Mutual Information Maximization on Disentangled Representations for Differential Morph Detection

Sobhan Soleymani, Ali Dabouei, Fariborz Taherkhani, Jeremy Dawson, Nasser M. Nasrabadi
West Virginia University
{ssoleyma, ad0046, ft0009}@mix.wvu.edu, {jeremy.dawson, nasser.nasrabadi}@mail.wvu.edu

Abstract

In this paper, we present a novel differential morph detection framework, utilizing landmark and appearance disentanglement. In our framework, the face image is represented in the embedding domain using two disentangled but complementary representations. The network is trained by triplets of face images, in which the intermediate image inherits the landmarks from one image and the appearance from the other image. This initially trained network is further trained for each dataset using contrastive representations. We demonstrate that, by employing appearance and landmark disentanglement, the proposed framework can provide state-of-the-art differential morph detection performance. This functionality is achieved by the using distances in landmark, appearance, and ID domains. The performance of the proposed framework is evaluated using three morph datasets generated with different methodologies.

1. Introduction

The main goal of biometric systems is automated recognition of individuals based on their unique biological and behavioral characteristics [36]. The human face is widely accepted as a means of biometric authentication. Although, the uniqueness of face images and user convenience of face recognition systems have resulted in their popularity, morphed face images have shown to pose a severe threat to them. This is because the main objective of morph attacks is to purposefully alter or obfuscate the unique correspondence between probe and gallery images [40]. The result of a morph attack is a face image which matches the probe images corresponding to two different face images. Therefore, the detection of morph images plays a major role in providing reliable face recognition.

The majority of morph generation frameworks focus on altering the position of the facial landmarks. These frameworks mainly utilize three steps: correspondence, warping, and blending. The first step aims to detect the corresponding landmarks of both the images. These sets of landmarks are then utilized to warp the images toward each other, *e.g.,* considering the landmarks of the morph image as the pairwise average of two face images. Finally, textures from the two images are combined either over the entire face image [41] or face patches [29]. Another trend of morph generation considers Generative Adversarial Networks (GANs) to construct images that can be matched with the two source images, such as AliGAN [12, 10] and StyleGAN [23, 50].

The face morphing algorithms can affect the face image in two broad aspects. First, they alter the position of the landmarks. On the other hand, they modify the appear-

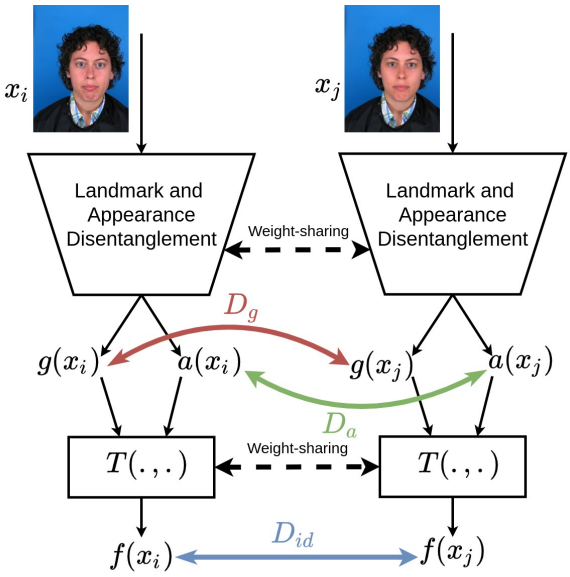

Figure 1. Trusted probe image, x_i, and image in question, x_j, are disentangled into landmark and appearance representations, using the disentanglement network trained on triplet of face images. In these triplets, the constructed intermediate face image inherits landmarks and the appearance from two different face images. Landmark, appearance, and ID representations are utilized to make the decision about the image in question.

ance of the face image by either blending two source images or generating samples using generative models. Although appearance corresponds to the soft biometrics of a subject which are not necessarily unique, such as ethnicity, hair color, and gender, it can still be interpreted to distinguish between face images with similar soft biometrics such as differences in the texture of the face images. However, deep differential morph detection frameworks focus on distinguishing the samples based on the ID information. Our proposed differential morph detection framework investigates both the locations of the landmarks and the appearance of the face image. Therefore, this approach restricts the attacker's morphing capability by studying both the changes resulted from altering the landmarks as well as modification in soft biometrics and texture information.

As presented in Figure 1, our proposed framework learns the disentangled representations for the landmarks and the appearance of a face image. While these representations are practically shown to be sufficient for face recognition [8], the proposed training setup ensures that the mutual information between representations of the real images from a subject is maximized. In this paper, we make the following contributions: i) we construct triplets of images in which an intermediate image inherits the landmarks from one image and the appearance from the other image, ii) these triplets are considered to train a disentangling network which provides disentangled representations for landmarks and face appearance, and iii) we train specific networks for each morph dataset by learning contrastive representations through maximizing the mutual information between real images from each subject.

2. Related Works

2.1. Facial Morphing

Facial morphing studies the possibility of creating artificial biometric samples which resemble the biometric information of two or more individuals [40]. Morph images can be generated with little technical experience using tools available on the internet and mobile platforms [40]. The overall purpose of face morphing is to generate a face image that will be verified against samples of two or more subjects in automated face recognition systems. One of the first efforts to study the generation of a morph image from two source images [13] has concluded that geometric alterations and digital beautification can cause an increase in the possibility of fooling recognition systems. Morph generation techniques can roughly be categorized into landmark-based [29, 43, 44] and generative models [10, 45]. Landmark-based frameworks focus on detecting the landmarks in both the images, translating these points toward each other, and blending the two face images. On the other hand, inspired by a learned inference model [12], Morgan [10] presents a face morphing attack based on automatic image generation using a GAN framework.

2.2. Morph Detection

Morph detection can be categorized into two main approaches [41]: single image morph detection and differential morph detection. Single image morph detection studies the possibility of detecting the morph image in the absence of a reference image. On the other hand, differential morph detection leverages the information extracted from a real image corresponding to the subject. Texture descriptors are the main feature extraction models for single image morph detection [51, 47, 38, 37, 34]. Recently, deep learning models have also been considered for this purpose [44, 43, 33]. The models mentioned can also be employed for differential morph detection when the extracted feature from the two images are compared [35, 39, 9]. Another trend of work for differential morph detection considers that subtracting the trusted image from the image in question should increase the classification score of the resulting image for one of the probe subjects [14, 15, 32].

2.3. Representation Disentanglement

The geometry of landmarks and visual appearance are the two main characteristics of the face that can be utilized for face recognition. Initially, the geometry of hand-crafted face landmarks were basis for face recognition [17]. Neural network approaches have provided state-of-the-art face recognition performance, with several deep models using the location of landmarks for varying face recognition purposes [21, 7]. On the other hand, the effect of appearance in face recognition is widely studied, including soft biometrics such as gender, age, ethnicity, and hair color [18, 16]. Recently, an unsupervised approach using a coupled autoencoder model for disentangling the appearance and geometry of face images was developed [46]. In this framework, each autoencoder learns the geometry or appearance representation of the face, while the reconstruction loss is considered as the supervision for disentangling. Another similar work [55] has incorporated variational autoencoders to improve the disentangling. Another recent generative model [24] presents an unsupervised algorithm for training GANs that learns the disentangled style and content representations of the data.

2.4. Mutual Information and Deep Learning

Among the first works that studied the application of mutual information in deep learning, [31] showed that GAN training loss can be recovered by minimizing the estimated divergence between the generated and true data distributions. The authors in [3] expanded the mutual information maximization techniques to estimate the mutual informa-

tion between two random variables via a neural network. The authors in [5] and [6] used mutual information to quantify the separation of distributions of positive and negative pairings in learning binary hash codes. The authors in [25] introduced RankMI algorithm,an information-theoretic loss function and a training algorithm for deep representation learning for image retrieval. The authors in contrastive representation distillation [48] proposed a contrastive-based objective function for transferring knowledge between deep networks. The authors in [2] propose an approach to self-supervised representation learning based on maximizing mutual information between features extracted from multiple views of a shared context.

3. Proposed Framework

Our proposed differential morph detection framework resonates with the morph generation frameworks in which the the landmarks of the real image are translated to landmarks of the target face image [29] or image generation by generative adversarial networks [10]. Disentangling appearance and landmark information has shown to be a powerful tool for face recognition [8]. These two domains provide the majority of the information content for differential morph detection as well. We aim to study the possibility of detecting the morph image based on its differences with the trusted image in both landmark and appearance domains. Therefore, to train our framework, we construct samples that inherit the appearance and landmarks from different samples. Then, we train a network that can disentangle these two types of information [8]. This framework is then trained for differential morph detection by maximizing the mutual information between representations of genuine pairs.

3.1. Landmark and Appearance Triplets

The first step in our proposed training consists of generating face images that inherit appearance from one image and landmarks from the other image. Then, these triplets of face images are used to train two deep networks. The first network aims to represent the appearance of the face image and the second network extracts the landmark information. The supervision for disentangling appearance and landmarks of faces is provided by constructing triplets of face images. Each triplet consists of two real face images from two different IDs. For convenience we denote these images as appearance image, x_i, landmark image, x'_i, and an intermediate face image generated using the appearance of the first face image and the landmarks of the second face image, $\widehat{x_i}$. To construct this intermediate face image, we translate the landmarks of the appearance image to the landmark image.

For this purpose, let x_i be a face image noted as an appearance image belonging to the class y_i and the set l_i describe the locations of its K landmarks. We find another face image x'_i from a different class corresponding to the closest set of landmarks l'_i as the landmark set. The distance between the sets of landmarks is calculated in terms of L_∞, to assure that x_i and x'_i have similar structures in order to minimize the distortion caused by the spatial transformation in the next step.

We use the thin plate spline (TPS) algorithm [4] to transfer the landmarks of the appearance face image to the landmarks of x'_i as:

$$\widehat{x_i} = \text{TPS}(x, l, l' + \delta_l), \tag{1}$$

where TPS and $\widehat{x_i}$ represent the spatial transformation and the deformed image noted as the intermediate face image. This face image has the appearance of x_i and the landmarks of x'_i. The set δ_l accounts for small perturbations in the localizing the landmarks in the morph generation framework.

3.2. Revisiting Landmarks and Appearance Disentanglement

As presented in Figure 2, in our proposed framework, two networks are defined as appearance network, a, and landmark network, g. These networks map the input face image to the appearance and landmark representations as: $a(.) : \mathbb{R}^{w\times h\times 3} \rightarrow \mathbb{R}^{d_a}$ and $g(.) : \mathbb{R}^{w\times h\times 3} \rightarrow \mathbb{R}^{d_g}$. It is worth mentioning that landmarks can be defined as the salient points in the face image. Although the landmark representation aims to represent the landmarks in the face image, it is trained through a classification setup to preserve the information required to distinguish between the input images regarding their geometrical differences. We define a third network, $f(.)$, that maps these two representations to a face ID representation as: $f(.) : \mathbb{R}^{d_a} \times \mathbb{R}^{d_g} \rightarrow \mathbb{R}^{d_f}$, where d_a, d_g, and d_f are the dimension of appearance, landmark, and face ID representations, respectively. This representations enables us to train the framework as a classification setup.

To provide enough information to distinguish between real and morph images, these three representations should satisfy three conditions: i) The appearance representation of the appearance and intermediate images should be similar: $a(x_i) \approx a(\widehat{x_i})$, ii) the landmark representation of the landmark and intermediate images should be similar: $g(x'_i) \approx g(\widehat{x_i})$, and iii) for both the non-manipulated images, x_i and x'_i, the face representations resulted from network f should preserve sufficient classification information. We address these three conditions in our initial training setup. The appearance-preserving loss function aims to enforce the first condition:

$$L_a^1(x_i, \widehat{x_i}) = -\frac{1}{N}\sum_i \Phi(a(x_i), a(\widehat{x_i})), \tag{2}$$

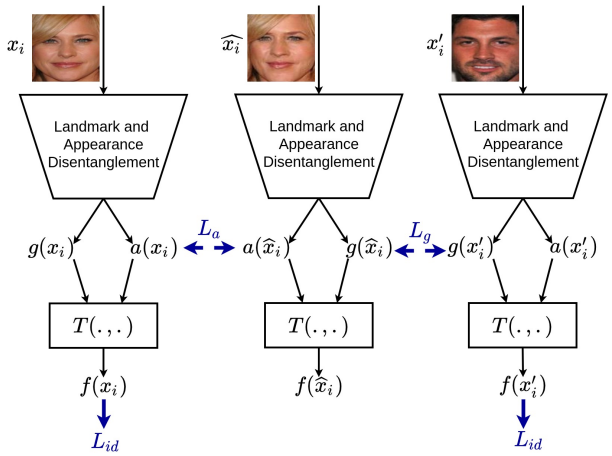

Figure 2. Face image $\widehat{x_i}$ is constructed by considering the appearance of x_i and the landmarks of x'_i. L_a enforces the appearance representations of x_i and $\widehat{x_i}$ to be similar. Similarly, L_g ensures that $g(\widehat{x_i})$ and $g(x'_i)$ are close to each other. A fully-connected layer of size 512 fed with the concatenation of g and a provides the ID representation for the input image.

where $\Phi(v_1, v_2)$ represents the cosine similarity between v_1 and v_2 as in [52, 28]: $\Phi(v_1, v_2) = \frac{v_1^T v_2}{||v_1||_2||v_2||_2}$ and N is the number of samples. Similarly, the landmark-preserving loss is defined as:

$$L_g^1(x'_i, \widehat{x_i}) = -\frac{1}{N}\sum_i \Phi(g(x'_i), g(\widehat{x_i})) + \max(0, \Phi(g(x'_i), g(x_i)) - \alpha_g \phi_g), \tag{3}$$

where $\phi_g = \frac{||l_i - l'_i||_2}{||l_i - \overline{l_i}||_2}$ is the normalized measure of the distance of landmark locations, and $\overline{l_i}$ is the mean of landmark locations along two axes. α_g is a scaling coefficient, scaling to form an angular loss which aims to maximize the cosine similarity of $g(x_i)$ and $g(\widehat{x_i})$ and dissimilarity of $g(x_i)$ and $g(x'_i)$.

In addition to the discussed training loss functions, we should assure that the appearance and landmark representations provide sufficient information for the identification of the real images, x_i and x'_i:

$$L_{id}^1(x_i) = \frac{-1}{N}\sum_i \log \frac{e^{s(\cos(m_1\theta_{y_i,i}+m_2)-m_3)}}{e^{s(\cos(m_1\theta_{y_i,i}+m_2)-m_3)} + \sum_{j\neq y_i} e^{s\cos(\theta_{j,i})}}, \tag{4}$$

where $f(x_i) = T(a(x_i), g(x_i))$ is the ID representation [11] for face image, $\cos(\theta_{j,i}) = \frac{W_j^T f(x_i)}{||W_j||_2||f(x_i)||_2}$, and W_j is the weight vector assigned to the i^{th} class. In this angular loss function, m_1, m_2, and m_3 are the hyperparameters controlling the angular margin, and s is the magnitude of angular representations. The training loss function is defined as:

$$L_t^1 = \sum_i L_{id}^1(x_i) + L_{id}^1(x'_i) + \lambda_a^1 L_a^1(x_i, \widehat{x_i}) + \lambda_g^1 L_g^1(x'_i, \widehat{x_i}), \tag{5}$$

where λ_a^1 and λ_g^1 are hyper-parameters scaling the appearance and landmark preserving loss functions.

3.3. Contrastive Morph Detection

Our proposed differential morph detection framework builds upon recent information-theoretic approaches to deep representation learning [25, 48]. We aim to maximize the mutual information between the real images from the same subject and minimize the mutual information between samples in an imposter pair during the training and make the decision during the test considering the distance between the representations of the pair of images in the embedding. To this aim, as presented in Figures 3, the joint training of the disentanglement and auxiliary networks provides embedding representations distinguishable enough to detect morphed face images in a differential morph detection setup. Our framework benefits from transferring knowledge from that recognition task on a large face dataset to the disentanglement network, which provides a faster training of both disentanglement and auxiliary networks.

To maximize the mutual information between real samples from the same subject in the embedding space, we follow the notation proposed in [25, 3]. Let x_i be an input face image and z_i^a and z_i^g be its corresponding appearance and landmark representations as:

$$z_i^a = a(x_i), z_i^g = g(x_i). \tag{6}$$

We aim to train $a(x)$ and $g(x)$ such that real images from the same subject are mapped closely in the embedding space. To this aim, we maximize the mutual information between the real images from the same subject in each embedding space using the functions $T^a(.)$ and $T^g(.)$. To construct our training samples we define a genuine set as:

$$P = \{(x_i, x_j) | c_i = c_j, r_i = r_j = 1\}, \tag{7}$$

where c_i and c_j represents the classes for the subjects and $r_i = 1$ represents the real images. On the other hand we define the imposter set as:

$$N = \{(x_i, x_j) | c_i \neq c_j \text{ or } r_i = 0 \text{ or } r_j = 0\}, \tag{8}$$

where $r_i = 0$ represents morphed images. It is worth mentioning that we define the above imposter set during the training. During the test phase, the imposer set consists of pairs in which both the samples belong to the same subject, while one of them is a real face image and the other is a morphed face image. In addition, for the genuine set, we

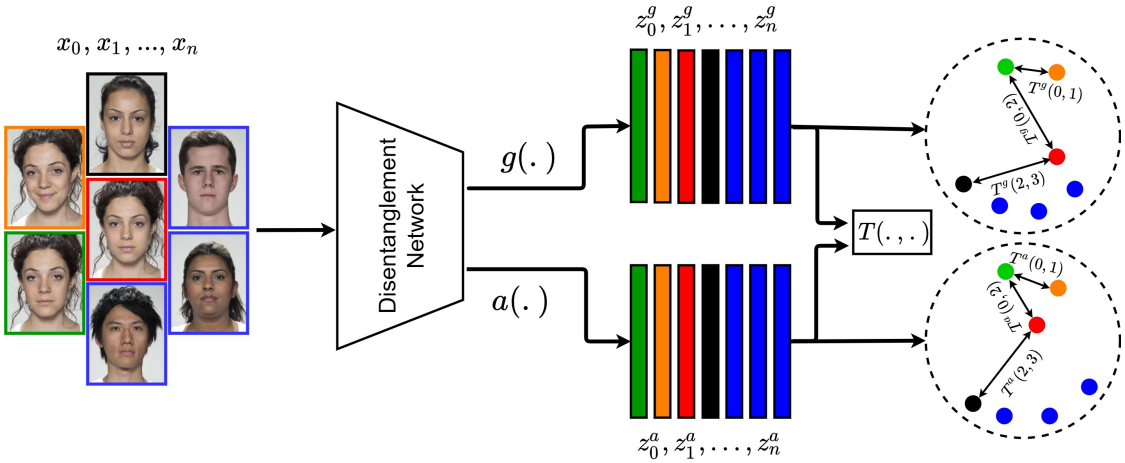

Figure 3. A pair of one trusted probe image, x_i, and an image in question, x_j, are fed into the disentanglement network. This network which is trained in combination with the auxiliary networks, $T^a(.,.)$ and $T^g(.,.)$, provides embedding representations that present high mutual information for genuine pairs and results in close representations for the samples in genuine pair and distant representations for samples in imposter pairs. Here, the morph image (red) is constructed displacing the landmarks of a real image (green) toward the landmarks of a visually similar image (black). The genuine pair consists of two real images from the same subject (orange and green), while the imposter pair in constructed using a real image and its corresponding morph image (green and red).

can define the joint distribution of x_i and x_j as:

$$p(x_i, x_j) = \sum_{k \in C} p(x_i, x_j, c = k, r_i = r_j = 1). \quad (9)$$

Assuming the high entropy of $p(c)p(r)$ for the imposter set, we can approximate the joint distribution of the samples as the product of their marginals:

$$\begin{aligned} &p(x_i)p(x_j) \approx \\ &\sum_{k \in c} \sum_{k' \neq k} \{p(x_i|c_i=k)p(x_j|c_j=k')p(c_i=k)p(c_j=k')\} \\ &+ \sum_{k \in c} \sum_{r_i \in r} \{p(x_i|c_i=k)p(x_j|c_j=k)p(c_i=k)p(c_j=k) \\ &p(r_j=0)\} + \sum_{k \in c} \{p(x_i|c_i=k)p(x_j|c_j=k')p(c_i=k) \\ &p(c_j=k')p(r_i=0)p(r_j=1)\}, \end{aligned} \quad (10)$$

where $r = \{0, 1\}$ represents morphed and real images. Considering the genuine and imposter pairs defined in equations 7 and 8, the appearance differential loss is defined to maximize the mutual appearance information between samples in a genuine pair as [3, 25]:

$$\begin{aligned} L_a^2 = &\frac{1}{||P||} \sum_{(x_i, x_j) \in P} T^a(z_i^a, z_j^a) \\ &- \log \frac{1}{||N||} \sum_{(x_i, x_j) \in N} e^{T^a(z_i^a, z_j^a)}. \end{aligned} \quad (11)$$

A similar loss is defied over the genuine and imposter pairs to calculate L_t^g as the differential landmark information loss. Then, the differential loss is defined as:

$$L_t^2 = \lambda_a^2 L_a^2 + \lambda_g^2 L_g^2 + L_{id}^1, \quad (12)$$

where L_{id}^1 provides the training for network T and subsequently $f(x_i)$.

4. Experiments

We study the performance of the proposed framework on three morph datasets. In our experiments we follow frameworks described in [41, 49]. Evaluation metrics for the differential morph detection are defined as: Attack Presentation Classification Error Rate (APCER) as the proportion of morph attack samples incorrectly classified as bona fide (non-morph), presentation and Bona Fide Presentation Classification Error Rate (BPCER) is the proportion of bona fide (nonmorph) samples incorrectly classified as morphed samples.

4.1. Training Setup

For all the datasets, DLib [26] is considered to detect and align faces, as well as extracting 68 landmarks. We train the model on the CASIA-WebFace [56] dataset. In the training set, for each image, the image from a different ID that provides closest landmarks to its landmarks in terms of L_2 norm is selected. Neighbor face is transformed spatially using Equation 1. This image is aligned again to compensate for the displacements caused by the spatial transformation.

All images are resized to 112×112 and pixel values are scaled to $[-1, 1]$.

We adopt ResNet-64 [19] as the base network architecture. To reduce the size of the model, the convolutional networks for extracting the landmark representation, $g(x)$, and the appearance representation, $a(x)$, are combined. This network produces feature maps of spatial size 7×7 and the depth of 512 channels. These feature maps are divided in depth into two sets, dedicated to the appearance and landmark representations, respectively. Each set of feature maps is reshaped to form a vector of size $12,544$ and passed to dedicated fully-connected layers. These layers of size 256 generate the final representations, $a(x)$ and $g(x)$. The ID representation is constructed by concatenation of these two representations fed to a fully-connected layer of size 512. The model is trained using Stochastic Gradient Descent (SGD) with the mini-batch size of 128 on two NVIDIA TITAN X GPUs. In Equation 4, following ArcFace [11] framework, m_1, m_2, and m_3 are set to 0.9, 0.4, and 0.15, respectively. In Equation 1, δ_l is sampled from $N(0, 3)$.

The initial value for the learning rate is set to 0.1 and multiplied by 0.9 in intervals of five epochs until its value is less than or equal to 10^{-6}. The model is trained for 600K iterations. We select $\alpha_g = 9.4$, $\lambda_a^1 = 1.3$, and $\lambda_g^1 = 0.75$. For training the network using Equation 11, each fully-connected layer of size 256 is fed to a fully-connected of size 64, and then to a single unit. Here, considering $\lambda_a^2 = \lambda_g^2 = 1$, the network is trained using the learning rate of 10^{-2} and is dropped similar to the rate mentioned above.

4.2. Results

MorGAN is constructed using the generative framework described in [10]. In this dataset, 500 bonafide images are considered. For each bona fide image two morph images are generated using two most similar identities to the bona fide image, resulting in $1,000$ morph images. In total this dataset consists of $1,500$ references, $1,500$ probes, and $1,000$ MorGAN morphing attacks. The database is randomly split into disjoint and equal train and test sets. All the images are of size 64×64.

VISAPP17-Splicing-Selected[1] is a subset of VISAPP17-Splicing dataset [30] containing genuine neutral and smiling face images as well as morphed face images. This dataset is generated by warping and alpha-blending [53]. To construct this dataset, facial landmarks are localized, the face image is tranquilized based on these landmarks, triangles are warped to some average position, and the resulting images are alpha-blended, where alpha is set to 0.5 making alpha-blending equal to average. Splicing morphs are designed to avoid ghosting artefacts usually present in the hair region, done by warping and blending of only face regions and inserting the blended face into one of the original face images. The background, hair and torso regions remain untouched. VISAPP17-Splicing-Selected dataset, which consists of 132 bona fide and 184 morph images of size 1500×1200, is constructed by selecting morph images without any recognizable artifacts.

[1]For simplicity, we refer to this dataset as VISAPP17.

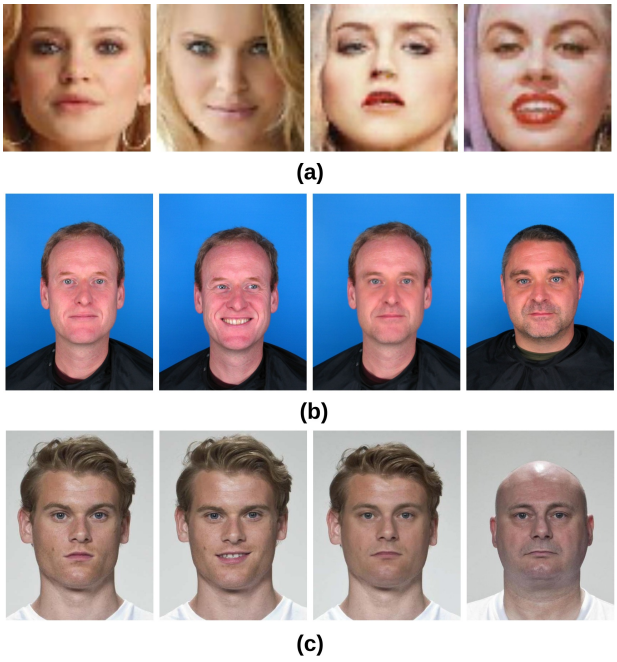

Figure 4. Samples from (a) MorGAN, (b) VISAPP17-Splicing-Selected, and (c) AMSL Face Morph Image Datasets. For each dataset, the first and second faces are the gallery and probe bona fide images and the third face is the morph image construed from the first and forth face images. The original sizes for face images in these datasets are 64×64, 1500×1200, and 531×413, respectively.

The AMSL Face Morph Image Dataset is created using the Face Research Lab London Set [1] and includes genuine neutral and smiling face images and morphed face images. The morphed face images are generated from pairs of genuine face images [30]. For all the morph images the proportions of both faces in the morphed face are the same. While generating morphed faces male, female, white, and Asian people are only morphed with their corresponding category. All images are down-scaled to 531×413 pixels and JPEG compression is applied to them to compress the images to 15kb [54]. This dataset includes 102 neutral or 102 smiling genuine face images and 2,175 morph images.

Differential Morph Detection: For the MorGAN dataset, we follow the train and test split presented in [10]. For the other two datasets, we consider a disjoint train and test split in which 50% of the subjects are used for training. The

Dataset	MorGAN			VISAPP17			AMSL		
	D-EER	5%	10%	D-EER	5%	10%	D-EER	5%	10%
LM-Dlib [9, 26]	12.53	20.71	10.17	17.88	26.64	22.71	14.45	20.67	18.55
BSIF+SVM [22]	10.17	14.22	8.64	16.42	28.77	25.37	12.75	20.71	16.26
LBP+SVM [27]	15.51	28.40	18.71	18.75	23.88	20.65	14.97	21.47	16.21
FaceNet [42]	16.14	38.38	26.67	9.51	29.82	6.91	8.43	25.74	5.68
ArcFace [11]	14.65	22.76	16.23	7.14	17.51	5.69	6.14	14.51	5.23
FaceNet+SVM	12.53	18.84	12.21	8.85	26.46	6.28	8.42	18.46	5.28
ArcFace+SVM [41]	10.82	15.47	12.43	5.38	7.45	4.78	3.87	6.12	3.28
Ours	**8.75**	**12.58**	**8.51**	**4.69**	**5.74**	**2.59**	**3.11**	**5.35**	**2.24**

Table 1. D-EER%, BPCER@APCER=5%, and BPCER@APCER=10% for the differential morph detection.

distance between face images x_i and x_j is defined as:

$$D = \Phi(f(x_i), f(x_j)) + \beta_a \Phi(a(x_i), a(x_j)) + \beta_g \Phi(g(x_i), g(x_j)), \quad (13)$$

where β_a and β_g are the scaling parameters used for decision making. We employ classical texture descriptors, BSIF [22] and LBP [27], with an SVM classifier. The LBP feature descriptors are extracted according to the original LBP image patches of 3×3. The resulting feature vector is then a normalized histogram of size 256, which encompasses all potential values of the LBP binary code. BSIF feature vectors are conducted on a filter size of 3×3 and 8 bits. The filters utilized for BSIF are pre-learned Independent Component Analysis (ICA) filters [20] that are utilized by the original BSIF paper to construct normalized histogram for each image. The feature vectors are then inputted to an SVM with an RBF kernel for classification. For all classical baseline models the difference between the feature representation of the image in question and the feature representation of the trusted image is fed to an SVM classifier.

In addition, we employ LM-Dlib [9, 26] as a model for the landmark displacement measure. In this framework, the distance between landmarks extracted by Dlib [26] are fed to an SVM. For deep models, the distance between the representations in the embedding domain is considered as the decision criteria. For all the model, the default parameters presented in the original papers are considered. It is worth mentioning that in this experiments we do not consider the prior knowledge on which of the images in the pair fed to the recognition framework is the trusted image. On the other hand, in Table 3, we assume that the differential morph detection framework is provided with the information regarding the trusted image.

For each the datasets, 10% of the training set is considered as the validation set. Then, the parameters to train the framework are selected based on the experiments described in Table 4 and Figure 5. Table 1 presents the performance of the proposed framework in comparison with four deep learning and three classical differential morph detection frameworks. In addition to outperforming the baseline models on all the datasets, the proposed framework outperforms the baseline models by a wide margin on the MorGAN dataset, which can be contributed to the disentanglement of landmark and appearance representations.

Train	Test	Algorithm	D-EER	5%	10%
MorGAN	VISAPP17	LM-Dlib [9, 26]	23.74	51.42	38.67
		BSIF+SVM [22]	19.21	51.25	39.41
		ArcFace+SVM [41]	11.67	22.36	14.86
		Ours	**8.55**	**12.68**	**8.57**
	AMSL	LM-Dlib [9, 26]	20.67	44.28	32.15
		BSIF+SVM [22]	17.27	38.54	24.71
		ArcFace+SVM [41]	10.48	22.49	14.90
		Ours	**7.95**	**11.26**	**8.81**
VISAPP17	MorGAN	LM-Dlib [9, 26]	16.82	38.54	24.8
		BSIF+SVM [22]	**13.52**	15.3	14.79
		ArcFace+SVM [41]	15.75	32.58	22.36
		Ours	13.85	**12.32**	**8.74**
	AMSL	LM-Dlib [9, 26]	18.83	38.86	24.78
		BSIF+SVM [22]	16.92	38.84	24.64
		ArcFace+SVM [41]	8.27	9.63	5.28
		Ours	**5.38**	**3.47**	**2.38**
AMSL	MorGAN	LM-Dlib [9, 26]	16.24	30.94	19.28
		BSIF+SVM [22]	**13.84**	**25.35**	**14.82**
		ArcFace+SVM [41]	16.34	38.62	24.51
		Ours	14.21	28.58	18.51
	VISAPP17	LM-Dlib [9, 26]	20.55	62.21	38.42
		BSIF+SVM [22]	20.36	51.28	32.95
		ArcFace+SVM [41]	10.65	14.36	9.81
		Ours	**5.21**	**8.26**	**4.17**

Table 2. Cross-dataset performance for differential morph detection: D-EER%, BPCER@APCER=5%, and BPCER@APCER=10%.

In Table 2, we study the performance of the networks trained on the training portion of one morph dataset and tested on the other datasets. As presented in this table, while outperforming the other models, the proposed framework provides high cross-dataset performance be-

Dataset	MorGAN			VISAPP17			AMSL		
	D-EER	5%	10%	D-EER	5%	10%	D-EER	5%	10%
LM-Dlib [9, 26]	8.14	10.67	7.83	15.67	22.87	20.32	11.67	16.98	14.63
BSIF+SVM [22]	6.07	9.15	4.63	13.87	23.53	20.12	10.53	16.53	13.86
LBP+SVM [27]	7.47	9.23	4.71	15.21	20.64	18.74	12.21	17.11	12.81
FaceNet [42]	8.11	14.52	7.59	7.32	24.54	5.21	7.46	22.12	5.17
ArcFace [11]	7.58	9.64	4.08	6.45	14.78	5.02	5.36	10.46	4.87
FaceNet+SVM	7.23	12.46	5.22	6.37	26.46	6.28	8.42	18.46	5.28
ArcFace+SVM [41]	5.35	6.71	3.50	4.52	5.98	4.05	3.27	5.56	2.69
Ours	4.71	5.32	3.85	3.74	4.91	2.17	2.82	4.97	2.82
Ours*	**4.06**	**5.04**	**3.42**	**3.45**	**4.25**	**1.85**	**2.36**	**4.16**	**1.47**

Table 3. The differential morph detection performance on three datasets, when the trusted image is known to the detection framework: D-EER%, BPCER@APCER=5%, and BPCER@APCER=10%.

tween VISAPP17 and AMSL. In addition, the proposed framework provides D-EER of 8.55% and 7.95% for cross-dataset performance on the network trained on MorGAN and tested on VISAPP17 and AMSL datasets, respectively. On the other hand, BSIF+SVM outperforms the other algorithms when testing the network trained on other two datasets and tested on MorGAN, which illustrates the same trend as the results provided in [10].

Table 3 studies the effect of the trusted images being known to the detection framework. For the baseline models, rather than comparing the representations of the trusted images and the image in question, the representation of the image in question is subtracted from the representation of the trusted image before feeding the difference to the SVM. For the proposed framework, we consider an additional algorithm, denoted as "Ours*", in which two dedicated instances of the framework are constructed for trusted images and images in question. In this algorithm, which outperforms the algorithm for which only one instance of the network is considered, we only train the network dedicated to the images in question. Table 4 provides the performance for the proposed framework on the validation sets when the scaling parameters in making the decision vary in Equation 13. As presented in this table, morph images constructed using landmark displacement are better detected for higher weights given to $g(x)$, while the MorGAN samples are best detected when $g(x)$ and $a(x)$ are given similar weights. In addition, Figure 5 provides the performance for three datasets when variance of the normal distribution to generate δ_l samples in Equation 1 varies from 0 to 6.

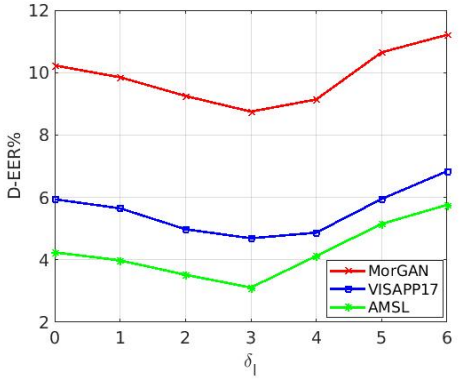

Figure 5. D-EER% for different variances of δ_l values in Equation 1.

	MorGAN	VISAPP17	AMSL
(4,1)	10.91	6.97	4.72
(3,1)	9.64	6.57	4.12
(2,2)	**8.75**	5.84	3.83
(1,3)	10.32	**4.69**	**3.11**
(1,4)	10.89	5.12	3.54

Table 4. The D-EER% for differential morph detection performance considering different scaling values (β_a and β_g) in Equation 13.

5. Conclusions

In this paper, we presented a novel differential morph detection framework which benefits from disentangling landmark representation and appearance representation in an embedding space. These two representations which are disentangled but complementary, are constructed using a disentanglement network trained using triplets of face images. Each triplet consists of two real images and an intermediate image which inherits the landmarks from one image and the appearance from the other image. We demonstrated that appearance and landmark disentanglement can be boosted using contrastive representations for each disentangled representation. This property provides the possibility of accurate differential morph detection, using distances in landmark, appearance, and ID domains. The performance of the proposed framework is studied using three morph datasets constructed with different methodologies.

References

[1] Face research lab london set: https://figshare.com/articles/face_research_lab_london_set/5047666.

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545, 2019.

[3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540, 2018.

[4] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585, 1989.

[5] Fatih Cakir, Kun He, Sarah Adel Bargal, and Stan Sclaroff. Mihash: Online hashing with mutual information. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 437–445, 2017.

[6] Fatih Cakir, Kun He, Sarah Adel Bargal, and Stan Sclaroff. Hashing with mutual information. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2424–2437, 2019.

[7] Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, and Nasser Nasrabadi. Fast geometrically-perturbed adversarial faces. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1979–1988, 2019.

[8] Ali Dabouei, Fariborz Taherkhani, Sobhan Soleymani, Jeremy Dawson, and Nasser Nasrabadi. Boosting deep face recognition via disentangling appearance and geometry. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 320–329, 2020.

[9] Naser Damer, Viola Boller, Yaza Wainakh, Fadi Boutros, Philipp Terhörst, Andreas Braun, and Arjan Kuijper. Detecting face morphing attacks by analyzing the directed distances of facial landmarks shifts. In *German Conference on Pattern Recognition*, pages 518–534, 2018.

[10] Naser Damer, Alexandra Mosegui Saladie, Andreas Braun, and Arjan Kuijper. MorGAN: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10, 2018.

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

[12] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

[13] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. The magic passport. In *IEEE International Joint Conference on Biometrics*, pages 1–7, 2014.

[14] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. Face demorphing. *IEEE Transactions on Information Forensics and Security*, 13(4):1008–1017, 2017.

[15] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. Face demorphing in the presence of facial appearance variations. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2365–2369. IEEE, 2018.

[16] Hiren J Galiyawala, Mehul S Raval, and Anand Laddha. Person retrieval in surveillance videos using deep soft biometrics. In *Deep Biometrics*, pages 191–214. 2020.

[17] Francis Galton. Personal identification and description. *Journal of anthropological institute of Great Britain and Ireland*, pages 177–191, 1889.

[18] Ester Gonzalez-Sosa, Julian Fierrez, Ruben Vera-Rodriguez, and Fernando Alonso-Fernandez. Facial soft biometrics for recognition in the wild: Recent works, annotation, and cots evaluation. *IEEE Transactions on Information Forensics and Security*, 13(8):2001–2014, 2018.

[19] 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.

[20] Aapo Hyvärinen, Jarmo Hurri, and Patrick O Hoyer. *Natural image statistics: A probabilistic approach to early computational vision.*, volume 39. Springer Science & Business Media, 2009.

[21] Seyed Mehdi Iranmanesh, Ali Dabouei, Sobhan Soleymani, Hadi Kazemi, and Nasser Nasrabadi. Robust facial landmark detection via aggregation on geometrically manipulated faces. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 330–340, 2020.

[22] Juho Kannala and Esa Rahtu. Bsif: Binarized statistical image features. In *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*, pages 1363–1366, 2012.

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[24] Hadi Kazemi, Seyed Mehdi Iranmanesh, and Nasser Nasrabadi. Style and content disentanglement in generative adversarial networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 848–856, 2019.

[25] Mete Kemertas, Leila Pishdad, Konstantinos G Derpanis, and Afsaneh Fazly. Rankmi: A mutual information maximizing ranking loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14362–14371, 2020.

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

[27] Shengcai Liao, Xiangxin Zhu, Zhen Lei, Lun Zhang, and Stan Z Li. Learning multi-scale block local binary patterns for face recognition. In *International Conference on Biometrics*, pages 828–837, 2007.

[28] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding

for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[29] Andrey Makrushin, Tom Neubert, and Jana Dittmann. Automatic generation and detection of visually faultless facial morphs. In *International Conference on Computer Vision Theory and Applications*, volume 7, pages 39–50, 2017.

[30] Tom Neubert, Andrey Makrushin, Mario Hildebrandt, Christian Kraetzer, and Jana Dittmann. Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images. *IET Biometrics*, 7(4):325–332, 2018.

[31] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279, 2016.

[32] Fei Peng, Le-Bing Zhang, and Min Long. FD-GAN: Face de-morphing generative adversarial network for restoring accomplice's facial image. *IEEE Access*, 7:75122–75131, 2019.

[33] Kiran Raja, Sushma Venkatesh, RB Christoph Busch, et al. Transferable deep-cnn features for detecting digital and print-scanned morphed face images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 10–18, 2017.

[34] Raghavendra Ramachandra, Sushma Venkatesh, Kiran Raja, and Christoph Busch. Towards making morphing attack detection robust using hybrid scale-space colour texture features. In *2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA)*, pages 1–8, 2019.

[35] Ulrich Scherhag, Dhanesh Budhrani, Marta Gomez-Barrero, and Christoph Busch. Detecting morphed face images using facial landmarks. In *International Conference on Image and Signal Processing*, pages 444–452, 2018.

[36] Ulrich Scherhag, Andreas Nautsch, Christian Rathgeb, Marta Gomez-Barrero, Raymond NJ Veldhuis, Luuk Spreeuwers, Maikel Schils, Davide Maltoni, Patrick Grother, Sebastien Marcel, et al. Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting. In *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7, 2017.

[37] Ulrich Scherhag, Ramachandra Raghavendra, Kiran B Raja, Marta Gomez-Barrero, Christian Rathgeb, and Christoph Busch. On the vulnerability of face recognition systems towards morphed face attacks. In *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2017.

[38] Ulrich Scherhag, Christian Rathgeb, and Christoph Busch. Morph deterction from single face image: A multi-algorithm fusion approach. In *Proceedings of the 2018 2nd International Conference on Biometric Engineering and Applications*, pages 6–12, 2018.

[39] Ulrich Scherhag, Christian Rathgeb, and Christoph Busch. Towards detection of morphed face images in electronic travel documents. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 187–192, 2018.

[40] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Ralph Breithaupt, and Christoph Busch. Face recognition systems under morphing attacks: A survey. *IEEE Access*, 7:23012–23026, 2019.

[41] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, and Christoph Busch. Deep face representations for differential morphing attack detection. *arXiv preprint arXiv:2001.01202*, 2020.

[42] 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.

[43] Clemens Seibold, Wojciech Samek, Anna Hilsmann, and Peter Eisert. Detection of face morphing attacks by deep learning. In *International Workshop on Digital Watermarking*, pages 107–120, 2017.

[44] Clemens Seibold, Wojciech Samek, Anna Hilsmann, and Peter Eisert. Accurate and robust neural networks for security related applications exampled by face morphing attacks. *arXiv preprint arXiv:1806.04265*, 2018.

[45] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Adversarial generative nets: Neural network attacks on state-of-the-art face recognition. *arXiv preprint arXiv:1801.00349*, 2(3), 2017.

[46] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–665, 2018.

[47] Luuk Spreeuwers, Maikel Schils, and Raymond Veldhuis. Towards robust evaluation of face morphing detection. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1027–1031. IEEE, 2018.

[48] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.

[49] Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, Luuk Spreeuwers, Raymond Veldhuis, and Christoph Busch. Detecting morphed face attacks using residual noise from deep multi-scale context aggregation network. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 280–289, 2020.

[50] Sushma Venkatesh, Haoyu Zhang, Raghavendra Ramachandra, Kiran Raja, Naser Damer, and Christoph Busch. Can GAN generated morphs threaten face recognition systems equally as landmark based morphs?-vulnerability and detection. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2020.

[51] Lukasz Wandzik, Gerald Kaeding, and Raul Vicente Garcia. Morphing detection using a general-purpose face recognition system. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1012–1016, 2018.

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

[53] George Wolberg. Image morphing: a survey. *The visual computer*, 14(8-9):360–372, 1998.

[54] Andreas Wolf. Portrait quality (reference facial images for mrtd). *Version: 0.06 ICAO, Published by authority of the Secretary General*, 2016.

[55] Xianglei Xing, Ruiqi Gao, Tian Han, Song-Chun Zhu, and Ying Nian Wu. Deformable generator network: Unsupervised disentanglement of appearance and geometry. *arXiv preprint arXiv:1806.06298*, 2018.

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