Feature Ensemble Networks with Re-ranking for Recognizing Disguised Faces in the Wild

Arulkumar Subramaniam*[1], Ajay Narayanan Sridhar*[2], Anurag Mittal[1]
[1]Indian Institute of Technology Madras, India,
[2]Indian Institute of Information Technology Design & Manufacturing Kancheepuram, India.
aruls@cse.iitm.ac.in, narayananajay99@gmail.com, amittal@cse.iitm.ac.in

Abstract

*Recognizing a person's face images with intentional/unintentional disguising effects such as make-up, plastic surgery, artificial wearables (hats, eye-glasses) is a challenging task. We propose a **F**eature **E**nsem**B**le **Net**work (FEBNet) for recognizing Disguised Faces in the Wild (DFW). FEBNet encompasses multiple base networks (SE-ResNet50, Inception-ResNet-V1) pretrained on large-scale face recognition datasets (MS-Celeb-1M, VGGFace2) and fine-tuned on DFW training dataset. During the fine-tuning phase, we propose to use two novel objective functions, namely, 1) Category loss, 2) Impersonator Triplet loss along with two prevalent objective functions: Identity loss, Inter-person Triplet loss. To further improve the performance, we apply a state-of-the-art re-ranking strategy as a post-processing step. Extensive ablation studies and evaluation results show that FEBNet significantly outperforms the baseline models.*

1. Introduction

Face recognition is an important and challenging biometry-aligned computer vision task that deals with matching person's faces[47, 17]. The task has attracted notable attention from the research community due to its wide range of applications in surveillance, robotics, access control, human-computer interaction and so on. The task involves significant challenges such as unconstrained illumination, scale, view-point and pose variations, image distortions, wide range of intentional and/or unintentional disguising effects[33, 32] on faces.

With the advent of deep learning, the performance of face recognition algorithms [26, 37, 39, 30] has surpassed human level performance in several benchmark datasets. However, such state-of-the-art algorithms do not perform well in scenarios involving complex variations such as illumination changes[22, 10, 35], disguising make-ups and impersonator faces[45, 2]. The disguising make-ups include use of tattoos, plastic surgery, artificial wearables such as hats, eye-glasses, nose rings, earrings, scarfs and so forth[33, 32]. Impersonator detection and recognition is an important task considering the ever-increasing demand for security and surveillance needs. Impersonator refers to a person who has a similar looking face as that of an iden-

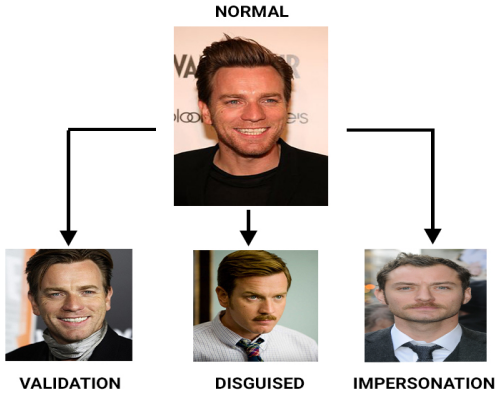

Figure 1. Illustration of disguised face recognition task. The representative images are taken from DFW-2018 dataset[33].

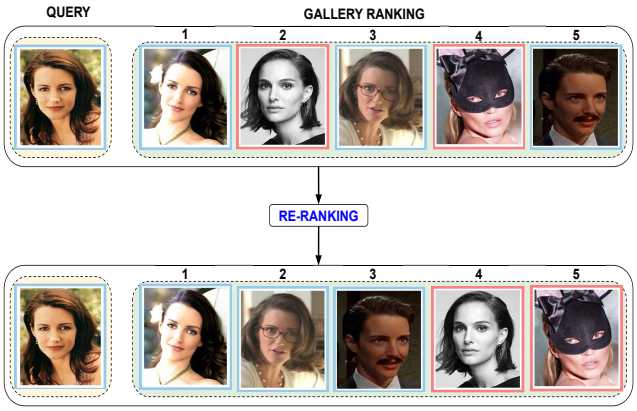

Figure 2. Application of re-ranking technique(s) [49] to the disguised face recognition task. The representative images are taken from DFW-2018 dataset[33]. Subject images with red colour are different from query's subject.

*Equal contribution

tity's face images (Fig. 1).

Most benchmark datasets for face recognition[12, 16, 4] are created based on scenes that are well-lit, with complete visible faces and with no disguising effects or wearables on faces. The face recognition systems trained with these datasets perform well on the face images that possess similar characteristics. However, in practical scenarios, the face images possess variety of illumination changes and disguising effects intentionally or unintentionally by people that makes the face recognition task harder and pretrained models to be ineffective. Further, the datasets collected especially to focus on the specific challenges such as illumination changes[11] or disguising make-up[33, 32] are considerably small in number of identities or images. To be able to perform well on such datasets, one of the widely used approach is to transfer the learned knowledge from large datasets[12, 16, 4] and further, fine-tune in the small dataset(s)[33, 32, 31, 11]. In a similar way, we explore the face recognition task focusing on disguised face images by first evaluating several pretrained models[46, 9] trained on several existing large benchmark datasets[12, 4] and show that they exhibit reasonable performance owing to the learned knowledge from large-scale datasets. In the next step, we fine-tune the base models on the Disguised Faces in the Wild-2018 (DFW-2018) training set and evaluate the results. During fine-tuning, we formulate two novel loss/objective functions namely, 1) Impersonator Triplet loss, 2) Category loss to systematically improve the classification performance of the identities and impersonators. Further, we formulate a **F**eature **E**nsem**B**le based **Net**work (FEBNet) which consists of multiple fine-tuned base model architectures to reduce variance of individual base models and promote discrimination among the identities.

A widely applicable post-processing technique in retrieval problems, such as person re-identification[48, 41], is to use the neighborhood information to improve the performance. For instance, k-reciprocal re-ranking[49] method uses mutual k-nearest neighbors followed by query expansion step to calculate the jaccard distance reflecting the neighborhood similarities. It has been proved by empirical studies that the re-ranking strategy gives superior performance[49, 34] and application of such post-processing could boost the performance in face recognition task as well (Fig. 2). However, the application of such post-processing methods is relatively unexplored in the face recognition task. To bridge this gap, in our work, we explore the usage of k-reciprocal re-ranking[49] as a post processing method and benchmark the performance improvements.

Our contributions are as follows:

1. We propose a **F**eature **E**nsem**B**le **Net**work (FEBNet) as an ensemble of multiple state-of-the-art face recognition networks for the problem of recognizing disguised faces in the wild.

2. We propose two novel loss functions namely 1) Impersonator Triplet loss, 2) Category loss to improve the performance in challenging impersonator recognition scenario.

3. We explore the usage of re-ranking strategy in the application of disguised face recognition.

4. We perform extensive ablation studies of the models and proposed objective functions to evaluate the importance of the individual constituting components.

2. Related works

Typical face recognition solutions consist of a three step approach namely: 1) Face and/or facial landmark points detection, 2) Face alignment and 3) Recognition. Our work focuses on improving recognition phase for the task of disguised face recognition. In this section, we briefly outline the related literature for the task.

Face recognition: Earlier approaches used hand-crafted features such as Local Binary Patterns (LBP)[1], Histogram of Oriented Gradients (HoG)[6], SIFT[21], SURF[25] followed by a suitable metric (learned or predefined) to compare the features. The recent works predominantly use deep learning architectures owing to their ability to learn features and metrics automatically end-to-end from the data. FaceNet[30] used ZF-Net[43] and GoogleNet[36] architectures with Triplet loss as objective function to optimize the model. The final descriptors are compared using L2 distance during test. DeepFace[37] used a 9-layer architecture with softmax cross-entropy loss and siamese loss for optimization. The face embeddings at the test time are compared using Chi-square distance. Recent approaches concentrate on improving the objective functions as well as exploring data augmentation strategies for better generalization. For instance, ArcFace[7] added an angular margin loss in phase domain of descriptors to promote generalization. Further, CosFace[40] used the cosine similarity metric with margin to improve the performance. Though a super-human level performance is achieved in face recognition[37, 30], the challenges remain in case of recognizing faces with illumination variations, disguised faces, etc. Our work focuses on improving performance in recognizing disguised faces in the wild.

Disguised face recognition: In a very few works in the literature focusing on disguised face recognition, pretrained models trained on large benchmark datasets are used to transfer the knowledge. Specifically, MiRA-Face[45] uses a combination of two CNNs for performing disguised face recognition by treating the aligned and unaligned images separately. A dimension reduction step based on Principal

Components Analysis (PCA) is carried out on the learned features before evaluation. Another similar work [2] performs feature extraction from the aligned faces[28] using two pretrained networks. Independent scores are calculated individually from the features of the models and the final score is obtained by averaging the individual scores. Our work follows a similar paradigm of transfer learning by utilizing pretrained networks, but differs by the addition of two novel loss functions and a re-ranking post-processing step.

Re-ranking and it's applications Successful application of re-ranking methods as a post-processing step can be noticed in several retrieval tasks. Re-ranking methods typically follow the philosophy of "Tell me who your friends are and I'll tell you who you are", i.e., they exploit the neighborhood information among the query and gallery instances as well as inter-gallery instances to improve the performance. K-nearest neighborhood method is prevalently used to capture the neighborhood information. Ondrej *et al*. [5] showed that using the average embedding of k-nearest neighbors to re-query the database improves performance. K-reciprocal nearest neighbor approach is introduced in [27] that showed the effectiveness of mutual k-nearest neighbors. Zhong *et al*. [49] used k-reciprocal nearest neighbors to determine a neighborhood based jaccard distance between query and gallery instances. The final distance is calculated as a combination of the jaccard distance and the original query-gallery distance. Application of this method to person re-identification task showed remarkable improvements[49, 34]. However, in face recognition such re-ranking methods are relatively unexplored. In our work, we demonstrate the effectiveness of [49] in the task of recognizing disguised faces.

3. Proposed pipeline

In this work, we explore the methods to transfer the knowledge from publicly available pretrained face recognition models[46, 9] to the task of recognizing disguised faces. The pipeline of the proposed method is shown in the Figure 3.

3.1. Pre-processing

In the pre-processing stage (Fig. 3), the face images are aligned using one of the two (dlib[18], MTCNN[44]) landmark detection and alignment routine depending on the base models.

dlib[18] face alignment In this module, we make use of the latest Dlib package's ResNet34[13] to predict the facial landmark location. Compared to the combination of HoG and Linear SVM, the deep model based landmark detection is more accurate and reasonably faster. The pixel values from the predicted landmark locations are mapped to canonical position to align the face image.

MTCNN[44] face alignment Multi-Task Cascaded Convolution Neural Networks is a joint face detection and alignment module that utilizes the inherent correlation between these tasks to improve the performance. The network consists of 3 cascaded stages that perform coarse-to-fine prediction of face and landmark location in real-time.

3.2. Base model architectures

In the proposed architecture, three pretrained base models are used in combination to tackle the disguised face recognition task. The individual models are explained below:

IR50 [46] This model is an extension of SE-ResNet50 [15] model trained on MS-Celeb-1M [12] dataset with objective function as Arc loss[7] and Focal loss[20]. MS-Celeb-1M [12] dataset contains 100K celebrity identities and around ∼5M images in total. During pre-training, MTCNN's face detection and alignment method is used to detect and crop the face image to size 112×112. Further, while fine-tuning on training set (refer to Section 4.1), we use two instances of IR50, one with Dlib's[18] face alignment, the other with MTCNN's[44] face alignment to have complementary alignment methods (referred as IR50_D and IR50_M respectively from now on).

FaceNet-Incep-ResNet-v1 In this architecture (referred as FaceNet from now on), the Inception[36] model with residual connections is pretrained with the dataset "VGGFace2" using person classification loss (cross-entropy). VGGFace2[4] is a large dataset consisting of 8631 identities and $\sim$ 3.08M images in total. During pre-training, the MTCNN's face detection and alignment method is used to detect and crop the face image to size 160×160.

3.3. Objective functions

The IR50 and FaceNet models are pretrained using MS-Celeb-1M dataset[12] and VGGFace2[4] respectively with the aid of Arc loss[7] and Focal loss[20]. Further, the pretrained base models are fine-tuned using training dataset with the aid of four objective functions as follows:

Identity Loss (L_{id}) The cross-entropy loss is used to calculate the loss between the softmax probabilty output p_i from the model and the target identity. i.e., Given a person's face image I_i and the target identity t_i as an one-hot vector, the Identity loss is defined as,

$$L_{id} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} t_{ij} \log p_{ij} \tag{1}$$

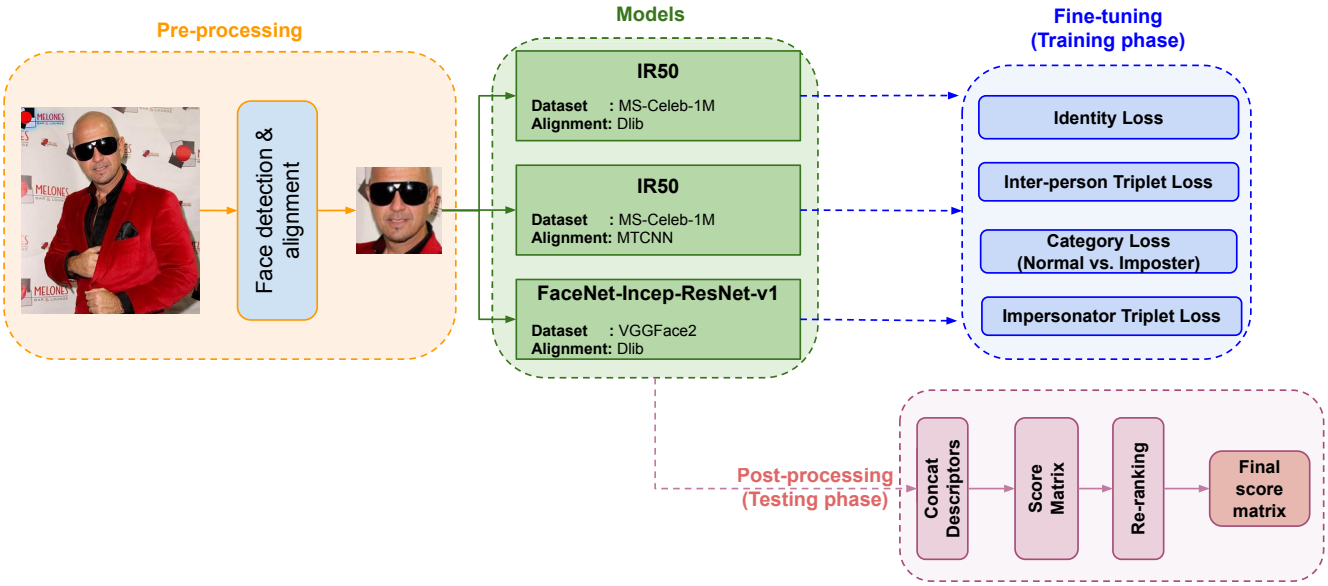

Figure 3. Illustration of our FEBNet model pipeline. The **pre-processing** step consists of face detection and alignment. In **training phase**, **the base models** are fine-tuned on DFW-2018 training set using the objective functions explained in Section 3.3. In **the testing phase**, for each test image, the (L2 normalized) descriptors are extracted from the base models and concatenated to get the final descriptor. Further, the re-ranking[49] method is applied on the score matrix to get the final score matrix.

Here, N = number of face images in the mini-batch, M = number of identities in train-set.

Inter-person Triplet Loss (L_{trip}) To promote small intra-class distance and high inter-class distance, the triplet loss is applied on the face image embeddings. For each face image I_i in the mini-batch, hard positive I_{i+} and hard negative I_{i-} instances are mined[14] within the mini-batch. The triplet loss is calculated by:

$$L_{trip} = \frac{1}{N}\sum_{i=1}^{N} max(0, d(I_i, I_{i+}) - d(I_i, I_{i-}) + m) \quad (2)$$

Here m = margin parameter, $d(i, j)$ is the metric distance between embeddings i & j (In this paper, Euclidean distance).

Category Loss (L_{cat}) To discriminate the impersonator images of the identities, we employ a binary classification loss to classify the face images into two classes namely 1) Normal-validation-disguise class, 2) Impersonator class. The binary classification loss is formulated as a cross-entropy function between the predicted category probability p and the one-hot encoding of target class y, as follows:

$$L_{cat} = -y \log p - (1 - y) \log(1 - p) \quad (3)$$

Impersonator Triplet Loss (L_{imp}) Similar to Inter-person Triplet Loss (L_{trip}), we also employ an impersonator triplet loss to distinguish a particular identity from it's impersonator. For this purpose, each face image's embedding I_i in the mini-batch is considered as 'anchor', a random instance of same person in the mini-batch is selected as 'positive' (I_{i+}) and impersonator images are considered as 'negative' (I_{imp}). L_{imp} is defined as:

$$L_{imp} = \frac{1}{N}\sum_{i=1}^{N} max(0, d(I_i, I_{i+}) - d(I_i, I_{imp}) + m) \quad (4)$$

Here m = margin parameter, $d(i, j)$ is the metric distance between embeddings i & j (In this paper, Euclidean distance).

The overall objective function/total loss is given by:

$$L = \gamma_1 L_{id} + \gamma_2 L_{trip} + \gamma_3 L_{imp} + \gamma_4 L_{cat} \quad (5)$$

The ratios $\gamma_1 = 1.0, \gamma_2 = 0.5, \gamma_3 = 0.1, \gamma_4 = 0.01$ are selected using validation set (refer to Section 4.1).

3.4. Post-processing, Testing

The L2-normalized feature vectors are extracted from the base models independently and concatenated to get the final feature descriptor. The final descriptors are L2-normalized again and score matrix (S_{fused}) is calculated by considering Euclidean distance between feature descriptors. Further, Re-ranking[49] is applied on S_{fused} to get the re-ranked score matrix.

3.4.1 Re-ranking for disguised face recognition

Re-ranking methods are prevalent in retrieval solutions and are proved to improve the retrieval performance[49, 34].

Albeit the significant performance improvements in other vision tasks, the performance gain of such re-ranking methods in the context of face recognition is less explored and unknown. In our work, we introduce the usage of re-ranking method, specifically, k-reciprocal re-ranking method [49] to the disguised face recognition task and evaluate the performance in the validation dataset. The hyper-parameters for the re-ranking method are empirically selected based on the validation set. The ablation studies on these hyper-parameters are shown in Table 5.

In the setup for re-ranking, the "query" images (Q) are to be matched with a set of "gallery" (G) images already in the database. The k-reciprocal re-ranking strategy consists of the following steps, as proposed and explored in [49]:

1. k-reciprocal nearest neighbors: In the first step, for each query q_i, the k-nearest gallery neighbors (k_1) i.e., $g_1, g_2 \dots g_{k_1}$ are refined to get k-reciprocal nearest neighbors list. Specifically, for each query, from the k-nearest neighbor list, the candidates for which the query is also a k-nearest neighbor are filtered. Such candidates are classified as k-reciprocal nearest neighbors.

2. k-reciprocal nearest neighbors expansion: In the second step, the k-reciprocal neighbor list of each query q_i is expanded further with k-reciprocal neighbors of its gallery candidates if both query and the gallery candidate has significant neighbors (k_2) in common.

3. Jaccard distance (D_{jac}) calculation: Given that the final k-reciprocal nearest neighbors are obtained for each of the query q_i, the neighborhood information of the query is encoded as a $d-$dimension feature vector where d = number of gallery/candidate images. Further, the jaccard distance is calculated by correlating the encoded feature vectors of query and gallery.

4. Distance fusion: The original distance matrix D_{orig} and Jaccard distance matrix (D_{jac}) are linearly combined (in terms of a ratio λ) to arrive at the final distance matrix (D_{final}). D_{final} is used further to rank the gallery images for each query to improve the retrieval results.

The hyper-parameters for re-ranking ($k_1 = 24, k_2 = 6, \lambda = 0.6$) are chosen by using validation dataset.

4. Experiments

4.1. Datasets, Implementation, Training

DFW-2018 Dataset[33]: DFW-2018 dataset consists of ∼11,000 images of 1000 subjects collected from the internet. Out of 1000 subjects, 400 subjects are selected as the **training set** and the remaining 600 subjects constitute the **validation set**. Each subject contains images of three types, 1) **Normal/Validation:** Non-disguised frontal face image of a subject, 2) **Disguised:** Face images of a subject having intentional or unintentional disguise, 3) **Impersonator:** Face images of an impersonator for a subject. (An image of any other person, intentionally or unintentionally, pretending to be the subject's identity)

DFW-2019 Dataset[32]: DFW-2019 dataset contains over 3800 images of 600 subjects, encompassing different disguise variations including variations due to bridal make-up and plastic surgery. DFW-2019 dataset serves as the **testing dataset**.

Training: The pre-trained base models are fine-tuned using DFW-2018 training dataset[33]. We use Stochastic mini-batch Gradient Descent (SGD) to optimize the model for 120 epochs with the hyper-parameters as follows: batch size = 120, momentum = 0.9, weight decay = 5e-4, margin for Inter-person triplet loss L_{trip} and Impersonator triplet loss L_{imp} = 0.6, initial learning rate = 0.01. The learning rate is decayed using a cosine annealing scheduler. The model with best validation accuracy is chosen to be the final model for testing.

During fine-tuning, we select mini-batches at each iteration in such a way that each distinct subject from a mini-batch has samples from all the three categories (Normal/Validation, Disguise, Impersonator). Also, we freeze the weights of all layers except the last fully-connected layer and each base model results in a feature dimension of 512.

4.2. Evaluation protocol

We follow the evaluation protocols from [33, 32] for disguised face recognition as illustrated in Table 1. Due to the inherent ambiguity of the 2D face images, it is very challenging to discriminate impersonated face images (protocol 1). The obfuscation and plastic surgery protocol are relatively easier as it is discriminating different identity's images.

We report the results of protocols 1, 2, 4 for the validation dataset and all the protocols in the testing dataset. We encourage the interested readers to refer to [33, 32] for more details. All the quantification numbers of the models are given by Genuine Acceptance Rate (GAR) @x% False Acceptance Rate (FAR), where $x \in \{1, 0.1, 0.01\}$.

4.3. Ablation studies

4.3.1 Performance of base models before fine-tuning

We report the performances of individual base models before fine-tuning in Table 2. Owing to the training on relatively large datasets[12, 4], the individual base models exhibit considerably good performance without fine-tuning. For example, FaceNet[36] model pretrained

Protocol	Usage	Genuine pairs	Imposter pairs
1. Impersonation	To evaluate the algorithm's performance under the impersonation case. i.e., discriminating the original identity's face from the person's face looking very similar to the identity	The combination of 'normal' and 'validation' face images of same identity	The combination of 'normal, validation , disguised' images with its own 'impersonator' images
2. Obfuscation	To correctly identify the person whose face has intentional/unintentional make-ups, wearable items, etc.,	The combination of (normal, disguise), (validation, disguise), and (disguise, disguise) images of the subject	The combination of 'normal, validation, disguised' images of one subject with 'normal, validation, disguised' images of another subject
3. Plastic surgery	To evaluate the algorithm to validate the faces with plastic surgery	The combination of (normal, disguise_p) and (validation, disguise_p) images of the subject. Here, disguise_p = plastic surgery images	The combination of 'normal, validation, disguised_p' images of one subject with 'normal, validation, disguised_p' images of another subject
4. Overall	To evaluate a given algorithm on the entire dataset	Super set of genuine pairs from Impersonation, Obfuscation, and Plastic surgery protocols	Super set of imposter pairs from Impersonation, Obfuscation and Plastic surgery protocols

Table 1. Evaluation protocols for recognizing disguised faces

on VGGFace2[4] dataset gives 79.83%, 72.48%, 72.61% GAR@1%FAR on protocols 1, 2, 4 respectively. The IR50 model[46] pretrained on MS-Celeb-1M dataset[12] + Dlib face alignment[18] outperforms FaceNet model by 17%, 8%, 8% GAR@1%FAR respectively on protocols 1, 2, 4 owing to the high capacity and generalizing ability of the ResNet model[13] over Inception model [36]. Surprisingly, IR50_M model gives much lower GAR@0.1%FAR on Protocol 1. It could be due to the accurate face alignment of MTCNN[44] that the feature extraction model is unable to distinguish between the impersonator and the real identity's face images.

Models	GAR					
	@1%FAR			@0.1%FAR		
	Protocol			Protocol		
	1	2	4	1	2	4
IR50_D ([46] + [18])	96.47	80.42	80.73	44.70	70.32	69.85
IR50_M ([46] + [44])	67.58	79.22	81.27	04.83	72.62	70.61
FaceNet[9, 30]	79.83	72.48	72.61	45.04	50.15	49.17

Table 2. Performance of base models without fine-tuning on training dataset

4.3.2 Performance of base models after fine-tuning

Transfer learning[3, 42] has been proved to be an effective way of reusing the knowledge gained from other tasks/datasets to the task/dataset at hand. In this spirit, to reuse the knowledge gained from large-scale face recognition task[46, 9, 30] to the disguised face recognition task, the pretrained base models are fine-tuned in the DFW-2018 training dataset and the results are illustrated in Table 3. During fine-tuning, we adopt the objective functions as mentioned in Section 3.3. The fine-tuning step lets the model familiarize with the training dataset at hand and align with the distribution of data of this particular task. Hence, the model's performance is typically increased after fine-tuning, as observed in many other computer vision tasks[24, 29]. For example, FaceNet model's GAR @1% FAR has increased by 0.5%, 1.32%, 1.76% in absolute scale on protocols 1, 2, 4 as noticed from Tables 2 and 3. Similarly, Protocol 2 and 4's GAR@1% FAR of IR50_D model is improved by 3% each, IR50_M model by 7% and 5% respectively. Out of all the protocols (1, 2 & 4), significant improvements are noticed for protocols 2 and 4.

Architecture			GAR					
			@1%FAR			@0.1%FAR		
			Protocol			Protocol		
IR50_D	IR50_M	FaceNet	1	2	4	1	2	4
		✓	80.33	73.80	74.37	45.37	52.57	51.87
	✓		66.38	81.81	82.27	05.71	73.87	72.97
	✓	✓	91.93	83.11	83.50	52.77	71.86	70.07
✓			93.94	83.16	83.37	48.40	70.12	69.05
✓		✓	93.61	84.30	84.44	53.10	71.24	69.66
✓	✓		94.62	85.42	85.56	53.44	75.07	73.72
✓	✓	✓	**95.79**	**86.19**	**86.25**	**56.30**	**75.25**	**73.42**

Table 3. Performance of various configurations of ensemble architectures. Here "IR50_D" denotes pretrained IR50 model fine-tuned with Dlib-aligned face images of DFW-2018 train set, "IR50_M" denotes pretrained IR50 model fine-tuned with MTCNN-aligned face images of DFW-2018 training dataset, "FaceNet" denotes pretrained FaceNet-Inception-ResNet-v1 network fine-tuned with Dlib-aligned face images of DFW-2018 training dataset.

4.3.3 Performance of ensembles of fine-tuned base models

In terms of bias-variance trade-off, organizing several models in an ensemble is an effective way to reduce the variance

of models [23]. In many cases, ensemble models are proved to be empirically superior than the individual models in several computer vision[38, 19] and speech recognition tasks [8]. Following similar footsteps, we explore to create ensembles by combining different fine-tuned base models and depict the results in Table 3. Ensemble models are created by concatenating feature descriptors from the constituting fine-tuned base models. In the disguised face recognition task, we observe that in majority of cases, the ensemble models perform on-par or superior to the individual constituent fine-tuned base models. First, we analyze the performances of ensemble models consisting of two base models, then further extend the analysis to the ensemble models containing all the three base models.

Ensemble of FaceNet model with other models: Combining FaceNet-Incep-ResNet-v1 (FaceNet) model with other models either performs on-par or increases the model's performance. Specifically, combining IR50_M (IR50 model with MTCNN face alignment) and FaceNet model increases the GAR@1% FAR up to 10% and GAR@0.1% FAR up to 7% on Protocol 1. Further, combining IR50_D (IR50 model with Dlib face alignment) and FaceNet model improves the GAR@1% FAR of Protocol 2, 4 by $\sim 1\%$ and GAR@0.1% FAR of Protocol 1 by 4.7% than the individual model's performance. As a result, we observe that FaceNet model's inclusion in the ensemble provides a positive boost to the overall performance.

Ensemble of IR50 models: The IR50_D model exhibits the highest performance compared to other models. We observe a degraded performance of IR50_M model in GAR@0.1% FAR of protocol 1, as similar to the non fine-tuned model's performance in Section 4.3.1. Regardless of such degraded performance, the ensemble of IR50_D and IR50_M performs superior than all of the ensemble models consisting of two individual models. We attribute this performance increase to the compatibility of the feature space of both IR50 models and increase in confidence score towards the retrieval performance.

Ensemble model consisting of three base models: Similar to the ensemble models with two base models, we further extend the analysis to ensembles with three base models. In the ensemble models with three base models, the performance remains on-par or superior to the overall highest performing model. We consider the best performing model (IR50_M + IR50_D + FaceNet) as our final model and name it as "FEBNet" that stands for "**F**eature **E**nsem**B**le **Net**work". In the following sections, we perform further ablation studies on "FEBNet" to demonstrate the influence of the proposed loss functions and re-ranking strategy.

4.3.4 Analysis of objective functions

In this section, we evaluate the effectiveness of newly proposed loss functions: Category loss and Impersonator triplet objective functions (refer to Section 3.3). We perform the experiments on our final model and illustrate the results in Table 4. In our empirical study, the person identity loss (L_{id}) and Inter-person triplet (L_{trip}) loss are kept constant and given more importance ($\gamma_1 = 1.0, \gamma_2 = 0.5$). The losses L_{id}, L_{trip} guide the stability of the training by driving the features to be discriminative and robust enough to classify the persons and to be able to promote lower intra-class distance as well as higher inter-class distance. Further in the experiments, we add the proposed losses of Category loss and Impersonator triplet loss to analyze the performance improvements.

We observe from Table 4 that the inclusion of Impersonator triplet loss increases GAR @1% FAR of Protocol 1 by 0.3% in absolute value and inclusion of category loss improves GAR @0.1% FAR of Protocol 1 by 0.6%. The inclusion of both of the losses improve the GAR@1% FAR of protocol 1 by 0.3% and GAR@0.1% FAR of protocol 1 by 1.35%. By improving the much harder GAR@0.1% FAR performance, It is evident that the proposed losses improve the performance significantly.

Losses		GAR					
		@1%FAR			@0.1%FAR		
		Protocol			Protocol		
L_{cat}	L_{imp}	1	2	4	1	2	4
		95.46	86.22	**86.42**	54.95	75.10	73.33
	✓	95.79	**86.37**	86.34	54.11	75.13	73.37
✓		95.12	86.31	86.39	55.63	75.16	73.29
✓	✓	**95.79**	86.19	86.25	**56.30**	**75.25**	**73.42**

Table 4. Performance comparison of various configurations of ensemble architectures with the proposed objective functions: Impersonator Triplet loss (L_{imp}), Category loss (L_{cat})

4.4. Application of re-ranking on face recognition

The re-ranking method described in Section 3.4.1 is dependent on the hyper-parameters k_1, k_2 and λ. We conduct empirical studies on validation dataset[33] with our final model to determine these hyper-parameters and show the quantitative observations in Table 5.

By comparing the performance of FEBNet from Table 3 (last row) with the re-ranking performances in Table 5, we observe that the application of re-ranking helps protocols 2 and 4 significantly while giving a slight improvement or on-par results in protocol 1. We fix the hyper-parameters of re-ranking to be $k_1 = 24, k_2 = 6, \lambda = 0.6$ based on the overall improvement in performance as shown in Table 5 and apply the re-ranking procedure during test set evaluation. Comparing our final model "FEBNet with re-ranking" to the state-of-the-art architectures (MiRA-Face[45] and UMDNets[2]) from DFW-2018 challenge[33] in Table 5, we observe that FEBNet outperforms them in the challenging measure of GAR@0.1% FAR and performs on-par in

Hyper-parameters			GAR					
			@1% FAR			@0.1% FAR		
k_1	k_2	λ	Protocol			Protocol		
			1	2	4	1	2	4
23	5	0.6	95.46	88.64	88.69	56.97	83.88	82.57
		0.7	96.30	88.35	88.49	57.14	82.88	81.68
	6	0.6	95.29	88.74	88.83	54.11	84.13	**82.88**
		0.7	95.96	88.41	88.60	53.78	83.21	82.00
24	5	0.6	95.46	88.68	88.75	**57.64**	83.85	82.44
		0.7	**96.47**	88.27	88.42	56.97	82.85	81.70
	6	0.6	95.83	**88.77**	**88.87**	56.13	**84.13**	82.77
		0.7	96.30	88.42	88.54	55.29	83.13	81.90
MiRA-Face[45]			95.46	**90.65**	**90.62**	51.09	80.56	79.26
UMDNets[2]			94.28	86.62	86.75	53.27	74.69	72.90
FEBNet+[49] (Ours)			**95.83**	88.77	88.87	**56.13**	**84.13**	**82.77**

Table 5. Hyper parameter search for re-ranking[49] method on the final model. Here, k_1 = the count for finding k-reciprocal nearest neighbors, k_2 = count for k-reciprocal nearest neighbor expansion, λ = ratio of importance given to original distance matrix with respect to jaccard distance during re-ranking. MiRA-Face[45] and UMDNets[2] are the present state of arts in DFW-2018 dataset.

GAR@1% FAR.

4.5. Test set evaluation

The final model "FEBNet with re-ranking" is evaluated in the test set[32] by the DFW-2019 competition organizers and the results are outlined in Table 6. We compare the performance of our model with two base models available with the test set namely 1) ResNet50 and 2) LightCNN-29v2.

Model	GAR							
	@0.1% FAR				@0.01% FAR			
	Protocol				Protocol			
	1	2	3	4	1	2	3	4
ResNet-50[32]	47.6	35.4	46.4	35.9	38.4	16.4	22.4	16.9
LightCNN-29v2[32]	**74.4**	55.6	69.2	55.7	**51.2**	36.9	47.2	36.5
FEBNet (ours)	54.8	**92.3**	**78.8**	**90.8**	42.4	**87.7**	**47.6**	**73.7**

Table 6. Test dataset results

We can observe from Table 6 that our model comfortably outperforms the ResNet-50 models in all of the protocols. Specifically, our model improves GAR@0.1% FAR / GAR@0.01% FAR of protocol 1 by 7.2% / 4%, protocol 2 by 56.9% / 71.3%, protocol 3 by 32.4% / 25.2% and protocol 4 by 54.9% / 56.8%. In case of the LightCNN-29v2 model, our model outperforms in all the protocols except protocol 1. Specifically, our model improves GAR@0.1% FAR / GAR@0.01% FAR of protocol 2 by 36.7% / 50.8%, protocol 3 by 9.6% / 0.4% and protocol 4 by 35.1% / 37.2%. We also depict the log scale Receiver Operating Characteristics (ROC) curve of our final model's performance in Test set in the Figure 4.

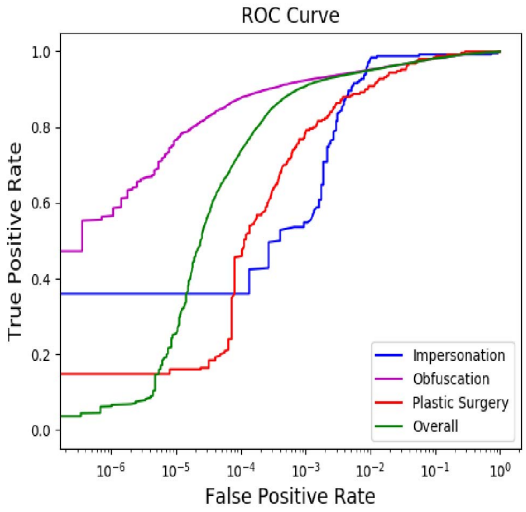

Figure 4. The ROC curve for the Disguised Faces in the Wild (DFW) test dataset

The test set[32] results illustrate the effectiveness of the proposed model to tackle disguised face recognition. Though LightCNN-29v2 outperforms our model in protocol 1 (impersonator distinguishing task), it fails to perform well on much easier (on human scale) task of distinguishing different persons (protocol 2). In contrast, our model strives for improved performance in disguised face recognition without degrading performance in the natural inter-person discriminative face recognition task.

5. Conclusion

In this work, we proposed a transfer learning based ensemble model for disguised face recognition. We started with fine-tuning the pretrained base models using two novel proposed loss functions. Then, we benchmarked the fine-tuned base model's performance in the validation dataset and further, we explored the combination of base models and arrived at the final model that achieves superior performance. Additionally, we also employed a less-explored strategy of re-ranking in face recognition to the task of disguised face recognition and verify that it improves the performance significantly by extensive empirical studies.

Acknowledgements: This work is supported by grants from PM's fellowship for Doctoral Research (SERB, India) & Google PhD Fellowship to Arulkumar Subramaniam.

References

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (12):2037–2041, 2006. 2

[2] Ankan Bansal, Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa. Deep features for recognizing disguised faces in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 10–16, 2018. 1, 3, 7, 8

[3] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36, 2012. 6

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018. 2, 3, 5, 6

[5] Ondrej Chum, James Philbin, Josef Sivic, Michael Isard, and Andrew Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 3

[6] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. 2005. 2

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018. 2, 3

[8] Li Deng and John C Platt. Ensemble deep learning for speech recognition. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014. 7

[9] Tim Esler. Pretrained pytorch face detection and recognition models. https://github.com/timesler/facenet-pytorch, 2018. 2, 3, 6

[10] Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, and Alberto Del Bimbo. Investigating nuisance factors in face recognition with dcnn representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 81–89, 2017. 1

[11] Mislav Grgic, Kresimir Delac, and Sonja Grgic. Scface–surveillance cameras face database. *Multimedia tools and applications*, 51(3):863–879, 2011. 2

[12] Adam Harvey and Jules LaPlace. Megapixels: Origins, ethics, and privacy implications of publicly available face recognition image datasets, 2019. 2, 3, 5, 6

[13] 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. 3, 6

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 4

[15] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[16] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 2

[17] Anil K Jain. *Handbook of face recognition*. Springer. 1

[18] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 3, 6

[19] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017. 7

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollr. Focal loss for dense object detection. *arXiv:1708.02002*, 2017. 3

[21] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[22] Mostafa Mehdipour Ghazi and Hazim Kemal Ekenel. A comprehensive analysis of deep learning based representation for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–41, 2016. 1

[23] David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, 11:169–198, 1999. 7

[24] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. In *IEEE Transactions on Knowledge and Data Engineering*, pages 1345–1359. IEEE, 2009. 6

[25] PM Panchal, SR Panchal, and SK Shah. A comparison of sift and surf. *International Journal of Innovative Research in Computer and Communication Engineering*, 1(2):323–327, 2013. 2

[26] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *bmvc*, volume 1, page 6, 2015. 1

[27] Danfeng Qin, Stephan Gammeter, Lukas Bossard, Till Quack, and Luc Van Gool. Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In *CVPR 2011*, pages 777–784. IEEE, 2011. 3

[28] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D. Castillo, and Rama Chellappa. An all-in-one convolutional neural network for face analysis. *arXiv:1611.00851*, 2016. 3

[29] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 806–813, 2014. 6

[30] Florian Schroff, Dmitry Kalenichenko, and James 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, 6

[31] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 53–58. IEEE, 2002. 2

[32] Maneet Singh, Mohit Chawla, Richa Singh, Mayank Vatsa, , and Rama Chellappa. Disguised faces in the wild 2019, technical report. 2019. 1, 2, 5, 8

[33] Maneet Singh, Richa Singh, Mayank Vatsa, N. Ratha, and Rama Chellappa. Recognizing disguised faces in the wild. In *IEEE Transactions on Biometrics, Behavior, and Identity Science, Volume 1, No. 2*, pages 97–108, 2019. 1, 2, 5, 7

[34] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided Contrastive Attention Model for Person Re-Identification. 2, 3, 4

[35] Arulkumar Subramaniam, Prashanth Balasubramanian, and Anurag Mittal. NCC-net: Normalized cross correlation based deep matcher with robustness to illumination variations. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1944–1953. IEEE, 2018. 1

[36] 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, 3, 5, 6

[37] 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, 2

[38] Josephine Sullivan Vahid Kazemi. One millisecond face alignment with an ensemble of regression trees. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1867–1874, 2014. 7

[39] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei 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. 1

[40] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. *arXiv preprint arXiv:1801.09414*, 2018. 2

[41] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2119–2128, 2018. 2

[42] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014. 6

[43] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 2

[44] 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, Oct 2016. 3, 6

[45] Kaipeng Zhang, Ya-Liang Chang, and Winston Hsu. Deep disguised faces recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 32–36, 2018. 1, 2, 7, 8

[46] Jian Zhao. High-performance face recognition library on pytorch. https://github.com/ZhaoJ9014/face.evoLVe.PyTorch, 2018. 2, 3, 6

[47] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003. 1

[48] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. *IEEE Transactions on Image Processing*, 2019. 2

[49] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017. 1, 2, 3, 4, 5, 8