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.
Dual-Proxy Modeling for Masked Face Recognition

Wang Shuhui*a,*, Mao Xiaochena

*aShanghai Institute of Measurement and Testing Technology, 1500 Zhang Heng Road, Shanghai, 201203, P.R.China

Abstract

With the recent worldwide COVID-19 pandemic, almost everyone wears a mask daily, leading to severe degradation in the accuracy of conventional face recognition systems. Several works improve the performance of masked faces by adopting synthetic masked face images for training. However, such methods often cause performance degradation on unmasked faces, raising the contradiction between the face recognition system’s accuracy on unmasked and masked faces. In this paper, we propose a dual-proxy face recognition training method to improve masked faces’ performance while maintaining unmasked faces’ performance. Specifically, we design two fully-connected layers as the unmasked and masked feature space proxies to alleviate the significant difference between the two data distributions. The cross-space constraints are adopted to ensure the intra-class compactness and inter-class discrepancy. Extensive experiments on popular unmasked face benchmarks and masked face benchmarks, including real-world mask faces and the generated mask faces, demonstrate our method’s superiority over the state-of-the-art methods on masked faces without incurring a notable accuracy degradation on unmasked faces.

© 2022 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
“Peer-review under responsibility of the scientific committee of the 7th International Conference on Intelligent, Interactive Systems and Applications”

Keywords: Masked face recognition; Dual-Proxy; neural network; deep learning

1. Introduction

The COVID-19 pandemic has become a global health crisis and a great challenge. To limit the risks of COVID-19, wearing medical face masks in public is an effective and efficient method. However, masked faces are heavily occluded, which causes verification difficulties for current face recognition applications, such as surveillance systems, face access control, and identifying passengers at airports. A few methods are proposed to improve the performance of mask face recognition. Anwar et al. [1] combined normal and masked samples into one batch to improve models’ masked face recognition performance. To address the lack of large-scale masked face datasets, MaskTheFace [1] is a

* Corresponding author.
E-mail address: wangsh@simt.com.cn; maoxc@simt.com.cn

1877-0509 © 2022 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
“Peer-review under responsibility of the scientific committee of the 7th International Conference on Intelligent, Interactive Systems and Applications”
10.1016/j.procs.2022.10.022
tool used to generate masked face datasets effectively by augmenting the normal face datasets with masks. However, it is inevitable for such models to witness an accuracy drop in normal face recognition. In this paper, we propose a novel method to address the masked face recognition problem, which can improve models’ masked face recognition performances and maintain normal face recognition abilities. Traditional face recognition methods verify a face image’s identity by finding the smallest distance between the face feature vector and proxies of individuals. Generally, each identity is aligned with one proxy, which can be considered as the class center. Masked faces are heavily occluded. Compared with the same identity’s normal face feature spaces, the loss of face key components leads to offset masked face features. This problem motivates us to consider dual-proxy operations, assigning each identity with a normal feature center and a masked feature center. Specifically, we introduce two fully-connected layers to represent the proxies of normal face centers and masked face centers. The two layers share the same backbone of convolution neural networks, which is shown in Figure 1. The two fully-connected layers facilitate cross-branch information back-propagation, rewarding normal and masked space losses to the backbone. Our method keeps face recognition models’ precision on various popular face recognition benchmark datasets and gains verification accuracy on masked face datasets generated from LFW [2] and IJB-C [3].

The remainder of this paper is organized as follows. Section 2 briefly reviews related work. Section 3 illustrates our proposed dual-proxy algorithm and analyzes its benefits. Experiments are shown in Section 4, and conclusions are drawn in Section 5.

2. Related work

Face recognition. Face recognition is a typical application in computer vision [4]. In recent years, the progress in deep learning and convolutional neural networks has boosted the performance of face recognition substantially [5][6][7][8]. A key component of CNN’s success is the support of large-scale labeled face datasets, such as MS1M [9] and VGGFace2 [10], which contain real-world face variations, including pose, expression, and illumination. With powerful computational resources like modern GPUs, CNN models can be trained to extract robust and invariant features from images taken under unconstrained environments [11]. Besides, margin-based softmax loss enforces greater intra-class compactness and inter-class discrepancy, resulting in more discriminating features. Several works focused on incorporating margin penalty into CNN framework, such as Arc-Face [12], CosFace [13].

A new research hot spot, masked face recognition, was created after the outbreak of the COVID-19 pandemic, as people have been used to wearing masks in their daily lives. The face blocking of masks raised a serious question on the accuracy of the facial recognition system used for different occasions, such as tracking office attendance, confirming identities in banks, and unlocking phones [14]. To solve the lack of masked face datasets, Anwar et al. proposed MaskTheFace [1] to generate a masked face dataset from a normal face dataset. Moreover, to identify people from occluded face images, traditional metric learning methods such as triplet loss and center loss are applied to close the distance between masked face images and non-occluded face images from the same person [15][16].

Cross-Domain Face recognition. Cross-domain face recognition aims to recognize the faces collected from multiple domain variations, such as visible light, low-resolution, and 3D images. Due to the apparent discrepancy between different data sources, domain adaptation and metric learning approaches are introduced to minimize the domain mismatch by aligning the feature distribution between source and target data [17]. The simplest form of parameter transfer learning refers to weight sharing for deep neural networks, where low-level parameters are assumed to be generalizable to both domains. In contrast, high-level features are sensitive to specific applications and are finetuned to the desired task. Other methods use sub-classes for each identity, which can be directly adopted by ArcFace and significantly increase its robustness. As a pioneering work, Qian et al. applied soft triplet loss to transfer the feature distribution, and Deng et al. proposed Sub-center ArcFace [18], which can differentiate face identities under noise.

3. Method

There are apparent feature distribution discrepancies between normal face images and masked face images, as the masked face images lose vital components in the lower half face. So it is hard to map normal facial and masked face features into a united feature space. We propose a face recognition training method based on dual-proxy modeling to solve this problem. Two fully-connected layers are used to represent the proxies of normal and masked class centers,
respectively. Our method effectively reduces the model convergence difficulty of training on a dataset that contains multiple face images of the same identity with and without masks. We verify accuracy gain over a wide range of masked face recognition tasks.

3.1. Framework

As shown in Figure 1, the framework mainly consists of three parts: data input module, feature extraction module, and dual-proxy module. The role of the data input module is to ensure the balance between masked face samples and normal face samples in the same batch during training. The feature extraction module uses a convolution neural network to convert input face images into feature embeddings. The dual-proxy module models two feature spaces, normal faces and masked faces. The loss function adopts cross-space constraints to ensure that the feature distance between masked face and normal face from the same ID is closer than the distance between different ID samples.

Let $\mathbf{x}$ be a set of face images at the input of a deep neural network. The masked samples from the same ID are collected as another batch size of the input, which is marked as $\mathbf{x}_m$. The two inputs share the same CNN network structure $\mathbf{F}$. After feeding two inputs into the deep neural network backbone, two intermediate data embeddings are obtained. We denote them as $\mathbf{F}(\mathbf{x}; \theta)$ and $\mathbf{F}(\mathbf{x}_m; \theta)$, where $\theta$ stands for the deep neural network’s parameters. Under normal conditions without ambiguity, we can write $\mathbf{F}(\mathbf{x})$ and $\mathbf{F}(\mathbf{x}_m)$ in short. $\mathbf{F}(\mathbf{x})$ and $\mathbf{F}(\mathbf{x}_m)$ are of the same dimension. Current SOTA deep face recognition methods mostly adopt margin-based softmax loss functions [12][19][20]. Taking ArcFace as an example, the loss function can be formulated as follows:

$$
\mathcal{L} = -\log \frac{e^{s(\cos(\theta_{yi} + m))}}{e^{s(\cos(\theta_{yi} + m))} + \sum_{j \neq y_i} e^{s(\cos(\theta_{yi}))}}
$$

(1)

where $\cos(\theta_{yi})$ and $\cos(\theta_{j})$ are the positive and negative cosine similarities. The parameters $s$ and $m$ denote the re-scaled feature norm and the feature margin between different classes, respectively.

In practice, we find it is hard to train a model with such loss functions when the training data contain normal faces and masked faces in the meanwhile. This phenomenon leads us to insert two fully-connected layers and corresponding loss functions to our framework, which converts feature maps output from the CNN backbone to different feature spaces, normal face space and masked face space. The loss functions are marked as $\mathbf{L}$ and $\mathbf{L}_m$. Now we can associate embeddings and loss functions in the same space together. Then we have $\mathbf{L}(\mathbf{F}(\mathbf{x}))$ and $\mathbf{L}_m(\mathbf{F}(\mathbf{x}_m))$, which measure the backbone’s feature extraction ability of two occasions. Moreover, the cross-space constraints are adopted to ensure the intra-class compactness and inter-class discrepancy. As a result, we add two more losses $\mathbf{L}(\mathbf{F}(\mathbf{x}_m))$ and $\mathbf{L}_m(\mathbf{F}(\mathbf{x}))$ to estimate cross-space face recognition process and control the distance between two spaces. The loss function of CNN backbone can be written as a combination of four loss items:
\[ L^{total} = L(F(x)) + L_m(F(x_m)) + L(F(x_m)) + L_m(F(x)) \]  

(2)

3.2. Dual-proxy module

The data flow in loss function module includes two parts: forward calculation of loss function value and gradient back-propagation. As shown in Figure 2, the arrows with solid line and dotted line represent the forward loss calculation and backward gradient update separately. The cross marks in the figure indicate that the gradients of the two cross-space losses \( L(F(x_m)) \) and \( L_m(F(x)) \) are cut off. The purpose of this conduct is to prevent the two fully-connected layers from getting closer. In the worst case, corresponding distributions will overlap to some extent during the training stage, thereby making the dual-proxy strategy lose efficacy. Using dual proxies to model normal samples and masked samples with the same ID, the model will not rigidly compress the two data distributions to the same class center, thereby improving the generalization ability of the model. In testing process, we use the output feature maps of backbone as image features. To identity each face, we compare the feature similarities between query images and gallery identities.

In summary, we establish a CNN model with two fully-connected layers imitating normal and masked face feature spaces, allowing the network backbone to consider cross-space information in both forward-propagation and back propagation. In contrast, fully-connected layers’ gradients are only affected by the distance between face feature vectors and correlative identity centers in either normal space or masked space. Our dual-proxy strategy makes full use of neural responses of normal and masked faces and improves the face recognition robustness of model backbone.

3.3. Masked face augmentation

After the sudden emergence of the COVID-19 epidemic, up till now, there are only a few institutions collected and published real-world masked face datasets [14, 1]. It is hard for researchers to train a masked face recognition deep neural network without enough data. To improve the model’s recognition accuracy of diverse mask-wearing faces in practice, we use the data augmentation method to extend masked faces’ diversity. Our method is based on MaskTheFace Tool [1]. MaskTheFace adds mask patches to uncovered faces, simulating the scene when individuals wear masks. As there are masked faces and normal faces in the same batch, their ratio has a direct impact on loss calculation: The more masked faces in training batches, the higher the trained model’s mask recognition precision ends to be. However, it is hard for the model to learn a balanced feature proxy between the normal face center and masked face center for each identity. In this work, we experimentally found that setting the ratio to 0.5 can achieve a good balance between the masked and normal face recognition performance.

4. Experiments

4.1. Dataset

Masked Face Dataset. We test our method’s performance on generated masked face dataset. We use MaskTheFace [1] to convert any existing normal face dataset to masked-face dataset LFW-mask and IJBC-mask.

Normal Face Dataset. We employ refined MS1M as our training data. We test our method’s performance on original and masked face images generated from several normal face datasets. IJB-C [3] is a common large-scale face benchmark that contains about 31,334 images of 3,531 identities, with an average of 6 images per identity varying in pose, age, ethnicity and illumination. We also test our method on several popular benchmarks, including LFW [2], CFP-FP [21], CPLFW [22], AgeDB [23], and CALFW [24].

4.2. Training settings

We build our dual-proxy face recognition model as Section 3.1 explains. Our deep model takes \(112 \times 112\) aligned faces following previous work [12]. We adopt ResNet50 and ResNet100 as the backbone [25, 26], applying Arcface
[12] as loss function. In our proposal, normal and masked face spaces share the same backbone, but have separate fully-connected layers. The gradient backpropagation process is conducted as Section 3.

Our framework is implemented in Pytorch [27]. We train models on 4 NVIDIA Tesla V100 GPU with batch size 512. The models are trained with the SGD algorithm, with momentum 0.9 and weight decay $5e^{-4}$. The learning rate starts from 0.1 and is divided by 10 at 10, 18, 24 epochs. The training process is finished at 26 epochs. We follow the common setting as Arcface to set scale $s = 64$ and margin $m = 0.5$. The margin parameters are set as $\alpha_1 = \alpha_2 = 0.4$.

Table 1. Verification comparison on benchmark face recognition datasets. ResNet50mix is trained with a mixed input of normal and masked images in MS1M.

| Methods (%) | LFW | CFP-FP | AgeDB | CPLFW | CALFW | LFW(N-M) | LFW(M-M) |
|-------------|-----|--------|-------|-------|-------|----------|----------|
| ResNet50    | 99.78 | 97.14  | 97.67 | 91.90 | 95.92 | 99.15    | 98.95    |
| ResNet50mix | 99.72 | 97.20  | 97.63 | 92.25 | 95.95 | 99.58    | 99.63    |
| ResNet50DualProxy | 99.75 | 97.03  | 97.77 | 92.35 | 96.00 | 99.68    | 99.58    |
| ResNet101   | 99.79 | 97.76  | 97.87 | 92.98 | 95.93 | 99.34    | 99.21    |
| ResNet101mix | 99.76 | 97.84  | 97.93 | 92.87 | 95.98 | 99.62    | 99.67    |
| ResNet101DualProxy | 99.73 | 97.93  | 98.03 | 92.98 | 95.98 | 99.75    | 99.70    |

Table 2. Verification TPR(@FAR=$1e^{-5}$ and $1e^{-6}$) comparison on normal IJB-C dataset and generated masked version with ResNet50 models.

| Datasets (%) | baseline | mix | Dual-Proxy |
|--------------|----------|-----|------------|
| Normal(FAR=$1e^{-5}$) | 92.59    | 90.92 | 91.74      |
| Normal(FAR=$1e^{-6}$) | 84.81    | 80.38 | 84.45      |
| Masked(FAR=$1e^{-5}$) | 62.68    | 88.23 | 89.43      |
| Masked(FAR=$1e^{-6}$) | 48.81    | 77.30 | 80.70      |

4.3. Experimental Results on benchmarks

The verification results in benchmark datasets are summarized in Table 1. We compare the experiment results between using Dual-Proxy Backward and both two components on IR50 and IR101. For fair comparison, besides
IR50, we also train IR50mix with a mixed input of normal and synthetic masked images as baseline. Training with batches mixed with unmasked and masked faces, we can observe mask recognition accuracy increased markedly and normal face recognition performance drop down in the meanwhile. Dual-Proxy structure further improves masked face recognition accuracy by aligning each identity with two sub-centers. In addition, we observe more significant accuracy gain when the network goes deeper. On the verification results of IJB-C dataset in Table 2, we observe very similar recognition accuracy drop phenomena for ResNet50mix. Our method helps the network to be trained better, which improves generated masked face recognition TPR@FAR = 0.0001 by 14.36% and keeps the ability of normal face recognition stable.

4.4. Ablation Study

As we mentioned in Section 3.1, the margin of ArcFace Loss has “butterfly effect” on the model training. The margin $m$ in Equation 2 has a geometric attribute as it can control the distance between face features and class center of individuals. When we mix masked faces and normal faces together, it’s essential to choose a suitable margin which can keep feature vectors close to corresponding space’s identity centers. We test different margin values on model training process and summarize the experimental results in Table 3. The results show that the value of margin has a distinct impact on model performance. Small margin leads to the overlapping of diverse identities. Big margin makes the same identity’s masked face features not similar enough to normal face features. Suitable margin helps the model to establish flexible embedding, which makes the masked face feature distribution different from but close enough to normal face distribution. Comparing verification results between 0.35, 0.4 and 0.5, we find setting $m = 0.4$ is the best choice.

5. Conclusions

In this paper, we propose a dual-proxy masked face recognition method, a straightforward yet practical approach to improve the ability of deep neural networks for occluded face recognition. We adopt two fully-connected layers after the model backbone to process neural responses from the normal and masked faces, which leads to multiple spaces’ information back-propagation throughout the backbone and modeling separative feature centers for each identity. Our method can be applied to many deep convolution neural networks and produce persistent masked faces accuracy gain on typical face recognition datasets.

Table 3. Ablation study of margin values. The margin values of Dual − Proxy, Dual − Proxy⋆ and Dual − Proxy⋄ are set as 0.40, 0.35 and 0.50 separately.

| Methods (%)    | LFW | CFP-FP | AgeDB | CPLFW | CALFW |
|----------------|-----|--------|-------|-------|-------|
| ResNet50       | 99.78 | 97.14 | 97.67 | 91.90 | 95.92 |
| ResNet50mix    | 99.72 | 97.20 | 97.63 | 92.25 | 95.95 |
| ResNet50Dual−Proxy | 99.75 | 97.03 | 97.77 | 92.35 | 96.00 |
| ResNet50Dual−Proxy⋆ | 99.77 | 97.10 | 97.45 | 92.27 | 96.13 |
| ResNet50Dual−Proxy⋄ | 99.80 | 97.59 | 97.53 | 91.87 | 95.90 |

References

[1] A. Anwar, A. Raychowdhury, Masked face recognition for secure authentication, arXiv preprint arXiv:2008.11104.
[2] G. B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, Tech. Rep. 07-49, University of Massachusetts, Amherst (October2007).
[3] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Nigge1, J. Anderson, J. Cheney, et al., Iarpa janus benchmark-c: Face dataset and protocol, in: 2018 International Conference on Biometrics (ICB), IEEE, 2018, pp. 158–165.
[4] W. Zhao, R. Chellappa, P. J. Phillips, A. Rosenfeld, Face recognition: A literature survey, ACM Comput. Surv. (2003) 399–458.
[5] W. Deng, J. Hu, J. Guo, Extended src: Undersampled face recognition via intraclass variant dictionary, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (9) (2012) 1864–1870.
[6] W. Deng, J. Hu, J. Guo, Face recognition via collaborative representation: Its discriminant nature and superposed representation, IEEE transactions on pattern analysis and machine intelligence 40 (10) (2017) 2513–2521.

[7] D. Chen, X. Cao, F. Wen, J. Sun, Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 3025–3032.

[8] Z. Cao, Q. Yin, X. Tang, J. Sun, Face recognition with learning-based descriptor, in: 2010 IEEE Computer society conference on computer vision and pattern recognition, IEEE, 2010, pp. 2707–2714.

[9] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, Ms-celeb-1m: A dataset and benchmark for large-scale face recognition, in: European conference on computer vision, Springer, 2016, pp. 87–102.

[10] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age, in: 2018 13th IEEE internationalconference on automatic face & gesture recognition (FG 2018), IEEE, 2018, pp. 67–74.

[11] Z. Zhu, P. Luo, X. Wang, X. Tang, Deep learning identity-preserving face space, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 113–120.

[12] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.

[13] 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: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5265–5274.

[14] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, et al., Masked face recognition dataset and application, arXiv preprint arXiv:2003.09093.

[15] C. Han, S. Shan, M. Kan, S.Wu, X. Chen, Face recognition with contrastive convolution, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 118–134.

[16] M. Kan, S. Shan, X. Chen, Multi-view deep network for cross-view classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4847–4855.

[17] Q. Qian, L. Shang, B. Sun, J. Hu, H. Li, R. Jin, Softtriple loss: Deep metric learning without triplet sampling, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6450–6458.

[18] J. Deng, J. Guo, T. Liu, M. Gong, S. Zafeiriou, Sub-center arcface: Boosting face recognition by large-scale noisy web faces, in: European Conference on Computer Vision, Springer, 2020, pp. 741–757.

[19] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: Deep hypersphere embedding for face recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 212–220.

[20] 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: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5901–5910.

[21] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, D. W. Jacobs, Frontal to profile face verification in the wild, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2016, pp. 1–9.

[22] T. Zheng, W. Deng, J. Hu, Cross-pose lfw: A database for studying crosspose face recognition in unconstrained environments, Tech. Rep. 18-01, Beijing University of Posts and Telecommunications (February 2018).

[23] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, S. Zafeiriou, Agedb: the first manually collected, in-the-wild age database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 51–59.

[24] T. Zheng, W. Deng, J. Hu, Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments, arXiv:1708.08197.

[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[26] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.

[27] A. Paszke, G. et al., Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.