# **DocFace: Matching ID Document Photos to Selfies**\*

Yichun Shi and Anil K. Jain Michigan State University East Lansing, Michigan, USA

shiyichu@msu.edu, jain@cse.msu.edu

# Abstract

Numerous activities in our daily life, including purchases, travels and access to services, require us to verify who we are by showing ID documents containing face images, such as passports and driver licenses. An automatic system for matching ID document photos to live face images in real time with high accuracy would speed up the verification process and reduce the burden on human operators. In this paper, we propose a new method, DocFace, for ID document photo matching using the transfer learning technique. We propose to use a pair of sibling networks to learn domain specific parameters from heterogeneous face pairs. Cross validation testing on an ID-Selfie dataset shows that while the best CNN-based general face matcher only achieves a TAR=61.14% at FAR=0.1% on the problem, the DocFace improves the TAR to 92.77%. Experimental results also indicate that given sufficiently large training data, a viable system for automatic ID document photo matching can be developed and deployed.

# **1. Introduction**

Identity verification plays an important role in our daily lives. For example, access control, physical security and international border crossing require us to verify our access (security) level and our identities. A practical and common approach to this problem involves the comparison of a subject's live face to the image on a photo ID document in order to verify identity. For example, immigration and customs officials look at the passport photos to confirm a traveler's identity. Clerks at supermarkets look at driver licenses to check a customer's age. This task of ID document photo matching is conducted in numerous scenarios, but it is primarily conducted by humans, which is time-consuming, costly and potentially error-prone. A study pertaining to the Sydney passport officers shows that even these trained offi-



(a) General face matching



(b) ID document photo matching

Figure 1: Example images from (a) LFW dataset [6] and (b) ID-Selfie-A dataset. Each row shows two pairs from the two datasets, respectively. Compared with the general unconstrained face recognition shown in (a), ID Document photo matching (b) does not need to consider large pose variations. Instead, it involves a number of other challenges such as aging and information loss via image compression.

cers perform poorly on matching unfamiliar faces to passport photos, with a 14% false acceptance rate [26]. Therefore, an automatic system that matches ID document photos to selfies<sup>\*</sup> efficiently with low error rates is required in these applications. In addition, automatic ID matching systems also enable remote authentication applications that are otherwise not feasible, such as onboarding new customers in a mobile app (by verifying their identities for account creation), or account recovery in the case of forgotten passwords. One application scenario of ID document photo matching system is illustrated in Figure 3.

A number of automatic ID document photo to selfies matching systems have been deployed at international borders. The earliest such system is SmartGates deployed in Australia [27] (See Figure 2). Due to the increasing number of travelers to Australia, the Australian government introduced SmartGate at most of its international airports for electronic passport control checks for ePassport holders. To

<sup>\*</sup>Technically, the word "selfies" refer to self-captured photos from mobile phones. But here, we define "selfies" as any self-captured live face photos, including those from mobile phones and kiosks.



(a) SmartGate (Australia) [27]

(b) ePassport gates (UK) [28]

(c) Automated Passport Control (US) [21]

Figure 2: Examples of automatic ID document photo matching systems at international borders.



Figure 3: An application scenario of the ID document matching system. The kiosk scans the ID document or reads its chip to obtain the document face photo and the camera takes another photo of the holder's live face (selfie). Then, through face recognition, the system decides whether the holder is indeed the owner of the ID document.

use the SmartGate, travelers only need to let a machine read their ePassport chips containing their digital photos and then capture their face images using a camera mounted at the SmartGate. After verifying a traveler's identity by face comparison, the gate is automatically opened for the traveler to enter Australia. Similar machines have also been installed in the UK (ePassport gates) [28], USA (US Automated Passport Control) [21] and other countries. However, all of the above border crossing applications read the subject's face image from ePassport's chip. If a traveler does not have an ePassport, he will still have to be processed by an inspector who will do the typical manual photo comparison. In addition to international border control, some businesses are providing face recognition solutions to ID document verification for online services [7] [13].

The problem of ID document face matching involves many difficulties that are different from general face recognition. For typical unconstrained face recognition tasks, the main difficulties lie in the pose, illumination and expression (PIE) variations. But in document photo matching, we are comparing a scanned or digital document photo to a digital camera photo of a live face. Assuming that the user is cooperative, both of the images are captured under constrained conditions and large PIE variations would not be present. Instead, low quality of document photos due to image compression<sup>1</sup>, and the time gap between document issue date and verification time remain as the main difficulties, as shown in Figure 1. In addition, since most modern face recognition systems use deep networks, another difficulty faced in our problem is the lack of a large training dataset (pairs of ID photos and selfies).

In spite of these numerous applications and associated challenges, there is a paucity of research on this topic. Indeed, only a few studies have been published on ID document matching [18][1][2][17], all of which are now dated. It is important to note that face recognition technology has made tremendous strides in the past five years, mainly due to the availability of large scale face training data and deep neural network models for face recognition. Verification Rate (VR) on the Labeled Faces in the Wild (LFW) dataset, one of the first public domain "faces in the wild" dataset, has increased from 41.66% in 2014 [11] to 98.65%in 2017 [5] at a False Accept Rate (FAR) of 0.1%. Hence, the earlier published results on ID document photo to live face matching are now obsolete. Advances in face recognition algorithms allow us to build more accurate and robust matchers for ID document matching.

In this paper, we first briefly review existing studies on the ID document photo matching problem and state-of-theart deep neural network-based face recognition methods. We then propose DocFace, a face matcher for ID document photos by exploiting transfer learning techniques. Our experiments use two datasets of Chinese Identity Cards with corresponding camera photos to evaluate the performance of (i) a Commercial-Off-The-Shelf (COTS) face matcher,

<sup>&</sup>lt;sup>1</sup>Most chips in e-Passports have a memory ranging from 8KB to 30KB; the face images need to be compressed to be stored in the chip. See https://www.readid.com/blog/face-images-in-ePassports

(ii) open source deep network face matchers, and (iii) the proposed method. The contributions of the paper are summarized below:

- An evaluation of published face matchers showing ID document photo matching is a non-trivial problem with different challenges from general face matching.
- A new recognition system with a pair of sibling networks for learning representations from heterogeneous face pairs.
- An open-source face matcher<sup>1</sup>, namely DocFace, for ID Document photo matching, which significantly improves the performance of existing general face matchers. The TAR on a private Chinese Identity Card dataset is improved from 61.14% to 92.77% at FAR=0.1%.

# 2. Related Works

# 2.1. ID Document Photo Matching

To the best of our knowledge, the first study on ID document face photo matching is attributed to Starovoitov et al. [18] [17]. Assuming all face images are frontal faces without large expression variations, the authors first localize the eyes with Hough Transform. Based on eye locations, the face region is cropped and gradient maps are computed as feature maps. The algorithm is similar to a general constrained face matcher, except it is developed for a document photo dataset. Bourlai et al. [1][2] considered ID document face recognition as a comparison between two degraded face images by scanning the document photo against high quality live face images. To eliminate the degradation caused by scanning, Bourlai et al. inserted an image restoration phase before comparing the photos using a general face matcher. In particular, they train a classifier to classify the degradation type for a given image, and then apply degradation-specific linear and nonlinear filters to restore the degraded images. Compared with their work on scanned documents, the document photos in our datasets are read from the chips in the Chinese Identity Cards. Additionally, our method is not designed for any specific degradation type but could be applied to any ID document photos. Concurrent with our work, Zhu et al. [31] have also worked on the same problem of ID photo matching. They formulate it as a bisample learning problem and develop a recognition system on a collection of 2.5M ID-selfie pairs.

#### 2.2. Deep Face Recognition

Since the success of deep neural networks in the ImageNet competition [10], all of the ongoing research and development in face recognition now utilizes deep neural networks to learn face representations [20] [19] [16] [5]. The popularity of deep neural networks could partially be attributed to a special property that the low-level image features are transferable, i.e. they are not limited to a particular task, but applicable to many image analysis tasks. Given this property, one can first train a network on a large dataset to learn salient low-level features, then train a domain specific neural network by transfer learning on a relatively small dataset. For example, Sankaranarayanan et al. [15] proposed to retrain networks by using a triplet probability embedding (TPE) loss function and achieved good results on the IJB-A benchmark [9]. Xiong et al. [29] proposed a framework named Transferred Deep Feature Fusion (TDFF) to fuse the features from two different networks trained on different datasets and learn a face classifier in the target domain, which achieved state-of-the-art performance on IJB-A dataset.

# 3. Datasets

In this section we briefly introduce the datasets that are used in this paper. Some example images of the datasets are shown in Figure 4. Due to privacy issues, we cannot release the ID-Selfie-A and ID-Selfie-B datasets. But by comparing our results with public face matchers, we believe it is sufficient to show the difficulty of the problem and advantage of the proposed method.

# 3.1. MS-Celeb-1M

The MS-Celeb-1M dataset [4] is a public domain face dataset facilitating training of deep networks for face recognition. It contains 8, 456, 240 images of 99, 892 subjects (mostly celebrities) downloaded from internet. In our transfer learning framework, it is used as the source domain to train a very deep network with rich low-level features. However, the dataset is known to have many mislabels. We use a cleaned version of MS-Celeb-1M with 5, 041, 527 images of 98, 687 subjects<sup>2</sup>. Some example images from this dataset are shown in Figure 4a.

### 3.2. ID-Selfie-A Dataset

Our first ID document-selfie dataset is a private dataset composed of 10,000 pairs of ID Cards photo and selfies. The ID card photos are read from chips in the Chinese Resident Identity Cards<sup>3</sup>. The selfies are from a stationary camera. Among the 10,000 pairs, we were able to align only 9,915 pairs, i.e. a total of 19,830 images. Assuming all the participants are cooperative, and hence there should be no failure-to-enroll case, we only keep these aligned pairs

<sup>&</sup>lt;sup>1</sup>The source code is available at https://github.com/ seasonSH/DocFace

<sup>&</sup>lt;sup>2</sup>https://github.com/AlfredXiangWu/face\_ verification\_experiment.

<sup>&</sup>lt;sup>3</sup>https://en.wikipedia.org/wiki/Resident\_ Identity\_Card



(a) MS-Celeb-1M

(b) ID-Selfie-A

(c) ID-Selfie-B

Figure 4: Example images in each dataset. The left image in each pair in (b) and each row in (c) is the ID photo and on its right are the corresponding selfies.

for our experiments. In experiments, we will conduct cross validation on this dataset to evaluate the efficacy of our method. Some example pairs from this dataset are shown in Figure 4b.

#### **3.3. ID-Selfie-B Dataset**

Our second ID document-selfie dataset is a private dataset composed of 10, 844 images from 547 subjects, each with one ID Card image and a varying number of selfies from different devices, including mobile phones. Compared with ID-Selfie-A, the selfies in this dataset are less constrained and some images have been warped or processed by image filters, as shown in Figure 4c. Out of these 547 subjects, some subjects do not have any selfie photos. After cleaning and alignment, we retain 10, 806 images from 537 subjects and use them for cross-dataset evaluation of the model trained on ID-Selfie-A dataset. There is no overlapping identities between ID-Selfie-A dataset and ID-Selfie-B dataset. See Figure 4c for example images in this dataset.

# 4. Methodology

# 4.1. Notation

We first train a network as *base model* on the source domain, i.e. unconstrained face dataset and then transfer its features to the target domain, ID and Selfie face images. Let  $X^s = \{(x_i^s, y_i^s) | i = 1, 2, 3, \cdots, N^s\}$  be the dataset of source domain, where  $x_i^s \in \mathbb{R}^{h \times w}$  and  $y_i^s = 1, 2, 3, \cdots, C$  are the  $i^{th}$  image and label, respectively, h and w are the height and width of images,  $N^s$  is the number of images, and C is the number of classes. The training dataset of the target domain is denoted by  $X^t = \{(x_{i1}^t, x_{i2}^t) | i = 1, 2, 3, \cdots, N^t\}$  where  $x_{i1}^t \in \mathbb{R}^{h \times w}$  and  $x_{i2}^t \in \mathbb{R}^{h \times w}$  refer to the ID image and selfie of the  $i^{th}$  subject in the source domain, respectively. Here,  $N^t$  is the number of ID-selfie pairs rather than the number of images. Function  $\mathcal{F}$ :  $\mathbb{R}^{h \times w} \to \mathbb{R}^d$  denotes the base model for the source

domain, where *d* is the dimensionality of the face representation. Similarly,  $\mathcal{G} : \mathbb{R}^{h \times w} \to \mathbb{R}^d$  represents the face representation network for ID photos and  $\mathcal{H} : \mathbb{R}^{h \times w} \to \mathbb{R}^d$  for selfies. An overview of the work flow is shown in Figure 5.

#### 4.2. Training on source domain

The source domain in our work is unconstrained face recognition, where we can train a very deep network on a large-scale dataset composed of different types of face images from a large number of subjects, i.e. MS-Celeb-1M. The objective is to train a base model  $\mathcal{F}$  so that resulting face representations maximizes inter-subject separation and minimizes intra-subject variations. To ensure its performance for better transfer learning, we utilize the popular Face-ResNet architecture [5] to build the convolutional neural network. We adopt the state-of-the-art Additive Maxmargin Softmax (AM-Softmax) loss function [22][3][24] for training the base model. For each training sample in a mini-batch, the loss function is given by:

$$\mathcal{L}_s = -\log \frac{\exp(s\cos\theta_{y_i,i} - m)}{\exp(s\cos\theta_{y_i,i} - m) + \sum_{j \neq y_i} \exp(s\cos\theta_{j,i})}$$
(1)

where

$$\cos \theta_{j,i} = W_j^T f_i$$
$$W_j = \frac{W_j^*}{\|W_j^*\|_2^2}$$
$$f_i = \frac{\mathcal{F}(x_i^s)}{\|\mathcal{F}(x_i^s)\|_2^2}.$$

 $W^* \in \mathbb{R}^{d \times C}$  is the weight matrix and m is a hyperparameter for controlling the margin. The scale parameter s can either be manually chosen or automatically learned [23]; we use automatically learned s for simplicity. During training, the loss in Equation (1) is averaged across all images in the mini-batch.

#### 4.3. Training on target domain

The target domain is a relatively small dataset composed of ID-selfie image pairs. The sources of these images are very different from those from the source domain, thus directly applying  $\mathcal{F}$  to these images will not work well. Because the ID images and selfies are from different sources, the problem can also be regarded as an instance of the heterogeneous face recognition [8]. A common approach in heterogeneous face recognition is to utilize two separate domain-specific models to map images from different sources into a unified embedding space. Therefore, we use a pair of sibling networks  $\mathcal{G}$  and  $\mathcal{H}$  for ID images and selfie images, respectively, which share the same architecture but could have different parameters. Both of their features are transferred from  $\mathcal{F}$ , i.e. they have the same initialization. Notice that although this increases the model size, the inference speed will remain unchanged as each image is only fed into one of the sibling networks.

Inspired by recent metric learning methods [16], we propose a *Max-margin Pairwise Score* (MPS) loss for training the heterogeneous face pair dataset. For each mini-batch of size M, M/2 ID-selfie pairs are randomly selected from all the subjects. For each pair, the MPS loss is given by:

$$\mathcal{L}_t = [\max_{j \neq i}(\max(\cos\theta_{j,i}, \cos\theta_{i,j})) - \cos\theta_{i,i} + m']_+ (2)$$

T

where

$$\cos \theta_{i,j} = g_i^{T} h_j$$
$$g_i = \frac{\mathcal{G}(x_{i1}^t)}{\|\mathcal{G}(x_{i1}^t)\|_2^2}$$
$$h_i = \frac{\mathcal{H}(x_{i2}^t)}{\|\mathcal{H}(x_{i2}^t)\|_2^2}.$$

The loss is averaged across all the M/2 pairs. Here, j iterates over all the other subjects in the batch.  $[x]_+ = max(0, x)$ . Hyper-paramter m' is similar to the m in the AM-Softmax. The idea of the MPS loss function in Equation (2) is to learn a representation by maximizing the margin between genuine pair similarities and imposter pair similarities. The MPS loss simulates the application scenario where the ID photos act like templates, while selfies from different subjects act like probes trying to be verified, or vice versa. Notice that, after the hardest imposter pair is chosen by maximum score, the MPS loss is similar to Triplet Loss [16] with one of the ID / selfie images as the anchor.

### 5. Experiments

#### 5.1. Experimental Settings

We conduct all of our experiments using Tensorflow r1.2. When training the base model on MS-Celeb-1M, we



Figure 5: Overview of the work flow of the proposed method. We first train a base model  $\mathcal{F}$  on a large scale unconstrained face dataset. Then the features are transferred to domain-specific models  $\mathcal{G}$  and  $\mathcal{H}$ , which are trained on an ID-Selfie dataset using the proposed MPS loss function.

use a batch size of 256 and keep training for 280K steps. We start with a learning rate of 0.1 and it is decreased to 0.01, 0.001 after 160K and 240K steps, respectively. When fine-tuning on the ID-Selfie-A dataset, we keep the batch size of 256 and train the sibling networks for 800 steps. We start with a lower learning rate of 0.01 and decrease the learning rate to 0.001 after 500 steps. For both the training stages, the model is optimized by Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a weight decay of (5e - 4). All the images are aligned via similarity transformation based on landmarks detected by MTCNN [30] and are resized to  $96 \times 112$ . We set margin parameters m and m' as 5.0 and 0.5, respectively. All the training and testing are run on a single Nvidia Geforce GTX 1080Ti GPU with 11GB memory. The inference speed of our model on this GPU is 0.003s per image.

By utilizing the MS-Celeb-V1 dataset and the AM-Softmax loss function in Equation (1), our Face-ResNet network achieved 99.67% accuracy on the standard verification protocol of LFW and a Verification Rate (VR) of 99.60% at False Accept Rate (FAR) of 0.1% on the BLUFR [11] protocol.

In the following subsections, five-fold cross validation is conducted for all experiments on ID-Selfie-A dataset to evaluate the performance and robustness of the methods. The dataset is equally split into 5 partitions, and in each fold, one split is used for testing while the remaining are used for training. In particular, 7,932 and 1,983 pairs are used for training and testing, respectively, in each fold. We use the whole ID-Selfie-B dataset for cross-dataset evaluation. Cosine similarity is used as comparison score for all

Model	Loss	Sibling Networks	VR(%) @FAR= 0.01%	VR(%) @FAR= 0.1%
FS	MPS	Yes	$0.03\pm0.04$	$0.07\pm0.04$
BM	-	No	$67.88 \pm 1.72$	$82.06 \pm 1.40$
TL	L2-Softmax	Yes	$70.53 \pm 1.73$	$85.15 \pm 1.38$
TL	AM-Softmax	Yes	$71.07 \pm 1.81$	$85.24 \pm 1.52$
TL	MPS	No	$85.71 \pm 1.29$	$92.51 \pm 1.13$
TL	MPS	Yes	$86.27 \pm 1.39$	$92.77 \pm 1.03$

Table 1: Performance of different approaches for developing an ID Face matcher on the ID-Selfie-A dataset. "FS", "BM" and "TL" refer to "from scratch", "base model" and "transfer learning", respectively. "VR" refers to Verification Rate. For the pre-trained model, because there is no training involved, we leave the loss function as blank.

experiments.

#### **5.2. Exploratory Experiments**

In this section, by using the ID-Selfie-A dataset, we compare different ways to develop an ID-Selfie face matcher. First, we evaluate the approaches without transfer learning: (1) a network trained from scratch with the same architecture and MPS loss function, and (2) the base model pretrained on MS-Celeb-1M but not fine-tuned. To justify the efficacy of the proposed MPS loss function, we fine-tune the base model on the ID-Selfie-A dataset using two other loss functions: L2-Softmax [14] and AM-Softmax [22], which achieved successful results in unconstrained face recognition. Finally, using the base model and MPS loss function, we compare the performance of sibling networks, i.e. different parameters for  $\mathcal{G}$  and  $\mathcal{H}$ , to that of a shared network, i.e.  $\mathcal{G} = \mathcal{H}$ . As mentioned in Section 5.1, all the experiments are conducted using five fold cross validation, and we report the average performance and standard deviation.

The results are shown in Table 1. Because the ID-Selfie-A dataset is such a small dataset, the model trained from scratch (FS) overfits heavily and performs very poorly. Similar results were observed even when we trained a smaller network from scratch. In comparison, the base model (BM) pre-trained on MS-Celeb-V1 performs much better, even before fine-tuning. This confirms that the rich features learned from unconstrained face datasets are transferable and can be helpful for developing domain specific matchers with small datasets. This performance is then further improved after transfer learning (TL). Although both L2-Softmax and AM-Softmax lead to an improvement in the performance, our proposed loss function (MPS) outperforms the pre-trained model even more significantly. This is because our loss function is specially designed for the problem, and directly maximizes the margin of pairwise score rather than classification probability. Finally, we find that a pair of sibling networks  $\mathcal{G}$  and  $\mathcal{H}$  slightly outperforms a shared network. This means learning separate domainspecific models for ID photos and selfies could help the system learn more discriminative low-level features and lead to

Method	VR(%) on ID-Selfie-A		VR(%) on LFW
	@FAR = 0.01%	@FAR = 0.1%	@FAR = 0.1%
COTS	$27.32 \pm 1.46$	$46.33 \pm 1.61$	92.01
CenterFace [25]	$28.02 \pm 1.93$	$60.10 \pm 1.68$	91.70
SphereFace [12]	$34.76 \pm 0.88$	$61.14 \pm 0.82$	96.74
DocFace	$86.27 \pm 1.39$	$92.77 \pm 1.03$	-

Table 2: Comparison of the proposed method with existing general face matchers on the ID-Selfie-A dataset under five-fold cross validation protocol. "VR" refers to Verification Rate. "TL" refers to Transfer Learning. For comparison, we report the performance of existing matchers on LFW according to BLUFR protocol [11]. The proposed model, DocFace is shown in italic style.

better face representations in the shared embedding space. But since the sibling networks also introduce more parameters, this improvement is limited by the small training dataset. A larger advantage of sibling networks might be observedd when training on more ID-Selfie samples.

#### 5.3. Comparison with Existing Matchers

To the best of our knowledge, there are no existing public face matchers in the domain of ID Document matching. Therefore, we evaluate the performance of existing general face matchers on the ID-Selfie-A dataset and compare them with the proposed method. To make sure our experiments are comprehensive enough, we compare our method not only with a Commercial-Off-The-Shelf (COTS) matcher, but also two open-source matchers representing the stateof-the-art unconstrained face recognition methods: Center-Face<sup>1</sup> [25] and SphereFace<sup>2</sup> [12]. During the five-fold cross validation, because the existing methods don't involve training, only the test split is used. For comparison, we also report the performance of the existing matchers on the unconstrained face dataset, LFW [6], using the BLUFR protocol [11].

The results are shown in Table 2. As one can see, although all the existing methods perform well on the unconstrained face dataset, there is a large performance drop when we test them on the ID-Selfie-A dataset. This is consistent with our observation in Section 1 that the characteristics of images and difficulties in the two problems are very different. In comparison, the proposed method significantly improves the performance on the target problem. Some false accept and false reject image pairs of our model are shown in Figure 6. From the figure, we can see that most of the selfies in the genuine pairs either have a obfuscations such as makeup or glasses, or a drastic appearance change due to aging. Additionally, many impostor pairs look surprisingly similar, and because of low quality of the ID image, it is hard to find fine-grained clues to tell that they are actually different people.

<sup>&</sup>lt;sup>1</sup>https://github.com/ydwen/caffe-face <sup>2</sup>https://github.com/wyliu/sphereface



(a) False accept pairs



(b) False reject paris

Figure 6: Examples of falsely classified images by our model on ID-Selfie-A dataset at FAR = 0.1%.

# 5.4. Effect of Dataset Size

In the previous sections, we fix the dataset size and conduct cross validation to test the performance of different matchers and training strategies. Here, we want to explore how much the size of the training dataset could affect our domain-specific network and whether there is a potential for improvement by acquiring more training data. We conduct the same five-fold cross validation, where in each fold we keep the test split unchanged but randomly select a subset of the ID-Selfie pairs in the training splits and report the average performance across the five folds. In particular, we select 1,000, 3,000, 5,000 and all (7,932) image pairs for training. The resulting TAR along with the dataset size is shown in Figure 7. For both FAR=0.01% and FAR=0.1%, the performance keeps increasing as the training dataset becomes larger. Notice that we are increasing the size of dataset linearly, which means the relative growth rate of the dataset size is decreasing, yet we can still observe a trend of increasing performance for larger datasets. This indicates that more performance gain would be expected if we could increase the size of the dataset by one or two orders of magnitude.

# 5.5. Cross-dataset Performance Evaluation

Although it is an important application to match an ID document photos to selfies from stationary cameras, in many other scenarios, the selfies could be captured by different devices, including mobile phones. An ideal model should perform robustly in both cases. Therefore, we train



Figure 7: Performance when training on subsets of different sizes on ID-Selfie-A dataset. The subsets are randomly selected from the training splits. The performance is reported by taking the average of the five folds.

Method	VR(%)@FAR= 0.01%	VR(%)@FAR = 0.1%
COTS	13.97	30.91
CenterFace [25]	17.69	35.20
SphereFace [12]	34.82	54.19
Base model	70.87	86.77
DocFace	78.40	90.32

Table 3: Cross-dataset evaluation on the ID-Selfie-B dataset. The "base model" is only trained on MS-Celeb-1M. The model *Doc-Face* has been fine-tuned on ID-Selfie-A. Our models are shown in italic style.

a model on the entire ID-Selfie-A dataset and test it on ID-Selfie-B dataset, whose selfies are from different sources. In testing, for subjects in ID-Selfie-B dataset that have more than one selfie images, we fuse their feature vectors by taking the average vector. The results are shown in Table 3. For comparison, we also show the performance of the existing methods. Our model performs the best on ID-Selfie-B, also higher than the base model which has not been fine-tuned on ID-Selfie-A dataset. This indicates that the face representation learned from the ID-Selfie-A dataset is not only discriminative for ID vs. stationary camera pairs, but also useful for other ID-selfie datasets. It also suggests that training on a mixed dataset of images from different sources could be helpful for the performance on all the sub-problems.

# 6. Conclusion

In this paper, we propose a new method, DocFace, which uses transfer learning techniques with a new loss function, Max-margin Pairwise Score (MPS) loss, to fine-tune a pair of sibling networks for the ID document photo matching problem. By using two private datasets<sup>1</sup>, we evaluate the performance of DocFace and existing unconstrained face

<sup>&</sup>lt;sup>1</sup>To our knowledge, no public-domain dataset for this problem is available. Due to privacy issues, we cannot release these two datasets.

matchers on the ID document matching problem. Experimental results show that general face matchers perform poorly on this problem because it involves many unique challenges. On the other hand, DocFace significantly improves the performance over general face matchers on this problem. We also show that accuracy on test set increases steadily with an increase in the size of the training set, which implies that additional training data could lead to better recognition performance.

# Acknowledgement

This work was supported in part by Visa Inc. We also thank ZKTeco for providing the ID-Selfie datasets.

### References

- T. Bourlai, A. Ross, and A. Jain. On matching digital face images against scanned passport photos. In *IEEE Int. Conf. Biometrics, Identity and Security (BIDS)*, 2009.
- [2] T. Bourlai, A. Ross, and A. K. Jain. Restoring degraded face images: A case study in matching faxed, printed, and scanned photos. *IEEE Trans. on TIFS*, 2011.
- [3] J. Deng, J. Guo, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv*:1801.07698, 2018.
- [4] 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.
- [5] A. Hasnat, J. Bohné, J. Milgram, S. Gentric, and L. Chen. Deepvisage: Making face recognition simple yet with powerful generalization skills. *arXiv*:1703.08388, 2017.
- [6] G. B. Huang, M. Ramesh, T. Berg, and E. 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.
- [7] Jumio. Netverify id verification. https://www.jumio. com/trusted-identity/netverify, 2018.
- [8] B. F. Klare and A. K. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE Trans. on PAMI*, 2013.
- [9] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *CVPR*, 2015.
- [10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [11] S. Liao, Z. Lei, D. Yi, and S. Z. Li. A benchmark study of large-scale unconstrained face recognition. In *IJCB*, 2014.
- [12] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017.
- [13] Mitek. Mitek id verification. https://www. miteksystems.com/mobile-verify, 2018.

- [14] R. Ranjan, C. D. Castillo, and R. Chellappa. L2constrained softmax loss for discriminative face verification. *arXiv*:1703.09507, 2017.
- [15] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa. Triplet probabilistic embedding for face verification and clustering. In *BTAS*, 2016.
- [16] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.
- [17] V. Starovoitov, D. Samal, and D. Briliuk. Three approaches for face recognition. In *International Conference on Pattern Recognition and Image Analysis*, 2002.
- [18] V. Starovoitov, D. Samal, and B. Sankur. Matching of faces in camera images and document photographs. In *ICASSP*, 2000.
- [19] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *CVPR*, 2015.
- [20] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.
- [21] U.S. Customs and Border Protection. Automated passport control (apc). https://www.cbp.gov/travel/ us-citizens/apc, 2018.
- [22] F. Wang, W. Liu, H. Liu, and J. Cheng. Additive margin softmax for face verification. arXiv:1801.05599, 2018.
- [23] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille. Normface: l<sub>2</sub> hypersphere embedding for face verification. arXiv:1704.06369, 2017.
- [24] H. Wang, Y. Wang, Z. Zhou, X. Ji, Z. Li, D. Gong, J. Zhou, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. arXiv:1801.09414, 2018.
- [25] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.
- [26] D. White, R. I. Kemp, R. Jenkins, M. Matheson, and A. M. Burton. Passport officers errors in face matching. *PloS one*, 2014.
- [27] Wikipedia. Australia smartgate. https://en. wikipedia.org/wiki/SmartGate, 2018.
- [28] Wikipedia. epassport gates. https://en.wikipedia. org/wiki/EPassport\_gates, 2018.
- [29] L. Xiong, J. Karlekar, J. Zhao, J. Feng, S. Pranata, and S. Shen. A good practice towards top performance of face recognition: Transferred deep feature fusion. *arXiv*:1704.00438, 2017.
- [30] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016.
- [31] X. Zhu, H. Liu, Z. Lei, H. Shi, F. Yang, D. Yi, and S. Z. Li. Large-scale bisample learning on id vs. spot face recognition. arXiv:1806.03018, 2018.