Mask Aware Network for Masked Face Recognition in the Wild

Kai Wang*[1], Shuo Wang*[2], Jianfei Yang[3], Xiaobo Wang[4], Baigui Sun[2], Hao Li[2], and Yang You †[1]

[1]National University of Singapore
[2]Alibaba Group, China
[3]Nanyang Technological University, Singapore
[4]Institute of Automation, Chinese Academy of Sciences

Abstract

Face recognition is one of the most important research topics for intelligence security system, especially in the COVID-19 era. Medical research has proven that wearing a mask is the most efficient way to avoid the risk of COVID-19. Nevertheless, classic face recognition systems often fail when dealing with masked faces. So it is essential to design a method that is robust to Masked Face Recognition (MFR). In this paper, to relieve the degraded performance of MFR, we propose Mask Aware Network (MAN) including a mask generation module and a loss function searching module. The mask generation module utilizes the face landmarks to obtain more realistic and reliable masked faces for training. The loss function searching module tries to match the most suitable loss for face recognition. On ICCV MFR challenge, our team victor-2021 achieves 5 first places (including 3 champions in standard face recognition and 2 champions in masked face recognition) and 1 third place by 3rd August 2021. These results demonstrate the robustness and generalization of our method in both standard or masked face recognition task.

1. Introduction

Face recognition (FR) has been a long-standing topic in the computer vision community since last century [4, 12, 35, 23, 20, 37, 27]. In past decades, many well-annotated datasets [38, 13] and advanced deep learning algorithms [12, 20] have been proposed, which make significant progress for face recognition. Existing face recognition models have strong robustness and generalization, but they still easily misidentify occluded, large-posture and overexposed faces.

Figure 1. The top row is the official mask augmentation, the bottom row is our mask augmentation. It is obviously that our augmentation can produce more variable masked faces than official method.

In the COVID-19 era, people always wear a mask to protect themselves from infection, which is very difficult for existing models to identify accurately. Masked faces may lead following problems: 1. The texture feature of the masked face area is not visible. 2. The masked areas for different or even the same people are not consistent at all times. Previous works [1, 18, 26, 16, 33, 3, 34] can be summarised as three categories: matching-based methods, restoration-based methods, and occlusion removal-based methods. Matching-based methods [16, 33, 9] try to compare the similarity between images using a matching process, but the performance is too sensitive to different sampling strategies. Restoration-based methods [3, 34, 8] aim to restore the masked regions of probe images according to gallery ones, it decreases the difficulties of recognition largely. In order to reduce the influence of occluded regions, occlusion removal-based methods[1, 18] first detect the occluded regions and then remove it directly. However, removing the occluded regions may break the shape feature of faces.

*Equally-contributed first authors
†Corresponding author (youy@comp.nus.edu.sg)

The previous works may focus on reducing the influence of masks, but ignore how to make sufficient augmentation using masks. To this end, we propose Mask Aware Network, consisting of Mask generation and loss function searching modules. Specifically, given an image, we first detect facial landmarks and then add mask under a given face with a random ratio (half and full masked). The detected facial landmarks can be regarded as the anchor points for our mask augmentation. As shown in Figure 1, different from other mask augmentation methods, our augmentation is more flexible and variable. Then, we feed these images into Convolutional Neural Network to extract the face feature. To enhance the feature discrimination, we utilize the LFS [29] to search the most suitable loss function for this task. Finally, our solution achieves 5 first places and 1 third place in ICCV Masked Face Recognition WebFace[39] track.

To sum up, the main contributions of this paper can be summarized as follows:

1. We propose the Mask Aware Network for masked face recognition, which consists of two well-designed module named Mask Generation and Loss Function Searching Modules.

2. Mask Generation module utilizes the facial landmarks as anchor points to improve the variety of masked faces, which is the crucial factor for our model performance.

3. Loss Function Searching module ensures that we can obtain the most suitable loss function for the masked face recognition problem.

2. Preliminary Knowledge

Deep Face Recognition. Face recognition has witnessed dramatically progresses due to the large-scale datasets, advanced architectures and loss functions. Large-scale datasets like CASIA-WebFace [38], VGGFace2 [5], *etc.* play the most crucial role in the data-driven deep learning methods. Based on these datasets, a variety of CNN architectures for improving the performances, such as VGGNet [19], GoogleNet [22], ResNet [11], AttentionNet [25] and MobileFaceNet [6] have been proposed. For the loss function, contrastive loss [21, 36] and triplet loss [19] may be good candidates. But they suffer from high computational cost and slow convergence. To this end, researchers attempt to explore new metric learning loss functions to boost the face recognition performance. Several margin-based softmax losses [15, 24, 30, 31, 7] have been exploited and obtain the state-of-the-art results.

Masked Face Recognition. Masked Face Recognition (MFR) has obtained more and more attention from research community, especially under the global-wise COVID-19 situation. As aforementioned, there are main three types MFR methods: Matching-based, Restoration-based and Occlusion Removal-based methods. Matching-based methods try to use the sampled face patches for similarity measure. Martinez et al.[16] sampled the face region into a fixed number of local patches. To make the matching progress more efficient, Duan et al. [9] propose facial landmarks-based matching strategy to ensure the patches for matching are aligned with each other. Matching-based method can not work when the selected patches are from masked area. So how to restore the masked area is also a good solution for MFR. Drira et al. [8] applied a statistical shape model to predict and restore the partial facial curves. Iterative closest point (ICP) algorithm has been used to remove occluded regions in [10]. In order to avoid a bad reconstruction process, another intuitive idea is to remove occluded region directly. This method needs to detect the masked area first and then removes it by the detected box. Different from these methods, our Mask Aware Network utilizes the mask and detected facial landmarks to do sufficient data augmentation for MFR.

3. Methodology

In this section, we first overview the MAN, and then present its two modules. We finally present the details of training and inference.

3.1. Overview of MAN

Masked Face Recognition challenge aims to build face recognition system that can perform well on both standard and masked face recognition tasks. To deal with these tasks, we design Masked Aware Network (MAN) and show it in Figure 2. Give a face dataset, we first define training and validation sets. For each image in training set, online mask generation module utilizes the detected facial landmarks to generate a masked face randomly, which can provide more variable masked faces. We then feed the masked (or non-masked) face into several backbones to extract the features. Note that all the backbones are with same structure but different parameters. To explore the most suitable loss for the current task, we define a loss function pool where different kind of loss functions are initialized with the same α. The α_i represents the probability of using i-th loss function in the corresponding backbone, which is determined by the performance of the corresponding backbone on the validation set. Finally, we synchronize the backbone parameters of the best performance on validation set to other backbones.

3.2. Mask Generation Module

In order to make variable augmentation of faces, different from other mask augmentation methods, we utilize the facial landmarks to generate masked face. The process can be formulated as follows,

$$M = G(I, L; \theta), \tag{1}$$

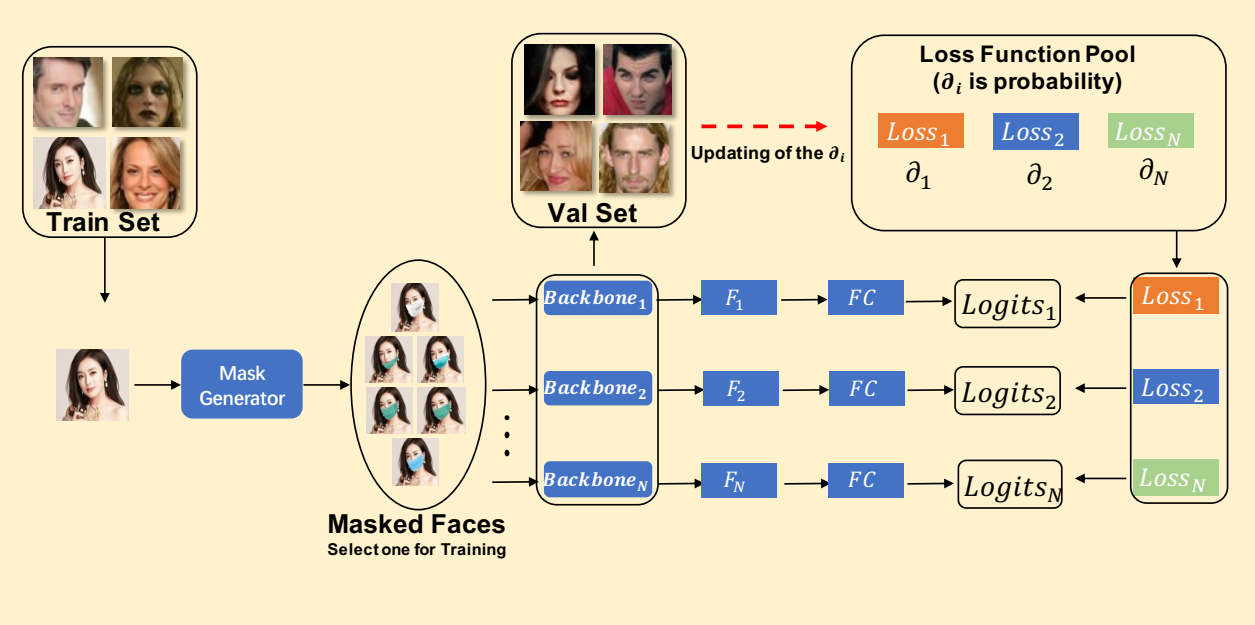

Figure 2. The pipeline of our proposed mask aware network. Given an image from train set, the online mask generator first add mask with 20% probability. Then, the image (or masked image) is fed into several backbones to extract the face recognition features. Note that, the backbones are with the same structure and supervised by different loss functions. The loss function pool contains several loss functions. The using probability of each function is initialized with the same α . The updating of α is determined by the results on the validation set. If the loss function achieves more better results on validation set, the using probability is updated with a larger value.

where G represents the operation of add mask. I is the input face image. L denotes the facial landmarks, θ is the parameters of the operation. We use face landmark detector to detect 96 facial landmarks and then choose 10 facial landmarks of the side face (5 points of left side and 5 points of right side). Based on these 10 facial landmarks, we can adjust the masked area randomly. For example, it is very normal that people take oblique masks. As shown in Figure 1, our mask generation can provide more reliable and stronger augmentations. To accelerate the speed of training, we revise the mask augmentation code to achieve online mask augmentation. We initialize the ratio of adding mask as 0.2. The masked or non-masked images are fed into different backbones respectively.

3.3. Loss Function Searching Module

Loss function plays a key role in face recognition, such as Center Loss [32], ArcFace [7], and so on. To explore the most suitable loss for masked face recognition task, we follow LFS [29] to design a loss function searching module. We first define a loss function pool, which consists of several loss functions and assign same probability value α as initialization. Then different loss functions are applied to different backbones. We evaluate the performance of each backbone on validation set and update the α. Specifically, we enlarge the α if the backbone obtain good performance on validation set and decrease the α when the backbone is not very good. After each epoch, we share the best model parameters to other backbones. The loss function searching module can ensure the better loss functions are with larger utilization probability.

The mask generation module can provide variable and reliable mask augmentation. Loss function searching module ensures the more efficient loss functions are assigned larger utilization weights, which can help us find the most suitable loss function for the masked face recognition task. Based on these two benefits, we call our method as mask aware network.

4. Results Summarisation

In this section, we first review the datasets used in the exploration. Then, we introduce the implementation details for reproducing our results. After that, we show the effectiveness of our proposed modules using the submission logs. Finally, we present the future planning of our team.

4.1. Datasets

Glint360K. Glint360K [2] is collected by DeepGlint company. They clean, merge and release a large and clean face recognition dataset. Baseline models trained on Glint360K can easily achieve state-of-the-art. Glint360K has 17 million images that consist of 360K face identities. The number of images and identities are much more than other previous face recognition datasets.

Table 1. Results of our submission. Submission #1 uses the Glint360K as training dataset. Submissions from #2 to #6 use the WebFace42M to train the mask aware network. #2 and #3 are the different checkpoint models using official mask augmentation. #4 apply online maske augmentation method for training. #5 and #6 use different initial learning rates. $(\cdot)$ represents the rank place among all the teams in ICCV-MFR challenge (by 3, August, 2021).

Submission	All-Masked(MFR)	Wild-Masked (MFR)	Controlled-Masked (MFR)	All (SFR)	Wild (SFR)	Controlled (SFR)
#1	0.3873	0.4425	0.2891	0.01951	0.0323	0.0021
#2	0.1226	0.1487	0.0817	0.0254	0.0420	0.0026
#3	0.110	0.1323	0.075	0.0191	0.0315	0.0022
#4	0.1055	0.127	0.0716	0.0175	0.0288	0.0021
#5	0.1032	0.1239	0.0699	0.0167	0.0273	0.002
#6	0.1017	0.1221	0.0691	0.0166	0.0272	0.0018

WebFace260M. WebFace260M [40] is collected by Tsinghua University from the Internet. They first use 5M celebrity names from MS1M and IMDB website to search their public images from Google engine. Then they remove 1M celebrity names that have no public images. 200 images per identity are downloaded for top 10% subjects, while 100, 50, 25 images are reserved for remaining 20%, 30%, 40% subjects, respectively. Finally, they collect 4M identities and 265M images. They perform Cleaning Automatically by Self-Training (CAST) pipeline to automatically clean the noisy WebFace260M and obtain a cleaned training set named WebFace42M, consisting of 42M faces of 2M subjects. The ICCV-MFR challenge provides the whole WebFace42M dataset to participants who outperform the baseline result.

4.2. Implementation Details

Mask Aware Network (MAN) is implemented with Pytorch toolbox. Specifically, we train MAN by stochastic gradient descent (SGD) optimizer with batch size 256 on 8 V100 Nvidia GPU. The initial learning rate is set to 0.1 with decay 1e-4 and momentum 0.9. We divide learning rate by 10 at 10, 15, 18 epochs, and stop training at 20 epochs. For online mask generation ratio α, we initialize it with 0.2. For loss function pool, we utilize different margin values of ArcFace [7] from 0.05 to 0.45. We expect loss function searching module can search the most suitable mragin value for masked face recognition task. As mentioned by [2], we select IJBC as our validation set.

4.3. Submission Results

We show our important submission logs in Table 1. Submission #1 utilizes the Glint360K as train set and non mask augmentation. Almost all the images from Glint360K are with high quality, so the results of SFR are much better than MFR. The submission #1 outperforms the baseline, so we apply the whole WebFace42M data for training. Submission #2 and #3 are the results of different checkpoints models of training on the whole dataset. Official mask augmentation method is used in submission #2 and #3. It can easily find the effectiveness of the Webface42M. Submissions #2 and #3 outperforms the submission #1 with a large margin in term of masked face recognition tracks. We add our online mask generation module to original backbone to generate more variable masked images for training. The results are shown in submission #4. We add loss function searching module and obtain the further improvement as shown in submissions #5 and #6. We use different initial learning rate in submissions #5 and #6. The results of submission #6 achieve 5 first places and 1 third place among all the teams.

4.4. Future Planning

Using deeper model. The detection and recognition time of our best model are 157ms and 496ms respectively. The organizers require the total time must less than 1s. Obviously, we can use more deeper or complex convolutional neural network to improve our results further.

Applying Neural Architecture Search to MFR task. We only search the most suitable loss function for masked face recognition task. We can also try to search the most suitable architecture of convolutional neural network for this task. Neural Architecture Search (NAS) can search the best structure based on your predefined searching space. We expect to search better models using NAS.

Knowledge Distillation. The organizers add a strict limitation of the inference time. The previous works [14, 17, 28] has proven a small model can use the knowledge distillation to obtain comparable performance with the large model. Specifically, we can first train a very large and good performance model, and then employ knowledge distillation method to train a small model.

5. Conclusion

We present the method of team victor-2021 on ICCV-MFR challenge. In this paper, we propose Mask Aware Network (MAN) including a mask generation module and a loss function searching module to reduce the degraded performance of masked face recognition task. Online mask generation module improves the robustness of face recognition model by producing variable cases of masked images.

Loss function searching module aims to find the most suitable loss function for this challenge. We show the submission logs to demonstrate the effectiveness of the two well-designed modules. We also introduce the future plans of our team. On ICCV MFR challenge, our team victor-2021 achieves 5 first places (including 3 champions in standard face recognition and 2 champions in masked face recognition) and 1 third place by 3rd August 2021. These results demonstrate the robustness and generalization of our method both in standard or masked face recognition task.

6. Acknowledge

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2021-08-008). We thank CSCS (Swiss National Supercomputing Centre) for supporting our project to get access to the Piz Daint supercomputer. We thank TACC (Texas Advanced Computing Center) for supporting our project to get access to the Longhorn supercomputer and the Frontera supercomputer. We thank LuxProvide (Luxembourg national supercomputer HPC organization) for supporting our project to get access to the MeluXina supercomputer.

References

[1] Soad Almabdy and Lamiaa Elrefaei. Deep convolutional neural network-based approaches for face recognition. *Applied Sciences*, 9(20):4397, 2019. 1

[2] Xiang An, Xuhan Zhu, Yang Xiao, Lan Wu, Ming Zhang, Yuan Gao, Bin Qin, Debing Zhang, and Ying Fu. Partial fc: Training 10 million identities on a single machine. *arXiv preprint arXiv:2010.05222*, 2020. 3, 4

[3] Parama Bagchi, Debotosh Bhattacharjee, and Mita Nasipuri. Robust 3d face recognition in presence of pose and partial occlusions or missing parts. *arXiv preprint arXiv:1408.3709*, 2014. 1

[4] Peter N. Belhumeur, João P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):711–720, 1997. 1

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. 2

[6] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018. 2

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 2, 3, 4

[8] Hassen Drira, Boulbaba Ben Amor, Anuj Srivastava, Mohamed Daoudi, and Rim Slama. 3d face recognition under expressions, occlusions, and pose variations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2270–2283, 2013. 1, 2

[9] Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Topology preserving structural matching for automatic partial face recognition. *IEEE Transactions on Information Forensics and Security*, 13(7):1823–1837, 2018. 1, 2

[10] Ashwini S Gawali and Ratnadeep R Deshmukh. 3d face recognition using geodesic facial curves to handle expression, occlusion and pose variations. *International Journal of Computer Science and Information Technologies*, 5(3):4284–4287, 2014. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[12] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang. Face recognition using laplacianfaces. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):328–340, 2005. 1

[13] Andrew Kae, Kihyuk Sohn, Honglak Lee, and Erik Learned-Miller. Augmenting CRFs with Boltzmann machine shape priors for image labeling. 2013. 1

[14] Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*, 2016. 4

[15] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 2

[16] Aleix M Martinez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Transactions on Pattern analysis and machine intelligence*, 24(6):748–763, 2002. 1, 2

[17] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 4

[18] G Nirmala Priya and RSD Wahida Banu. Occlusion invariant face recognition using mean based weight matrix and support vector machine. *Sadhana*, 39(2):303–315, 2014. 1

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[20] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014. 1

[21] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2892–2900, 2015. 2

[22] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with

convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[23] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 1

[24] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 2

[25] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 2

[26] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020. 1

[27] Kai Wang, Shuo Wang, Zhipeng Zhou, Xiaobo Wang, Xiaojiang Peng, Baigui Sun, Hao Li, and Yang You. An efficient training approach for very large scale face recognition. *arXiv preprint arXiv:2105.10375*, 2021. 1

[28] Xiaobo Wang, Tianyu Fu, Shengcai Liao, Shuo Wang, Zhen Lei, and Tao Mei. Exclusivity-consistency regularized knowledge distillation for face recognition. In *European Conference on Computer Vision*, pages 325–342. Springer International Publishing, 2020. 4

[29] Xiaobo Wang, Shuo Wang, Cheng Chi, Shifeng Zhang, and Tao Mei. Loss function search for face recognition. In *International Conference on Machine Learning*, pages 10029–10038. PMLR, 2020. 2, 3

[30] Xiaobo Wang, Shuo Wang, Shifeng Zhang, Tianyu Fu, Hailin Shi, and Tao Mei. Support vector guided softmax loss for face recognition. *arXiv preprint arXiv:1812.11317*, 2018. 2

[31] Xiaobo Wang, Shifeng Zhang, Zhen Lei, Si Liu, Xiaojie Guo, and Stan Z Li. Ensemble soft-margin softmax loss for image classification. *arXiv preprint arXiv:1805.03922*, 2018. 2

[32] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 3

[33] Renliang Weng, Jiwen Lu, and Yap-Peng Tan. Robust point set matching for partial face recognition. *IEEE transactions on image processing*, 25(3):1163–1176, 2016. 1

[34] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 1

[35] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2008. 1

[36] Yang Yang, Shengcai Liao, Zhen Lei, and Stan Z Li. Large scale similarity learning using similar pairs for person verification. In *Thirtieth AAAI conference on artificial intelligence*, 2016. 2

[37] Jinxing Ye, Xioajiang Peng, Baigui Sun, Kai Wang, Xiuyu Sun, Hao Li, and Hanqing Wu. Learning to cluster faces via transformer. *arXiv preprint arXiv:2104.11502*, 2021. 1

[38] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 1, 2

[39] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jia Guo, Jiwen Lu, Dalong Du, and Jie Zhou. Masked face recognition challenge: The WebFace260M track report. *arXiv:2108.07189*, 2021. 2

[40] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021. 4