (2014).
[12] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, et al., Partial fc: Training 10 million identities on a single machine, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1445–1449.
[13] J. Deng, J. Guo, X. An, Z. Zhu, S. Zaferiou, Masked face recognition challenge: The insightface track report, in: IEEE International Conference on Computer Vision Workshops, 2021, pp. 1437–1444.
[14] J. Deng, J. Guo, Mask data augmentation, 2021, (https://github.com/deepinsight/insightface/tree/master/recognition/tools/).
[15] F. Ding, P. Peng, Y. Huang, M. Geng, Y. Tian, Masked face recognition with latent part detection, in: Proceedings of the 28th ACM international conference on Multimedia, 2020, pp. 2281–2289.
[16] M. Geng, P. Peng, Y. Huang, Y. Tian, Masked face recognition with generative data augmentation and domain constrained ranking, in: Proceedings of the 28th ACM international conference on Multimedia, 2020, pp. 2246–2254.
[17] B. Huang, Z. Wang, G. Wang, K. Jiang, Z. He, H. Zou, Q. Zou, Masked face recognition datasets and validation, in: IEEE International Conference on Computer Vision Workshops, 2021, pp. 1487–1491.
[18] Y. Li, K. Guo, Y. Lu, L. Liu, Cropping and attention based approach for masked face recognition, Applied Intelligence 51 (5) (2021) 3012–3025.
[19] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.
[20] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Spherface: Deep hypersphere embedding for face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 212–220.

Baojin Huang received the B.S degree in School of Computer Science from Wuhan University of Technology, Wuhan, China in 2019. He is currently pursuing the Ph.D. degree with National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, Wuhan, China. His research interests include person search and face recognition.

Zhongyuan Wang received the Ph.D. degree in communication and information system from Wuhan University, Wuhan, China in 2008. Dr. Wang is currently a professor with the School of Computer Science, Wuhan University, Wuhan, China. He has been directing four projects funded by the National Natural Science Foundation of China. He has authored or co-authored over 80 refereed journal and conference papers and has been granted more than 30 invention patents. His research interests include biometrics and identification, computer vision and pattern recognition, image processing, etc.
GuangCheng Wang received the B.S. degree in School of Electrical and Optoelectronic Engineering from Changzhou Institute of Technology, Changzhou, China, in 2015, and M.S. degree in School of Information and Control Engineering from China University of Mining and Technology, Xuzhou, China, in 2018. He is currently pursuing the Ph.D. degree with the National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan, China. His research interests include image processing, quality assessment, face recognition and biometrics and identification.

Kui Jiang received the B.S. degree from College of Chemistry and Chemical Engineering, Xinjiang University, Urumchi, China, in 2017. He is currently pursuing the Ph.D. degree under the supervision of Prof. Zhongyuan Wang in the School of Computer, Wuhan University, Wuhan, China. His research interests include image/video processing and computer vision.

Zhen Han received the B.S. degree in computer science and technology and Ph.D. degree in computer application technology from Wuhan University, Wuhan, China, in 2002 and in 2009 respectively. Now he is an associate professor in school of computer, Wuhan University. His research interests include image/video compressing and processing, computer vision and artificial intelligence.

Tao Lu received the B.S and M.S. degrees from School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China in 2003 and in 2008 respectively, and the Ph.D. degree from National Engineering Research Center For Multimedia Software, Wuhan University, in 2013. Now he is an associate professor in School of Computer Science and Engineering, Wuhan Institute of Technology and a research member in Hubei Provincial Key Laboratory of Intelligent Robot. He was a Post-Doc in Department of Electrical and Computer Engineering, Texas A&M University from 2015 to 2017. His research interests include image/video processing, computer vision and artificial intelligence.

Chao Liang received the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2012. He is currently working as an Associate Professor with the National Engineering Research Center for Multimedia Software, Computer School, Wuhan University, Wuhan, China. His research interests focus on multimedia content analysis and retrieval, and computer vision and pattern recognition, where he has published over 60 papers, including premier conferences, such as CVPR, ACM MM, AAAI, IJCAI, and honorable journals, such as IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. Dr. Liang won the Best Paper Award of PCM 2014.