The Devil is in the Detail: Deep Feature Based Disguised Face Recognition Method

Shumin Zhu[1], Jianjun Qian[1(✉)], Yangwei Dong[1], and Waikeung Wong[2]

[1] PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{zhushumin,csjqian,dongyangwei}@njust.edu.cn

[2] Institute of Textiles and Clothing, The Hong Kong Polytechnic University, Hong Kong, China
calvin.wong@polyu.edu.hk

Abstract. Face recognition have been developed rapidly, launched by the breakthrough of Deep learning based face representation method. However, disguised face verification in the wild is still a challenge problem. To address this issue, we propose a novel deep feature based disguised face recognition scheme (DDFR). DDFR introduces the multi-scale residual network with AM-softmax loss for learning face representation. In training stage, we put the different occlusions (mask, sunglasses and scarf et al.) on clean face images to enhance the diversity of training set. Meanwhile, both aligned face image and un-aligned face image are combined to improve the discriminative power of feature representation for disguised face verification. Experimental results demonstrate that the proposed method achieves the better results than state-of-the-art methods on the DFW (disguised face in the wild) set.

Keywords: Disguised face verification · Face alignment · Add face occlusion · Multi-scale feature extraction

1 Introduction

Face recognition has been a hot topic in the field of computer vision and has been widely used in public security and finance. Recently, Deep learning based representation methods achieve the landmark breakthrough in face recognition [3,10,12,18]. This achievement is mainly due to the applications of a better convolutional neural network architecture [6,13,16,17] and a more restrictive loss function. However, disguised face recognition is still a challenging problem

This work was supported by the National Science Fund of China under Grant Nos. 61876083, 61876084, U1713208, and Program for Changjiang Scholars. Hong Kong Scholar program, The Hong Kong Polytechnic University (YZ2K).

Y. Peng et al. (Eds.): PRCV 2020, LNCS 12306, pp. 420–431, 2020.
https://doi.org/10.1007/978-3-030-60639-8_35

since it contains personal unconscious or conscious cover of the face to hide personal identity, as well as one person imitates another to deceive the face recognition system. In this task, large intra-class distance and small inter-class distance make unconstrained disguised face verification very difficult. To solve this problem, there are many works have been developed, ranging from the sparse representation based method to deep convolutional neural network, in the past decades.

In the literature, Wright et al. [19] proposed a sparse representation based method to handle face recognition with real-disguise. To further improve the robustness of the sparse model, robust sparse representation models are developed for robust face recognition by using M-estimator to characterize the error term [7,21]. Subsequently, Yang et al. [20] employed nuclear norm to describe the error term and presented a novel nuclear norm based matrix regression model to solve facial images contain disguise or occlusion. Qian et al. [11] introduced the low rank regularized term to ridge regression for solving disguise face recognition. However, these methods overlooked open sets of subjects, which is limited for real world applications. To facilitate the research of unrestricted disguised faces recognition, Kushwaha et al. [9] proposed a novel Disguised Faces in the Wild (DFW) dataset. DFW is mainly used to evaluate the performance of various methods in dealing with disguised face recognition. The authors of DFW also organize a competition [14] in conjunction with the International Conference on Computer Vision and Pattern Recognition (CVPR) 2018. Based on DFW, many deep learning based methods are developed to solve disguised face verification. Zhang et al. [23] proposed a two-stage training approach for this task. At the first stage, they employed generic aligned face images and unaligned face images to train two 64-layer DCNNs [10] in conjunction with AM-Softmax [18]. At the second stage, PCA is used to obtain the low dimensional compact feature representation. Smirnov et al. [15] proposed a new deep embedding learning method for disguised face recognition. They used general face images to train AEFANet with Auxiliary Embedding. Bansal et al.[1] combined ResNet-101 [6] and Inception-ResNet-v2 [16] with L2-constrained Softmax for handling disguised face verification (Fig. 1).

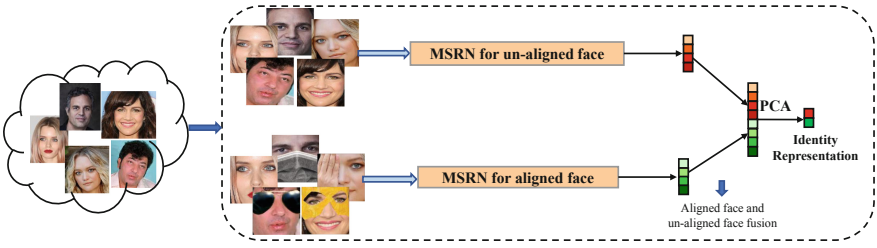


Fig. 1. The proposed Deep feature based disguised face recognition scheme (DDFR).

However, above mentioned methods use nearly ten million face images to learn the face representation model. To further improve the disguised face verification performance, this paper presents a novel method for capturing intrinsic feature of face image with little training data as possible. The main contributions of our work are as follows:

- We propose a multi-scale residual (MSR) block to capture more detailed facial features for improving the performance to distinguish the person and its impersonator.
- We put the different face occlusions (mask, sunglasses and scarf et al.) on clean face images to enhance the diversity of training set and increases the intra-class differences of training set.
- We combine the feature of aligned face image and unaligned face image to improve the discriminative power of feature representation for disguised face verification. Experimental results on the DFW dataset demonstrate that our method achieves better performance than state-of-the-art methods.

2 Deep Disguised Face Verification Framework

In this section, we introduce the multi-scale feature representation method for obtaining more detailed facial features. The facial occlusion synthesis scheme is proposed to enrich the diversity of training set. Finally, we fuse the features of aligned facial image and unaligned facial image into one disguised face verification framework.

2.1 Multi-scale Feature Extraction to Capture More Facial Details

As well known, ResNet is a good tool to capture the image feature with deeper network [6]. Based on this, W. Liu et al. developed ResNet-like block by combining 3×3 convolution kernels with a residual unit for face representation and achieved remarkable results [10].

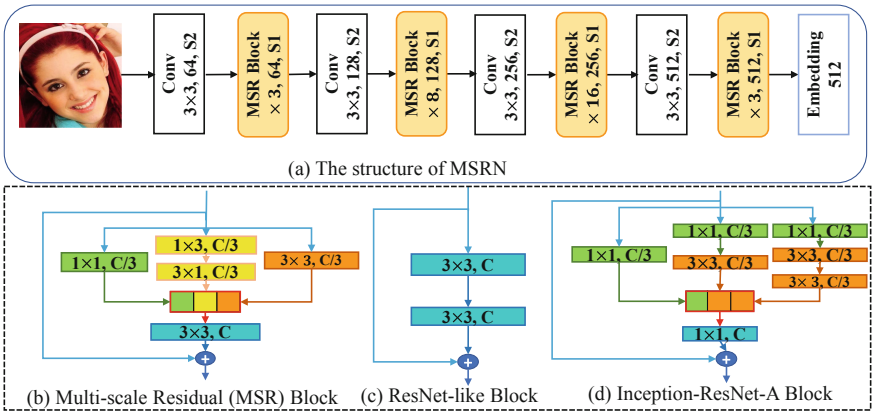
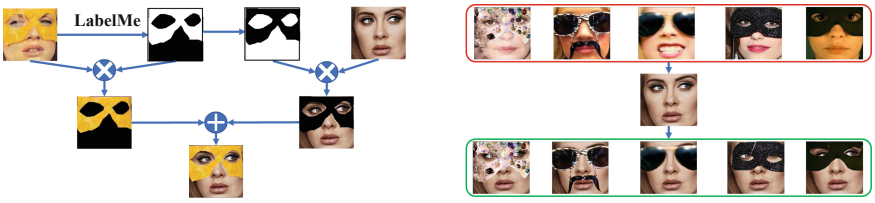


Fig. 2. (a) Our MSRN structure. (b) Our Multi-scale Residual Block. (c) ResNet-like Block. (d) Inception-ResNet-A Block.

ResNet-like block motivates us to design the multi-scale residual block for matching faces with intentional and unintentional disguises. Compared with ResNet-like block, our multi-scale residual block includes convolution kernels of different sizes and employs the different nonlinear transform to combine the features. The schema of Multi-scale Residual block has to be shown in Fig. 2 (b). Suppose that there are 64 feature maps and the size of each feature map is 80×80 as the input of this block. Based on this, we convolved the feature map with three different scale convolution kernels. The number of convolution kernels for each scale is 21. Then, we connect the convolved feature maps together and use 3×3 convolution kernels to further represent the connected feature maps. Here, the number of convolution kernels is 64. Finally, all the convolved feature and the input feature maps are added together as a whole.

The main difference between Multi-scale Residual block and Inception-ResNet block is that the Inception-ResNet block draws the idea of Network-in-Network to reduce dimension, it should use a 1×1 size convolution kernel before and after using a multi-scale convolution kernel. The structure of Inception-Resnet-A is shown in Fig. 2 (d). In addition, Multi-scale Residual block can capture rich facial feature than Inception-ResNet block. The experiments in Sect. 3 also support our view.



(a) The process of adding occlusion (b) Some synthetic occluded face images

Fig. 3. The pipeline of facial occlusion synthesis and some synthesized results.

(a) The occlusion styles: glasses, mouth-muffle, gestures, masks, and other materials from left to right.

(b) Some synthesized occluded face images are used to train the occlusion detection network.

(c) Some detection results of Res-Unet on DFW testing set.

Fig. 4. The various of occlusion styles. Our synthesize training set of Res-Unet and some detection results of Res-Unet.

2.2 Facial Occlusion Synthesis

In the task of disguised face recognition, the person can wear glasses and masks to weaken its distinguishability, and the pretender can also increase the similarity between itself and the imitated person in this way. In DFW, it is easy to lead the overfitting problem since the diguised face images in the training set are limited. In general, most methods use a large number of face images without disguise in training stage. To fit the complex face data in the disguised face test set, the previous work only expand the number and category of face images in the training set.

Here, we just put occlusions on partial of clean face images to enrich the diversity of the training set for improving the disguised face verification performance. To ensure the authenticity of facial occlusion synthesis, we employs the LabelMe to obtain the occlusion part and then synthesized the facial occlusion images as shown in Fig. 3 (a). Specifically, we collect face images with different occlusions from the training set of DFW, MAFA [5] and NUST-RF dataset. All face images are aligned by using the same face alignment method. There are 285 occlusion styles, including glasses (55), mouth-muffle (53), masks (53), gestures (55), and other materials (69). Some face images shows in Fig. 4(a). Then, the LabelMe is employed to mark the occlusion's position in the face image. Subsequently, we can obtain the synthesed occlusion face by combing the occlusion and clean face images.

To add an appropriate proportion of face images with occlusion to the training set, we propose that the rate of occluded face images in the training set should be consistent with the test set. We introduce an occlusion detection network (the backbone is Res18-Unet [4]) to detect whether the images in the test set are occluded and further count the proportion of face images containing occlusion in the test set. And then we randomly select the same proportion face images of each person in the training set to put occlusion on them. For occlusion detection network, the training set is composed of various occlusion styles and 100,000 face images without occlusion from CASIA-WebFace [22] dataset. Some occlusion detection results of the network on DFW are shown in Fig. 4 (c).

2.3 Aligned and Unaligned Face Feature Fusion

It is known that aligned face images have the advantage of eliminating posture changes compared to unaligned face images, which makes the model pay more attention to details such as facial texture. However, the unaligned face images possibly contain some irregular discriminative information. We think that the irregular discriminative information of unaligned face image and the discriminative information of aligned face image can be combined together to further improve the discriminative power of face representation. Finally, PCA is then used to achieve the low-dimensional feature vector.

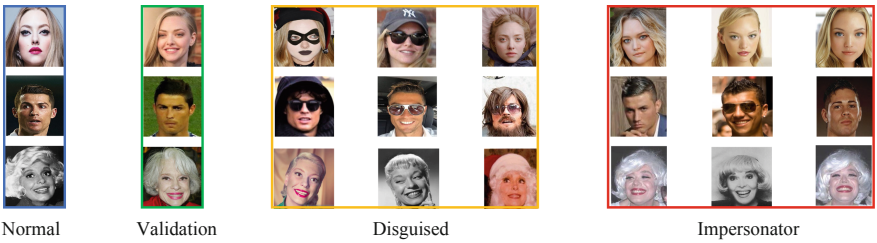


Fig. 5. Some example images of Disguised Faces in the Wild. It consists of four kinds of images: normal, validation, disguised and impersonator.

3 Experiments

3.1 Disguised Faces in the Wild Dataset

The Disguised Faces in the Wild (DFW) data set includes 11157 images from 1000 different identities. Among them, the training set contains 3386 face images of 400 people. And the test set includes 7,771 face images of 600 people. Most of the identities have four kinds of face images: Normal, Validation, Disguised and Impersonator. These four types images are shown in Fig. 5.

There are three pre-defined protocols. Protocol 1 aims to evaluate the performance of face verification methods under impersonation only. There are 25,046 pairs of face images for this protocol. In Protocol 2, the given face verification methods are evaluated for disguises via obfuscation only. The number of image pairs is 9,041,283 for this protocol. And Protocol 3 is used for evaluating the given methods on the whole dataset. The total number of image pairs for this protocol is 9,066,329.

3.2 Experimental Settings

Mini Training Set. The mini training set is designed to facilitate the ablation study. This training set is composed of CASIA-WebFace [22], PubFig [8] and the training set of DFW [9]. We removed the identities overlap between training set and testing set strictly according to provided identity names. More details for removing the overlap identities can be found in AM-Softmax paper [18]. The final generic mini training dataset includes 10,397 identities and 444,895 (0.44M) face images.

Big Training Set. This set is an extension of the mini training set. The expanded images are all from the VGGFace2 data set [2]. Compared with Mini training set, there are another 8047 persons and each person have about 200 face images. The final expanded big training set includes 18444 people and about 2,039,485 (2.04M) face images.

Training Setting. In our experiments, all face images are resized to 160×160. AM-Softmax is used in our model. The parameters m (cosine margin constrain) is 0.35 and s (norm-scale of features) is 30. The batch size is 24 for the mini training set and the large training set. For the mini training set, the learning rate starts from 0.1 and is divided by 10 at 140K, 180K iterations and the maximum iterations is 200K. For the large training set, to make full use of this data set, the learning rate starts from 0.1 and is divided by 10 at 700K, 900K iterations, and the maximum iterations are 1M. In addition, we use random horizontal flipping for data augmentation. For PCA, we select the first 250 eigenvectors to form the projection matrix.

3.3 Experimental Results

We compare our verification results on DFW with state-of-the-art methods published in DFW competition. The compared results are shown in the Table 1. And the ROC curves on the mini training set and the big training set are shown in the Fig. 6.

We can see from the Table 1 that when the training set is 0.44M, our method outperforms Occlusionface on three protocols. Specifically, our method is 13% higher than the Occlusionface at 0.1% FAR on Protocol-1. On Protocol-2 and Protocol-3, our method is 6% and 7% higher than the Occlusionface at 1%FAR and 0.1%FAR, respectively. It is worth mentioning that even with 0.44M training set, our method is nearly 6% higher than UMDNet whose training set is 5.6M at 0.1% FAR on Protocol-1. And the results of our method on other protocols are also similar with UMDNets.

When the number of face images in the training set increases to 2.04M, the results of our method on three protocols are obviously better than UMDNets. Compared with the AEFRL whose training set is 8.3M, even through our methods is 0.5% lower than AEFRL at 0.1% FAR on Protocol-1, our method is 2%–4% higher than AEFRL at FARs on Protocol-2 and Protocol-3. Compared with MiRA-Face whose training set is 7.6M, our method is nearly 1% higher than MiRA-Face at 1% FAR and is 9% higher at 0.1% FAR on Protocol-1. On Protocol-2 and Protocol-3, the results of our method are only a tiny difference with MiRA-Face. Overall, our method outperforms other state-of-the-art methods with less training set.

3.4 Ablation Study

Effect of Face Alignment Scale. We use two different face alignment scales (one contains hair and face contours, and the other remove these two parts) to train our multi-scale residual network model, and compare their verification results on DFW test set.

From Table 2, we can see that the alignment scale of the face without hair shows a clear advantage on three protocols. On Protocol-1, the alignment scale of without hair part is nearly 5% higher than the face images with hair part at 0.1% FAR. On Protocol-2 and Protocol-3, the alignment scale without hair is

Table 1. Verification accuracy (%) of our method and other published methods on three protocols

Method	**Protocol-1** GAR		**Protocol-2** GAR		**Protocol-3** GAR	
	@1%FAR	@0.1%FAR	@1%FAR	@0.1%FAR	@1%FAR	@0.1%FAR
AEFRL (8.3M)	**96.80**	57.64	87.82	77.06	87.90	75.54
MiRA-Face (7.6M)	95.46	51.09	**90.65**	80.56	**90.62**	79.26
UMDNets (5.6M)	94.28	53.27	86.62	74.69	86.75	72.90
Occlusionface (0.52M)	93.44	46.21	80.45	66.05	80.80	65.34
DDFR (0.44M)	93.86	59.23	86.14	73.43	86.35	72.47
DDFR (2.04M)	96.30	**60.84**	90.19	**80.61**	90.30	**79.57**

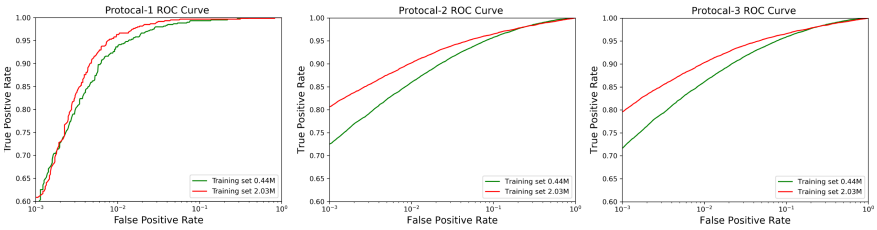


Fig. 6. ROC Curves for three protocols on the mini training set and the big training set

2%–5% higher than that with hair at FARs. In general, the face images without the hair part can achieve better performance than that with hair cause it can remove the interference of hairstyle and beard.

Table 2. Verification accuracy (%) of our model trained with different face align scale

Method	**Protocol-1** GAR		**Protocol-2** GAR		**Protocol-3** GAR	
	@1%FAR	@0.1%FAR	@1%FAR	@0.1%FAR	@1%FAR	@0.1%FAR
With hair scale	91.30	51.55	80.47	64.52	80.86	63.88
Without hair scale	**92.87**	**55.48**	**83.09**	**69.07**	**83.38**	**68.50**

Effect of Different Feature Extraction Block. To compare the feature extraction capabilities of MSR block, Resnet-like block and Inception-Resnet-A block, we prepare three different models. The first one is MSRN, then we replace the MSR block in the MSRN with Resnet-like block and Inception-Resnet-A block to construct Resnet-like model and Inception-Resnet-A model, respectively. The structures of these three blocks are shown in Fig. 2. And the verification results of the three models on the DFW test set are shown in Table 3 (Fig. 7).

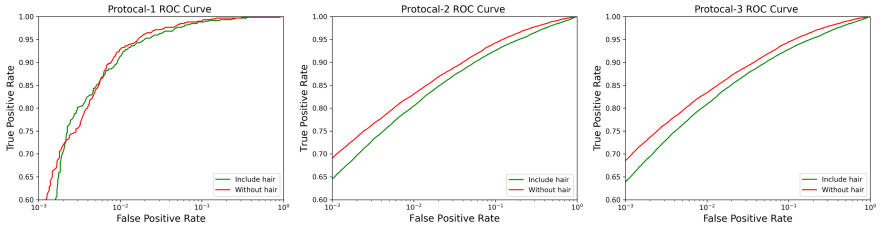


Fig. 7. ROC curves for three protocols with different face alignment scale

We can find that whether on Protocol-1 (impersonation) or Protocol-2 (obfuscation), MSR Block has stronger feature representation capabilities than the other two network Blocks. Especially on Protocol-1, the model with the MSR Block is 4.5% higher than the ResNet-like Block model and 7% higher than the Inception-Resnet-A Block model at 0.1% FAR (Fig. 8).

Table 3. Verification accuracy (%) of ResNet-Like Block (R-L Block), Inception-Resnet-A Block (I-R-A Block), and our MSR Block on three protocols.

Method	**Protocol-1** GAR		**Protocol-2** GAR		**Protocol-3** GAR	
	@1%FAR	@0.1%FAR	@1%FAR	@0.1%FAR	@1%FAR	@0.1%FAR
R-L Block	90.88	50.98	81.48	67.14	81.89	66.72
I-R-A Block	91.93	48.44	81.77	67.51	82.11	67.04
MSR Block	**92.71**	**55.48**	**83.09**	**69.07**	**83.38**	**68.50**

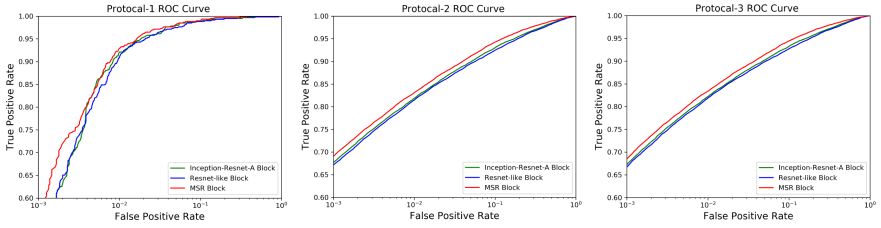


Fig. 8. ROC curves of different blocks on three protocols

Effect of Increasing the Rate of Occluded Face Images. To understand the impact of different proportions of synthetic occluded face images in the training set on the model's disguised face verification results, we compare three different rates 5%, 20% and 100%. Among them, 5% is also the rate of occlusion images in the DFW test set.

We can see that when we keep the rate of the occluded face images in training set consistent with the proportion of the occluded face images in the DFW test set, the effectiveness of our model on disguised face verification tasks is improved. However, when we increase the proportion of synthetic occluded face images synthesized in the training set to 20% and 100%, the verification performance of the model on the DFW test set decreases. On Protocol-1, our model trained with 5% synthetic occluded face images is 2% higher than the model trained with ordinary face images at 0.1% FAR. On Protocol-2 and Protocol-3, our model trained with 5% synthetic occluded face images is 1% higher than the model trained with ordinary face images at 0.1% FAR. Therefore, it is effective to add synthetic occlusion face images to the training set. However, the ratio of the added synthetic occlusion images should be consistent with the occlusion image ratio in the test set (Table 4).

Table 4. Verification accuracy (%) of our model trained with or with out occlusion faces

Method	**Protocol-1** GAR		**Protocol-2** GAR		**Protocol-3** GAR	
	@1%FAR	@0.1%FAR	@1%FAR	@0.1%FAR	@1%FAR	@0.1%FAR
Ordinary face	93.64	57.20	85.95	72.47	86.09	71.65
Occluded face (20%)	93.03	55.66	86.11	72.46	86.26	71.58
Occluded face (100%)	91.54	52.85	82.52	65.77	82.74	65.06
Occluded face (5%)	**93.86**	**59.24**	**86.14**	**73.43**	**86.35**	**72.48**

Effect of Facial Feature Fusion. We train aligned MSRN model and unaligned MSRN model by using aligned and unaligned face images, respectively. Then, we merge the aligned and unaligned face features in the embedding layer and use PCA to achieve the compressed feature. The verification results of these three methods on the DFW test set are shown in the Table 5. And the ROC curves of these three methods on the three protocols are shown in Fig. 9. On Protocol-1, the feature fusion method is 4% higher than aligned MSRN model and 2% higher than unaligned MSRN model at 0.1% FAR. On Protocol-2, the feature fusion method is 3% higher than aligned MSRN model and 5% higher than unaligned MSRN model at 1% FAR. And at 0.1% FAR, the GAR of feature fusion method is 4% higher than aligned MSRN model and 8% higher than unaligned MSRN model. In general, the feature fusion method has more advantages than the single feature method.

Table 5. Verification accuracy (%) of our model trained with aligned face images, unaligned face images and feature fusion of both.

Method	**Protocol-1** GAR		**Protocol-2** GAR		**Protocol-3** GAR	
	@1%FAR	@0.1%FAR	@1%FAR	@0.1%FAR	@1%FAR	@0.1%FAR
Unalign	92.70	57.20	80.75	64.91	81.03	64.21
Align	92.87	55.48	83.09	69.06	83.38	68.50
Fusion+PCA	**93.86**	**59.24**	**86.14**	**73.43**	**86.35**	**72.47**

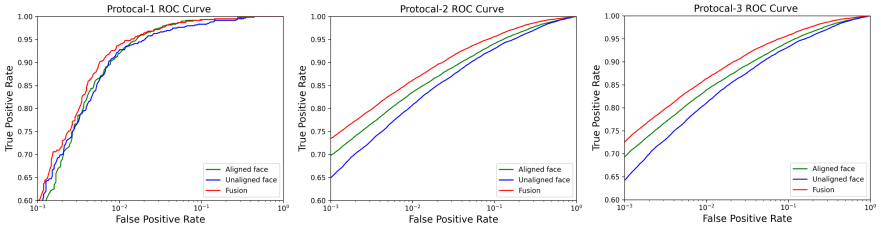


Fig. 9. ROC curves of aligned face feature, unaligned face feature and feature fusion of both for three protocols

4 Conclusion

In this article, we propose a multi-scale residual network and a method of adding real occlusion to the training set for disguised face verification. Compared with the deep metric learning method, our proposed method enriches the facial feature by using multi-scale residual blocks and increases the diversity of training set samples by adding real occlusion on a clean face. Another advantage is that our method uses fewer training samples and achieves better results than the state-of-the-art methods. In future work, we will further investigate how to design efficient deep face representation model with little training set as possible. It is always interesting in developing attention neural network to handle disguised face representation.

References

1. Bansal, A., Ranjan, R., Castillo, C.D., Chellappa, R.: Deep features for recognizing disguised faces in the wild. In: CVPR Workshops, pp. 10–16 (2018)
2. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: a dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74. IEEE (2018)
3. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: additive angular margin loss for deep face recognition. In: CVPR, pp. 4690–4699 (2019)
4. Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C.: ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogramm. Remote Sens. **162**, 94–114 (2020)

5. Ge, S., Li, J., Ye, Q., Luo, Z.: Detecting masked faces in the wild with LLE-CNNs. In: CVPR, pp. 2682–2690 (2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
7. He, R., Zheng, W.S., Hu, B.G.: Maximum correntropy criterion for robust face recognition. TPAMI **33**(8), 1561–1576 (2010)
8. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV, pp. 365–372. IEEE (2009)
9. Kushwaha, V., Singh, M., Singh, R., Vatsa, M., Ratha, N., Chellappa, R.: Disguised faces in the wild. In: CVPR Workshops, pp. 1–9 (2018)
10. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: SphereFace: deep hypersphere embedding for face recognition. In: CVPR, pp. 212–220 (2017)
11. Qian, J., Yang, J., Zhang, F., Lin, Z.: Robust low-rank regularized regression for face recognition with occlusion. In: CVPR Workshops, pp. 21–26 (2014)
12. Ranjan, R., Castillo, C.D., Chellappa, R.: L2-constrained softmax loss for discriminative face verification. arXiv preprint arXiv:1703.09507 (2017)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
14. Singh, M., Singh, R., Vatsa, M., Ratha, N.K., Chellappa, R.: Recognizing disguised faces in the wild. IEEE Trans. Biometrics Behav. Identity Sci. **1**(2), 97–108 (2019)
15. Smirnov, E., Melnikov, A., Oleinik, A., Ivanova, E., Kalinovskiy, I., Luckyanets, E.: Hard example mining with auxiliary embeddings. In: CVPR Workshops, pp. 37–46 (2018)
16. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In: AAAI (2017)
17. Szegedy, C., et al.: Going deeper with convolutions. In: CVPR, pp. 1–9 (2015)
18. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Process. Lett. **25**(7), 926–930 (2018)
19. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. TPAMI **31**(2), 210–227 (2008)
20. Yang, J., Luo, L., Qian, J., Tai, Y., Zhang, F., Xu, Y.: Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. TPAMI **39**(1), 156–171 (2016)
21. Yang, M., Zhang, L., Yang, J., Zhang, D.: Regularized robust coding for face recognition. IEEE Trans. Image Process. **22**(5), 1753–1766 (2013). https://doi.org/10.1109/TIP.2012.2235849
22. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014)
23. Zhang, K., Chang, Y.L., Hsu, W.: Deep disguised faces recognition. In: CVPR Workshops, pp. 32–36 (2018)
24. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)