ArcFace for Disguised Face Recognition

Jiankang Deng
Imperial College&FaceSoft
UK
j.deng16@imperial.ac.uk

Stefanos Zafeiriou
Imperial College&FaceSoft
UK
s.zafeiriou@imperial.ac.uk

Abstract

Even though deep face recognition is extensively explored and remarkable advances have been achieved on large-scale in-the-wild dataset, disguised face recognition receives much less attention. Face feature embedding targeting on intra-class compactness and inter-class discrepancy is very challenging as high intra-class diversity and inter-class similarity are very common on the disguised face recognition dataset. In this report, we give the technical details of our submission to the DFW2019 challenge. By using our RetinaFace for face detection and alignment and ArcFace for face feature embedding, we achieve state-of-the-art performance on the DFW2019 challenge.

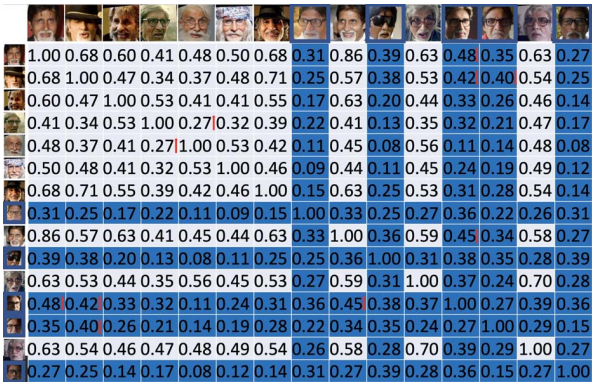
1. Introduction

Face representation using Deep Convolutional Neural Network (DCNN) embedding is the method of choice for face recognition [22, 4]. DCNN maps the face image, typically after a pose normalisation step [6, 28, 15, 16, 7, 8, 9], into a feature embedding that has intra-class compactness [20, 3] and inter-class discrepancy.

Even though remarkable advances have been achieved on large-scale unconstrained face recognition, it has often been observed that most of the face recognition systems are susceptible to spoofing techniques or disguises. Although anti-spoofing is well explored [33, 29, 19], disguised face recognition [13, 24] receives less attention.

Disguised face recognition presents the challenge of face verification and identification under both intentional and unintentional distortions. For instance, a criminal may intentionally attempt to conceal his identity by using external disguise accessories (*e.g.* sunglasses or breathing mask), thereby resulting in a challenging genuine (positive) match problem for an authentication system. In addition, an individual might intentionally attempt to impersonate another person by professional make up, resulting in a challenging imposter (negative) unauthorised login for the face recognition system. In [13, 24], both obfuscation and impersonation are considered as the disguised face recognition problem. Obfuscation gives rise to intra-variance while impersonation results in high inter-similarity, which poses great challenge for current face recognition system.

(a) Amitabh Bachchan



(b) Cosine Similarity Matrix by ArcFace

Figure 1. Sample images from the DFW2019 dataset. (a) Images within blue boxes (Impersonation) are not Amitabh Bachchan. (b) The cosine similarity matrix is predicted by ArcFace [4]. For negative pairs (in blue), the similarity score should not be too high. For positive pairs (in white), the similarity score should not be too low. We use red arrows to mark the challenging pairs.

In Figure 1(a) illustrates sample images of "Amitabh Bachchan" from the DFW2019 dataset. The face images are detected and aligned by RetinaFace [5], and the cosine similarity matrix is predicted by ArcFace [4]. For this particular subject, there are positive pairs with low similarity scores (< 0.3) and negative pairs with high similarity scores (> 0.4). For negative pairs (in blue), the similarity score should not be too high. For positive pairs (in white), the

similarity score should not be too low. However, we find some challenging pairs (marked by red arrows). We can also easily find the confusion between positive and negative pairs around the interval of [0.3,0.4]. It can be observed from Figure 1(b), disguised face images result in increased intra-class variations, while the impersonator images render lower inter-class variability.

In this report, we first set up the baseline by using our RetinaFace [5] and ArcFace [4]. RetinaFace is a practical state-of-the-art face detector, which also outputs five facial landmarks for face normalisation. ArcFace is a state-of-the-art face feature embedding method. Then, we explore some extra intra and inter loss to further improve the performance on disguised face recognition. Finally, our solution achieves state-of-the-art performance on the DFW2019 challenge.

2. Related Work

Deep Face Recognition. Face recognition via deep learning has achieved a series of breakthroughs in recent years. Triple loss [22] pioneered employing the margin penalty on triplets and obtained state-of-the-art performance on face recognition. The margin penalty can enhance intra-class compactness and inter-class discrepancy at the same time as the Triplet loss targets on that the Euclidean distance between positive pairs should be closer enough by a clear margin than the Euclidean distance between negative pairs. Even though the motivation of the Triplet loss matches the target of face feature embedding, the training procedure of the Triplet loss is very challenging as (1) there is a combinatorial explosion in the number of face triplets especially for large-scale datasets, leading to a significant increase in the number of iteration steps, and (2) semi-hard sample mining is tricky for the model training. Since the margin penalty is a confirmed effective method for deep face recognition and the image-to-image comparison in Triple loss [22] is too tricky, recent improvement works [17, 27, 26, 4] focused on incorporating margin penalty into a more feasible framework, softmax loss, which has extensive image-to-class comparisons within the multiplication step between the embedding feature and the linear transformation matrix. As illustrated in Figure 2, we give the motivations of translating Triplet loss [22] into Arcface loss [4]. Naturally, each line of the linear transformation matrix is viewed as the class centre to represent each class in SphereFace [17], CosFace [27, 26] and ArcFace [4].

Disguised Face Recognition. Disguised face recognition [13, 24] is a special case of deep face recognition but it is more challenging as (1) the intra-class distance can be very large due to unintentional obfuscation (*e.g.* aging, heavy make-up, plastic surgery and occlusion) on the face region, and (2) the inter-class distance can be very small due to the intentional impersonation. The combination of both unintentional and intentional disguises render the problem of disguised face recognition an arduous task. In the DFW2018 challenge [13, 24], some methods (in Table 1) were proposed to solve the problem of disguised face recognition. However, only common deep face recognition methods were used in the DFW2018 challenge to solve the particular problem of disguised face recognition except for some fine-turning on the training dataset of DFW2018.

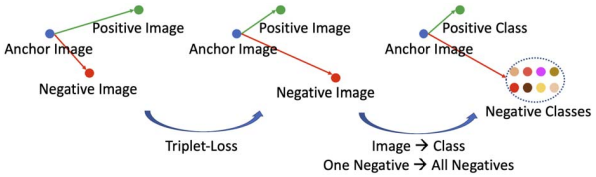


Figure 2. Motivations of translating Triplet loss [22] into Arcface loss [4]. Image-to-class comparison is more efficient and stable than image-to-class comparison as (1) class number is much smaller than image number, and (2) each class can be represented by a smoothed vector which is updated online during training.

3. Our Solution

3.1. Face Detection and Alignment by RetinaFace

RetinaFace [5] is a single-stage face detection method which can jointly predict face boxes and five facial landmarks. For any training anchor, RetinaFace minimises the following multi-task loss:

$$L_1 = L_{cls}(p_i, p_i^*) + \lambda_1 p_i^* L_{box}(t_i, t_i^*) + \lambda_2 p_i^* L_{pts}(l_i, l_i^*). \quad (1)$$

The loss-balancing parameters λ_1 and λ_2 are set to 0.25 and 0.1, respectively. Please refer to [5] for more details.

3.2. Face Feature Embedding by ArcFace

ArcFace [4] is an additive angular margin loss designed on the softmax loss. Based on the feature and weight normalisation, ArcFace adds an additive angular margin penalty m between x_i and W_{y_i} to simultaneously enhance the intra-class compactness and inter-class discrepancy.

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n} e^{s\cos\theta_j}}. \quad (2)$$

Please refer to [4] for more details.

3.3. Further Improvement by Intra and Inter Loss

ArcFace views each class as a smoothed vector, which is efficient and stable during training. However, the pre-trained model by ArcFace can not ideally project all face images of one subject into one point in the high dimension space during testing. In fact, facial appearance variations due to subject(*e.g.* aging, heavy make-up, plastic surgery

Method	Brief Description
VGGFace	Pre-trained VGG-Face model + Cosine distance
AEFRL	MTCNN + 4 networks + Cosine distance
ByteFace	Weighted ensemble of 3 CNNs (Centre loss, SphereFace, Joint Bayesian)
DDRNET	Inception Network with Centre Loss
DisguiseNet	Siamese network on pretrained VGG-Face + Cosine distance
DR-GAN	MTCNN + DR-GAN + Cosine distance
LearnedSiamese	Siamese Neural Network
MEDC	MTCNN + Ensemble of 3 CNNs (Centre loss, SphereFace) + Average Cosine distance
MiRA-Face	MTCNN + RSA Detector + Ensemble of CNNs
OcclusionFace	MTCNN + Fine-tuned ResNet-28
Tessellation	Siamese network with triplet loss model
UMDNets	All-In-One Detector + ensemble of 2 CNNs
WVU_CVL	MTCNN + CNN + Softmax

Table 1. Solutions from the DFW2018 challenge [13, 24]. By default, original face boxes given by the dataset were used, and some solutions used other face detection and alignment methods (*e.g.* MTCNN [32], RSA [18] and All-In-One [21]) to get the face crops.

and occlusion) and environment(*e.g.* camera pose, illumination, blur and low resolution) can significantly change intra-class distance. The largest intra-class distance between samples is larger than the largest image-to-class distance, and the smallest inter-class distance between samples is smaller than the smallest image-to-class distance. Therefore, we can go back to image-to-image comparison method to further improve the ArcFace model. To further improve the performance of ArcFace on disguised face recognition, we try to explore intra-loss and inter-loss by hard sample mining in this paper.

Sampling Probability Update. When the ArcFace model has roughly converged, most of the samples (around 90%) in the dataset have been well classified and do not contribute to the network training. To improve the pre-trained ArcFace model, we assign sampling probability to each sample [14]. During training, when the sample is correctly classified in this iteration, we pass the signal to the data layer and reduce its sampling probability. Otherwise, we increase its sampling probability. Therefore, the samples which are correctly classified will be gradually ignored and the samples which are incorrectly classified will be repeatedly learned as the training progresses. We also set a minimum sampling probability, in case simple samples are never sampled. To avoid over-fitting on the noise data, we also add feedback for noisy samples, as the noisy samples are continuously mis-classified and have large sampling probability. For each sample in mini-batch, if the cosine similarity between its feature and its corresponding centre is lower than a threshold, we will pass the message to the data layer to drastically reduce the sampling probability of this sample.

Intra-Loss and Inter-Loss. In order to enhance the discriminative power of the deeply learnt features, we add intra-loss [10] and inter-loss [10] into ArcFace to minimise the intra-class variations and keep inter-classes distances within the batch. When x_i and x_j are from the same class, their cosine distance should be higher than the threshold θ (*e.g.* 0.3). By contrast, when x_i and x_j are from the different class, their cosine distance should be smaller than the threshold θ (*e.g.* 0.3).

$$L_3 = L_2 + \frac{1}{N^2 - N} \sum_{i,j,i \neq j}^{N} \left(\xi + y_{ij} \left(\theta - {x_i}^T x_j\right)\right)_+, \quad (3)$$

where m is the batch size, $y_{ij} \in \{\pm 1\}$ indicates whether the faces x_i, and x_j are from the same class or not, $(u)_+ := max(u, 0)$ is the hinge loss [11], θ is the threshold to distinguish whether the faces are from the same person or not, and ξ is the error margin besides the classification hyper-plane. In this paper, θ is set as 0.3 and ξ is set as 0.1. As we employ the intra-loss and inter-loss, the sampling strategies changes from global random face images sampling into similar inter-class identities sampling and then diverse intra-class images sampling [10]. When single intra-loss or inter-loss is used, we simply change the comparison between image pairs.

4. Ablation Experiments

4.1. Which training dataset is better?

For face feature embedding, we employ four recent large-scale in-the-wild datasets (*e.g.* CASIA [30], VGG2 [2] MS1MV2 [12] and Asian [1]). In Table 3, we compare the performance of models trained on different datasets. As we can see from the results, combining datasets together can achieve best performance. In addition, we find the BN-FC-BN structure is better than the BN-Dropout-FC-BN structure to get the final 512-D embedding feature. By increasing the capacity of the network (from ResNet100 to ResNet140), we can also improve the performance.

Datasets	#Identity	#Image/Video
CASIA [30]	10K	0.5M
VGGFace2 [2]	9.1K	3.3M
MS1MV2 [12]	85K	5.8M
Asian [1]	94 K	2.83M

Table 2. Face datasets for the training of face feature embedding network.

Methods	1e-05	**1e-04**	1e-03	1e-02	1e-01
VGG2-Res100	18.61	59.62	80.40	90.57	96.43
CASIA-Res100	21.75	65.07	81.02	90.09	95.92
MS1MV2Asian-Res100	18.08	76.23	88.57	93.94	97.39
MS1MV2-Res100	20.99	79.33	89.19	93.73	96.91
All-Res100	21.44	81.92	90.75	94.72	97.35
All-Res100FC	21.39	82.54	91.33	94.99	97.27
All-Res140	20.95	82.83	91.54	95.04	97.51

Table 3. Verification accuracy (%) of our methods under protocol-3 (Overall) of the DFW2019 validation dataset.

4.2. Which ensemble strategy is better?

We select the top 3 models (*e.g.* All-Res100, All-Res100FC and All-Res140)from Table 3 and explore the best ensemble strategies. In Table 4, we consider two ensemble strategies: feature ensemble and score ensemble. For feature ensemble, we concatenate features from different models with weights to construct new features. For instance, feat-cb-equal3 denotes three feature ensemble with equal weights, and feat-cb-532 denotes three feature ensemble with weights of [0.5, 0.3, 0.2]. High weight is assigned to better model (All-Res140). Feat-cb-equal2 denotes two feature ensemble with equal weights and feat-cb-64 denotes two feature ensemble with weights of [0.6, 0.4]. For score ensemble, we use weighted average scores from different models. We use similar abbreviations for the weights. After balancing the performance and the efficiency, we finally combine features from two models (All-Res140 and All-Res100FC) with weights of [0.6, 0.4]. In the following experiments, we call this ensemble feature as the ArcFace feature.

Methods	1e-05	**1e-04**	1e-03	1e-02	1e-01
feat-cb-equal3	21.23	83.29	91.60	95.12	97.49
feat-cb-532	20.96	83.36	91.70	95.12	97.51
feat-cb-433	21.16	83.47	91.65	95.14	97.52
feat-cb-equal2	21.08	83.28	91.66	95.16	97.49
feat-cb-64	21.16	83.33	91.76	95.11	97.52
score-cb-equal3	21.23	83.29	91.60	95.12	97.49
score-cb-532	20.99	83.52	91.72	95.14	97.51
score-cb-433	21.13	83.43	91.62	95.15	97.50
score-cb-equal2	21.08	83.28	91.66	95.16	97.49
score-cb-64	21.12	83.24	91.75	95.14	97.50

Table 4. Verification accuracy (%) of our methods under protocol-3 (Overall) of the DFW2019 validation dataset.

4.3. Could Intra and Inter Loss improve the performance?

As shown in Table 5, both the proposed intra-loss and inter-loss can obviously improve the performance. By combining the intra-loss and inter-loss, our solution finally achieves GAR of 85.54% at 1e-4 FAR.

Methods	1e-05	**1e-04**	1e-03	1e-02	1e-01
ArcFace(R140+R100FC)	21.16	83.33	91.76	95.11	97.52
ArcFace+Intra	19.77	85.94	92.46	95.66	97.86
ArcFace+Inter	23.91	84.52	94.70	97.60	98.55
ArcFace+Intra&Inter	20.52	86.54	94.31	97.93	98.76

Table 5. Verification accuracy (%) of our methods under protocol-3 (Overall) of the DFW2019 validation dataset.

5. DFW2019 Challenge

5.1. DFW2019 Benchmark

The DFW2019 benchmark [13, 24] contains training, validation and testing datasets.

The training and validation datasets include 1,000 identities collected from the Internet. Most of the subjects are adult famous personalities of Caucasian or Indian ethnicity. The training and validation datasets comprise of 11,157 face images including different kinds of images for a given subject, that is, *normal*, *validation*, *disguised*, and *impersonator*. Each subject contains at least 5 and at most 26 face images. In Table 6 we give the statistics of the training and validation datasets. Overall, the training and validation datasets contain 1,000 normal images, 903 validation images, 4,814 disguised images, and 4,440 impersonator images.

For the testing dataset of the DFW2019 benchmark, there are 3840 images of 600 subjects, encompassing different disguise variations including variations due to bridal make-up and plastic surgery. Table 7 presents the statistics of the DFW2019 dataset. The labels of the testing dataset are kept by the organiser for fair comparison in the challenge.

Category	Training Set	Validation Set
Subjects	400	600
Images	3,386	7,771
Normal Images	400	600
Validation Images	308	595
Disguised Images	1,756	3,058
Impersonator Images	922	3,518

Table 6. Statistics of the training and validation sets of the DFW2019 benchmark. Disguised images have same identity as the normal image, while impersonator images have different identity.

Category	Subjects	Images
Bridal	100	200
Plastic Surgery	250	500
Other	250	3140
Total	600	3840

Table 7. Statistics of the testing sets of the DFW2019 benchmark.

5.2. Validation Results

The DFW2019 validation dataset has three protocols for evaluation. All three protocols correspond to face verification protocols, where a face recognition model is expected to classify a pair of face images as genuine (positive) or imposter (negative). Detailed description of each protocol is given below:

Protocol-1 (Impersonation) evaluates a face recognition model for its ability to distinguish impersonators from genuine users with high precision. A genuine (positive) pair for this protocol is a normal image with a validation image of the same subject. For imposter (negative) pairs, the impersonator images of a subject are partnered with the normal, validation, and disguised images of the same subject. In total, there are 595 positive pairs and 24,451 negative pairs.

In Table 8, the performance of each method is reported in terms of Genuine Acceptance Rate (GAR) at 1% and 0.1% False Acceptance Rate (FAR). For the task of impersonation on the validation dataset, the winner of DFW2018, AE-FRL [25], presents a GAR of 96.80% and 57.64% at 1% and 0.1% FARs. By contrast, ArcFace significantly outperforms all other algorithms by achieving 98.66% and 60.84% at 1% and 0.1% FARs, respectively.

Algorithm	**GAR**	
	@1%FAR	**@0.1%FAR**
VGG-Face	52.77	27.05
ByteFace	75.53	55.11
DDRNET	84.20	51.26
DenseNet + COST	92.1	62.2
DR-GAN	65.21	11.93
LearnedSiamese	57.64	27.73
MEDC	91.26	55.46
OcclusionFace	93.44	46.21
UMDNets	94.28	53.27
WVU_CL	81.34	40.00
AEFRL	96.80	57.64
MiRA-Face	95.46	51.09
ArcFace	**98.66**	**60.84**

Table 8. Verification accuracy (%) of ArcFace and the baselines on protocol-1 (impersonation) of the validation dataset. Results of other methods are from [24].

Protocol-2 (Obfuscation) evaluates a face recognition model for its ability to distinguish intentional or unintentional disguises, wherein a person attempts to hide identity. The genuine (positive) pairs are constituted by all pairs generated using the normal and validation images with the disguise images, and the pairs generated between the disguise images of the same subject. The cross-subject imposter (negative) pairs are created by combining the normal, validation, and disguised images of one subject with the normal, validation, and disguised images of a different subject. The impersonator images are not used in this protocol. In total, there are 13,302 positive pairs and 9,027,981 negative pairs.

Table 9 summarises the verification accuracy for all the models, along with the proposed ArcFace. For the task of obfuscation on the validation dataset, the winner of DFW2018, MiRA-Face [31] achieves the best accuracy of 90.65% and 80.56% at 1% and 0.1% FARs. ArcFace obviously outperforms all other algorithms by a clear margin of at least 4.43% for GAR@1%FAR and 11.64% for GAR@0.1%FAR. Compared to the previous protocol (impersonation), the difference in the verification accuracy at the two FARs is relatively less, which indicates that recognition systems suffer less in case of obfuscation, as compared to impersonation at stricter FARs.

Algorithm	**GAR**	
	@1%FAR	**@0.1%FAR**
VGG-Face	31.52	15.72
ByteFace	76.97	21.51
DenseNet + COST	87.1	72.1
DDRNET	71.04	49.28
DisguiseNet	66.32	28.99
DR-GAN	74.56	58.31
LearnedSiamese	37.81	16.95
MEDC	81.25	65.14
OcclusionFace	80.45	66.05
Tessellation	1.23	0.18
UMDNets	86.62	74.69
WVU_CL	78.77	61.82
AEFRL	87.82	77.06
MiRA-Face	90.65	80.56
ArcFace	**95.08**	**92.20**

Table 9. Verification accuracy (%) of ArcFace and the baselines on protocol-2 (obfuscation) of the validation dataset. Results of other methods are from [24].

Protocol-3 (Overall) evaluates a face recognition model for its ability to distinguish disguises as well as impersonators at the same time. The genuine (positive) and imposter (negative) pairs created in the above two protocols are combined to generate the evaluation data for this protocol. In total, there are 13,897 positive pairs and 9,052,432 negative pairs.

Table 10 presents the GAR values of all other baseline methods as well as ArcFace. As with the protocol-2, Arc-

Face significantly outperforms the results of the winner of DFW2018, MiRA-Face, by an obvious margin of 4.49% and 12.5% at 1% and 0.1% FARs.

Algorithm	GAR	
	@1%FAR	@0.1%FAR
VGG-Face	33.76	17.73
ByteFace	75.53	54.16
DenseNet + COST	87.6	71.5
DDRNET	71.43	49.08
DisguiseNet	60.89	23.25
DR-GAN	74.89	57.30
LearnedSiamese	39.73	18.79
MEDC	81.31	63.22
OcclusionFace	80.80	65.34
Tessellation	1.23	0.17
WVU_CL	79.04	60.13
UMDNets	86.75	72.90
AEFRL	87.90	75.54
MiRA-Face	90.62	79.26
ArcFace	**95.11**	**91.76**

Table 10. Verification accuracy (%) of ArcFace and the baselines on protocol-3 (overall) of the validation dataset. Results of other methods are from [24].

ROC of ArcFace. Figure 3 presents the Receiver Operating Characteristic (ROC) curves of ArcFace for all three protocols on the DFW2019 validation dataset. Table 11 presents the GAR values of ArcFace at more challenging FAR (*e.g.* 1e-4). As we can see from Figure 3 and Table 11, ArcFace achieves 83.33% GAR at 1e-4 FAR, which is very impressive. However, GAR is very low under small FAR (in Figure 3), which indicates the dataset is very challenging.

Protocols	1e-5	1e-4	1e-3	1e-2	1e-1
1(Impersonation)	5.71	14.96	60.84	98.66	99.83
2(Obfuscation)	78.42	88.01	92.20	95.08	97.45
3(Overall)	21.16	83.33	91.76	95.11	97.52

Table 11. Verification accuracy (%) of ArcFace on protocol-1 (impersonation), protocol-2 (obfuscation), and protocol-3 (overall) of the DFW2019 validation dataset.

Score Distribution of ArcFace. In Figure 4, we show the cosine distance distribution of negative and positive pairs. The DFW2019 dataset is so challenging that there are a large number of positive pairs with low cosine similarity. In Figure 5, we give some extreme challenging positive and negative pairs, which can not be even distinguished by human.

5.3. Test Results

The DFW2019 test dataset has four protocols for evaluation. All four protocols correspond to face verification protocols, where a face recognition model is expected to classify a pair of face images as genuine (positive) or imposter (negative). Detailed description of each protocol is given below:

Protocol-1 (Impersonation) evaluates a face recognition model for its ability to distinguish impersonators from genuine users with high precision. In Table 12, the performance of our methods is reported in terms of GAR at different FARs. For the task of impersonation on the test dataset, the ArcFace baseline, presents a GAR of 99.20% and 72.40% at 1% and 0.1% FARs. Interestingly, the proposed intra loss obviously decreases the GAR to 56.80% and 17.60% at 0.1% and 0.01%FARs, which indicates that compressing intra variance can has side effect on inter discrepancy. In addition, the proposed inter loss has not improved the GAR compared to the performance of the baseline ArcFace, which indicates impersonators can construct very challenging negative pairs. In conclusion, impersonation is a very challenging problem for current deep face recognition methods.

Methods	1e-4	1e-3	1e-2	1e-1
ArcFace	**44.80**	**72.40**	**99.20**	**99.20**
ArcFace+Intra	17.60	56.80	**99.20**	**99.20**
ArcFace+Inter	**44.80**	**72.40**	**99.20**	**99.20**
ArcFace+Intra&Inter	17.60	56.80	**99.20**	**99.20**

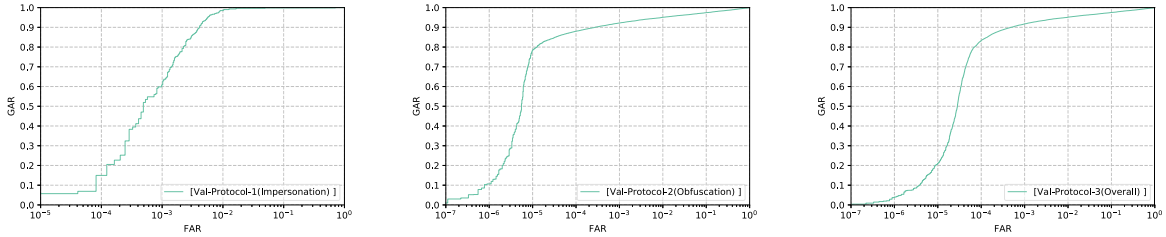
Table 12. Verification accuracy (%) of our methods on protocol-1 (impersonation) of the DFW2019 test dataset.

Protocol-2 (Obfuscation) evaluates a face recognition model for its ability to distinguish intentional or unintentional disguises, wherein a person attempts to hide identity. In Table 13, the performance of our methods is reported in terms of GAR at different FARs. Without impersonator images, the performance of our methods on the Protocol-2 (Obfuscation) are impressive at strict FARs. Both the proposed intra loss and inter loss can significantly improve the results. More specifically, the intra loss improves GAR from 91.43% to 94.48% at 1e-4 FAR, and the inter loss improves GAR to 97.99%. As these two losses are complementary to each other, their combination further improve the GAR to 98.43% at 1e-4 FAR.

Methods	1e-5	1e-4	1e-3	1e-2	1e-1
ArcFace	84.00	91.43	95.73	98.06	99.13
ArcFace+Intra	90.55	94.48	97.03	98.42	99.31
ArcFace+Inter	96.14	97.99	98.73	99.15	99.40
ArcFace+Intra&Inter	**97.68**	**98.43**	**98.92**	**99.31**	**99.47**

Table 13. Verification accuracy (%) of our methods on protocol-2 (obfuscation) of the DFW2019 test dataset.

Protocol-3 (Plastic Surgery) evaluates a face recognition model for its ability to distinguish identities after plastic surgery, wherein a person takes a plastic surgery which can intentionally or unintentionally change the identity. In Table 14, the performance of our methods is reported in terms of GAR at different FARs. Once again, the proposed in-



(a) protocol-1 (impersonation) (b) protocol-2 (obfuscation) (c) protocol-3 (overall)

Figure 3. ROC curves of ArcFace under protocol-1 (impersonation), protocol-2 (obfuscation), and protocol-3 (overall) of the DFW2019 validation dataset.

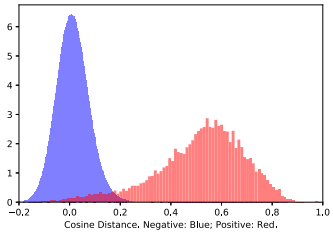


Figure 4. Cosine distance distribution of negative pairs and positive pairs. Cosine distance is predicted by ArcFace under protocol-3 (overall) of the DFW2019 validation dataset. Density is set to true for better visualisation, thus please neglect the y value.

(a) The ground-truth is same person.

(b) The ground-truth is different person.

Figure 5. Extreme challenging pairs (failure cases) for ArcFace under protocol-3 (overall) of the DFW2019 validation dataset.

tra loss presents side effect, and the verification results are slightly decreased compared to the results of ArcFace. By contrast, the proposed inter loss significantly improves the GAR from 87.60% to 95.60% at 1e-4 FAR. After combining the intra and inter loss, the side effect from the intra loss is alleviated by the inter loss. Finally, the combination solution obtains GAR of 95.60% at 1e-4 FAR.

Methods	1e-5	1e-4	1e-3	1e-2	1e-1
ArcFace	72.40	87.60	94.80	98.00	99.60
ArcFace+Intra	75.60	86.40	93.60	97.60	99.60
ArcFace+Inter	**91.60**	**95.60**	**98.40**	**99.60**	**100.00**
ArcFace+Intra&Inter	88.00	**95.60**	**98.40**	**99.60**	**100.00**

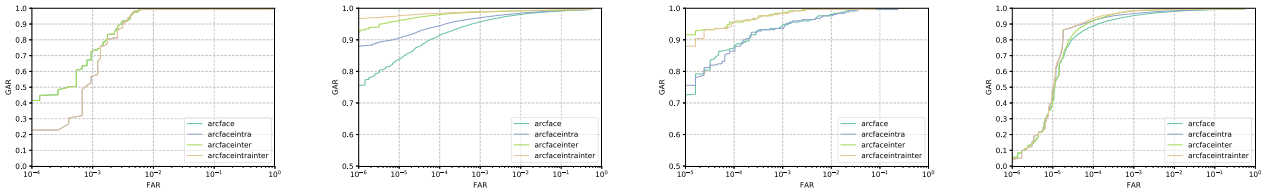
Table 14. Verification accuracy (%) of our methods on protocol-3 (plastic surgery) of the DFW2019 test dataset.

Protocol-4 (Overall) evaluates a face recognition model for its ability to distinguish disguises as well as impersonators at the same time. In Table 15, the performance of our methods is reported in terms of GAR at different FARs. At FAR of 1e-4, ArcFace sets up a strong baseline performance of 88.86%. The proposed intra loss and inter loss separately improve the GAR to 92.19% and 92.00%. When combining these two losses, our solution finally achieves GAR of 93.64% at 1e-4 FAR.

Methods	1e-5	1e-4	1e-3	1e-2	1e-1
ArcFace	41.36	88.86	95.29	98.03	99.15
ArcFace+Intra	47.76	92.19	96.71	98.37	99.31
ArcFace+Inter	41.36	92.00	98.30	99.14	99.41
ArcFace+Intra&Inter	**48.67**	**93.64**	**98.45**	**99.30**	**99.47**

Table 15. Verification accuracy (%) of our methods on protocol-4 (overall) of the DFW2019 test dataset.

ROC of Our Methods. In Figure 6, we give ROC curves of our methods under protocol-1 (impersonation), protocol-2 (obfuscation), protocol-3 (plastic surgery) and protocol-4 (overall) of the DFW2019 test dataset. Based on our observations, we have following conclusions: (1) ArcFace is a very general and excellent face recognition algorithm which can enhance intra-class compactness and inter-class discrepancy to some extend. (2) Exploring challenging negative pairs (*e.g.* the proposed inter loss) to improve deep face



(a) protocol-1 (impersonation) (b) protocol-2 (obfuscation) (c) protocol-3 (plastic surgery) (d) protocol-4 (overall)

Figure 6. ROC curves of our methods under protocol-1 (impersonation), protocol-2 (obfuscation), protocol-3 (plastic surgery) and protocol-4 (overall) of the DFW2019 test dataset.

Method	Impersonation			Obfuscation		Plastic Surgery		Overall	
FAR	$1e-4$	$1e-3$	$1e-2$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e-3$
ResNet-50 [23]	38.40	47.60	-	16.42	35.38	22.40	46.40	16.89	35.96
LightCNN-29v2 [23]	**51.20**	**74.40**	-	36.90	55.56	47.20	69.20	36.50	55.74
ArcFace	44.80	**72.4**	**99.2**	91.4	95.7	87.6	94.8	88.6	95.2
ArcFace+Intra	17.60	56.8	99.2	94.4	97.0	86.4	93.6	92.1	96.7
ArcFace+Inter	44.80	72.4	99.2	97.9	98.7	95.6	98.4	92.0	98.3
ArcFace+Intra&Inter	17.60	56.8	**99.2**	**98.4**	**98.9**	**95.6**	**98.4**	**93.6**	**98.4**

Table 16. Verification accuracy (%) of our methods under protocol-1 (impersonation), protocol-2 (obfuscation), protocol-3 (plastic surgery) and protocol-4 (overall) of the DFW2019 test dataset.

recognition is one of the general and effective approaches. (3) Exploring challenging positive pairs (*e.g.* the proposed intra loss) can be tricky and unstable. Enhancing intra-class compactness on the face images with extremely large appearance variations can affect the manifold optimisation. (4) Disguised face recognition is still a very challenging problem in the field of deep face recognition.

Official Evaluation of Our Methods. In Table 16, we give the verification results reported by the organisers. Under four protocols of the DFW2019 test dataset, the organisers are more interested in GAR at 1e-3 and 1e-4 FARs. On the Impersonation track, our solution is worse than the baseline method (LightCNN-29v2 [23]) provided by the organiser. On other tracks, our solution achieves significant better results than the baseline methods.

6. Conclusions

Disguised face recognition features high intra-class diversity and inter-class similarity. In this report, we give the technical details of our submission to the DFW2019 challenge. By using our RetinaFace for face detection and alignment and ArcFace for face feature embedding, we achieve state-of-the-art performance on the DFW2019 challenge.

Acknowledgements. Jiankang Deng acknowledges financial support from the Imperial President's PhD Scholarship and GPU donations from NVIDIA. Stefanos Zafeiriou acknowledges support from EPSRC Fellowship DEFORM (EP/S010203/1), FACER2VM (EP/N007743/1) and a Google Faculty Fellowship.

References

[1] http://trillionpairs.deepglint.com/overview. 3, 4

[2] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018. 3, 4

[3] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, 2018. 1

[4] J. Deng, J. Guo, N. Xue, and S. 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. 1, 2

[5] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019. 1, 2

[6] J. Deng, Q. Liu, J. Yang, and D. Tao. M3 csr: Multi-view, multi-scale and multi-component cascade shape regression. *Image and Vision Computing*, 47:19–26, 2016. 1

[7] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *International Journal of Computer Vision*, 127(6-7):599–624, 2019. 1

[8] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou. Joint multi-view face alignment in the wild. *IEEE Transactions on Image Processing*, 28(7):3636–3648, 2019. 1

[9] J. Deng, Y. Zhou, S. Cheng, and S. Zaferiou. Cascade multi-view hourglass model for robust 3d face alignment. In *2018 13th IEEE International Conference on Automatic*

Face & Gesture Recognition (FG 2018), pages 399–403. IEEE, 2018. 1

[10] J. Deng, Y. Zhou, and S. Zafeiriou. Marginal loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 60–68, 2017. 3

[11] C. Gentile and M. K. Warmuth. Linear hinge loss and average margin. In *Advances in neural information processing systems*, volume 11, pages 225–231, 1998. 3

[12] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 3, 4

[13] V. Kushwaha, M. Singh, R. Singh, M. Vatsa, N. Ratha, and R. Chellappa. Disguised faces in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–9, 2018. 1, 2, 3, 4

[14] H. Liu, X. Zhu, Z. Lei, and S. Z. Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[15] Q. Liu, J. Deng, and D. Tao. Dual sparse constrained cascade regression for robust face alignment. *IEEE Transactions on Image Processing*, 25(2):700–712, 2015. 1

[16] Q. Liu, J. Deng, J. Yang, G. Liu, and D. Tao. Adaptive cascade regression model for robust face alignment. *IEEE Transactions on Image Processing*, 26(2):797–807, 2016. 1

[17] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. 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

[18] Y. Liu, H. Li, J. Yan, F. Wei, X. Wang, and X. Tang. Recurrent scale approximation for object detection in cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 571–579, 2017. 3

[19] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu. Deep tree learning for zero-shot face anti-spoofing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[20] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and 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*, pages 51–59, 2017. 1

[21] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 17–24. IEEE, 2017. 3

[22] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1, 2

[23] M. Singh, M. Chawla, R. Singh, M. Vatsa, and R. Chellappa. Disguised faces in the wild 2019. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019. 8

[24] M. Singh, R. Singh, M. Vatsa, N. K. Ratha, and R. Chellappa. Recognizing disguised faces in the wild. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(2):97–108, 2019. 1, 2, 3, 4, 5, 6

[25] E. Smirnov, A. Melnikov, A. Oleinik, E. Ivanova, I. Kalinovskiy, and E. Luckyanets. Hard example mining with auxiliary embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 37–46, 2018. 5

[26] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018. 2

[27] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 2

[28] J. Yang, J. Deng, K. Zhang, and Q. Liu. Facial shape tracking via spatio-temporal cascade shape regression. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 41–49, 2015. 1

[29] X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li, and W. Liu. Face anti-spoofing: Model matters, so does data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[30] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014. 3, 4

[31] K. Zhang, Y.-L. Chang, and W. Hsu. Deep disguised faces recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 32–36, 2018. 5

[32] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. 3

[33] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li. A dataset and benchmark for large-scale multi-modal face anti-spoofing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1