

Towards Large-Pose Face Frontalization in the Wild

Xi Yin^{†*}, Xiang Yu[‡], Kihyuk Sohn[‡], Xiaoming Liu[†] and Manmohan Chandraker^{§‡}

[†]Michigan State University

[§]University of California, San Diego

[‡] NEC Laboratories America

{yinxil, liuxm}@cse.msu.edu, {xiangyu, ksohn, manu}@nec-labs.com

Abstract

Despite recent advances in face recognition using deep learning, severe accuracy drops are observed for large pose variations in unconstrained environments. Learning pose-invariant features is one solution, but needs expensively labeled large-scale data and carefully designed feature learning algorithms. In this work, we focus on frontalizing faces in the wild under various head poses, including extreme profile views. We propose a novel deep 3D Morphable Model (3DMM) conditioned Face Frontalization Generative Adversarial Network (GAN), termed as FF-GAN, to generate neutral head pose face images. Our framework differs from both traditional GANs and 3DMM based modeling. Incorporating 3DMM into the GAN structure provides shape and appearance priors for fast convergence with less training data, while also supporting end-to-end training. The 3DMM-conditioned GAN employs not only the discriminator and generator loss but also a new masked symmetry loss to retain visual quality under occlusions, besides an identity loss to recover high frequency information. Experiments on face recognition, landmark localization and 3D reconstruction consistently show the advantage of our frontalization method on faces in the wild datasets.¹

1. Introduction

Frontalization of face images observed from extreme viewpoints is a problem of fundamental interest in both human and machine facial processing and recognition. Indeed, while humans are innately skilled at face recognition, newborns do not perform better than chance on recognition from profile views [31], although this ability seems to develop rapidly within a few months of birth [8]. Similarly, dealing with profile views of faces remains an enduring challenge

^{*}This work was supported by a research gift from NEC Laboratories America to Michigan State University.

¹Detail results and resource can be referred to: <http://cvlab.cse.msu.edu/project-face-frontalization.html>.

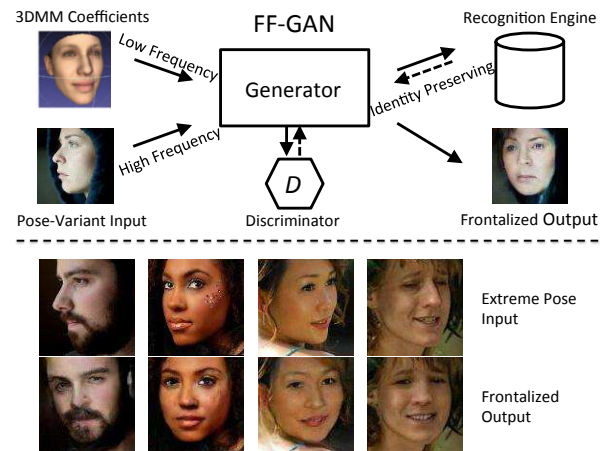


Figure 1: The proposed FF-GAN framework. Given a non-frontal face image as input, the generator produces a high-quality frontal face. Learned 3DMM coefficients provide global pose and low frequency information, while the input image injects high frequency local information. A discriminator distinguishes generated faces against real ones, where high-quality frontal faces are considered real. A face recognition engine is used to preserve identity information. The output is a high quality frontal face that retains identity.

for computer vision too. The advent of deep convolutional neural networks (CNNs) has led to large strides in face recognition, with verification accuracy surpassing human levels [28, 27] on datasets such as LFW. But even representations learned with state-of-the-art CNNs suffer for profile views, with severe accuracy drops reported in recent studies [16, 26]. Besides aiding recognition, frontalization of faces is also a problem of independent interest, with potential applications such as face editing, accessorizing and creation of models for use in virtual and augmented reality.

In recent years, CNNs have leveraged the availability of large-scale face datasets to also learn pose-invariant representations for recognition, using either a joint single framework [18] or multi-view pose-specific networks [46, 16]. Early works on face frontalization in computer vision rely

on frameworks inspired by computer graphics. The well-known 3D Morphable Model (3DMM) [1] explicitly models facial shape and appearance to match an input image as closely as possible. Subsequently, the recovered shape and appearance can be used to generate a face image under novel viewpoints. Many 3D face reconstruction methods [22, 43, 23] build upon this direction by improving speed or accuracy. Deep learning has made inroads into data-driven estimation of 3DMM models too [46, 16, 15], circumventing some drawbacks of early methods such as over-reliance on the accuracy of 2D landmark localization [40, 41]. Due to restricted Gaussian assumptions and nature of losses used, insufficient representation ability for facial appearance prevents such deep models from producing outputs of high quality. While inpainting methods such as [44] attempt to minimize the impact on quality due to self-occlusions, they still do not retain identity information.

In this paper, we propose a novel generative adversarial network framework FF-GAN that incorporates elements from both deep 3DMM and face recognition CNNs to achieve high-quality and identity-preserving frontalization, using a single input image that can be a profile view up to 90° . Our framework targets more on face-in-the-wild conditions which show more challenge on illumination, head pose variation, self-occlusion and so on. To the best of our knowledge, this is the first face frontalization work that handles pose variations at such extreme poses.

Noting the challenge of purely data-driven face frontalization, Section 3.1 proposes to enhance the input to the GAN with 3DMM coefficients estimated by a deep network. This provides a useful prior to regularize the frontalization, however, it is well known that deep 3DMM reconstructions are limited in their ability to retain high-frequency information. Thus, Section 3.2 proposes a method to combine 3DMM coefficients with the input image to generate an image that maintains both global pose accuracy and retains local information present in the input image. In particular, the generator in our GAN produces a frontal image based on a reconstruction loss, a smoothness loss and a novel symmetry-enforcing loss. The aim of the generator is to fool the discriminator, presented in Section 3.3, into being unable to distinguish the generated frontal image from a real one. However, neither the 3DMM that loses high-frequency information, nor the GAN that only aligns domain-level distributions, suffice to preserve identity information in the generated image. To retain identity information, Section 3.4 proposes to use a recognition engine to align the feature representations of the generated image with the input. A balanced training with all the above objectives results in high-quality frontalized faces that preserve identity.

We extensively evaluate our framework on several well-known datasets including Multi-PIE, AFLW, LFW, and IJB-A. In particular, Section 5 demonstrates that the face verifi-

cation accuracy on LFW dataset that uses information from our frontalized outputs exceeds previous state of the art. We observe even larger improvements on the Multi-PIE dataset especially for the viewpoints beyond 45° and being the sole method among recent works to produce high recognition accuracies for 75° - 90° . We present ablation studies to analyze the effect of each module and several qualitative results to visualize the quality of our frontalization.

To summarize, our key contributions are:

- A novel GAN-based end-to-end deep framework to achieve face frontalization even for extreme viewpoints.
- A deep 3DMM model to provide shape and appearance regularization beyond the training data.
- Effective symmetry-based loss and smoothness regularization that lead to generation of high-quality images.
- Use of a deep face recognition CNN to enforce that generated faces satisfy identity-preservation, besides realism and frontalization.
- Consistent improvements on several datasets across multiple tasks, such as face recognition, landmark localization and 3D reconstruction.

2. Related Work

Face Frontalization Synthesizing a frontal view of a face from a single image with large pose variation is very challenging because recovering the 3D information from 2D projections is ambiguous and there exists self-occlusion. A straightforward method is to build 3D models for faces and directly rotate the 3D face models. Seminal works date back to the 3D Morphable Model (3DMM) [1], which models both the shape and appearance as PCA spaces. Such 3DMM fitting helps boost 3D face reconstruction [22] and 3D landmark localization [14, 43] performance. Hassner et al. [12] apply a shared 3D shape model combined with input images to produce the frontalized appearance. Ferrari et al. [9] use 3DMM to fit to the input image and search for the dense point correspondence to complete the occlusion region. Zhu et al. [44] provide a high fidelity pose and expression normalization method based on 3DMM.

Besides model-based methods, Sagonas et al. [24] propose frontalization as a low-rank constraint optimization problem to target landmark localization. Some deep learning based methods also show promising performance in terms of pose rectification. In [36], a recurrent transform unit is proposed to synthesize discrete 3D views. A concatenated network structure is applied in [38] to rotate the face, where the output is regularized by the image level reconstruction loss. In [5, 2], a perception loss is designed to supervise the generator training. Our method is also based on deep learning. We incorporate 3DMM into a deep network and propose a novel GAN based framework to jointly utilize the geometric prior from 3DMM and high frequency information from the original image. A discriminator and a face

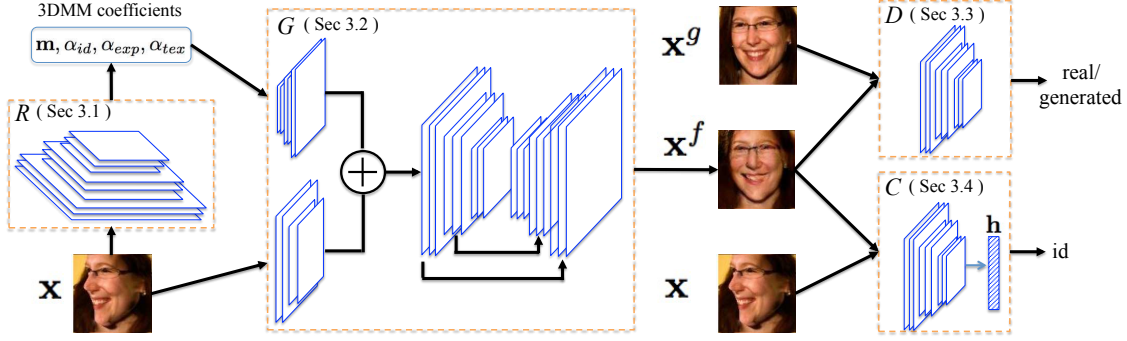


Figure 2: The proposed FF-GAN for large-pose face frontalization. R is the reconstruction module for 3DMM coefficients estimation. G is the generation module to synthesize a frontal face. D is the discrimination module to make real or generated decision. C is the recognition module for identity classification.

recognition engine are proposed to regularize the generator to preserve identity.

Generative Adversarial Networks Introduced by Goodfellow et al. [10], the Generative Adversarial Network (GAN) maps from a source data distribution to a target distribution using a minimax optimization over a generator and a discriminator. GANs have been used for a wide range of applications including image generation [6, 7, 29], 3D object generation [33], etc. Deep Convolutional GAN (DC-GAN) [21] extended the original GAN multi-layer perceptron network into convolutional structures. Many methods favor conditional settings by introducing latent factors to disentangle the objective space and thereby achieve better synthesis. For instance, Info-GAN [4] employs the latent code for information loss to regularize the generative network.

A recent method, DR-GAN [30], has been proposed concurrent to this work. It also uses a recognition engine to regularize the identity, while using a pose code as input to the encoder to generate an image with specific pose. Instead of explicitly injecting the latent code, our framework learns the shape and appearance code from a differentiable 3DMM deep network, which supports the flavor of end-to-end joint optimization for the GAN. Unlike our framework, it is the encoder in [30] that is required to be identity-preserving, which suffices for reconstruction, but results in loss of spatial and high-frequency information crucial for image generation.

Pose-Invariant Feature Representation While face frontalization may be considered an image-level pose-invariant representation, feature representations invariant to pose have also been a mainstay of face recognition. Early works employ Canonical Correlation Analysis (CCA) to analyze the commonality among pose-variant samples. Recently, deep learning based methods consider several aspects, such as multiview perception layers [46], to learn a model separating identity from viewpoints. Feature pooling across different poses [16] is also proposed to allow a single network structure for multiple pose inputs. Pose-invariant feature disentanglement [20] or identity preservation [45, 39] methods aim to factorize out the non-identity part with care-

fully designed networks. Some other methods focus on fusing information at the feature level [3] or distance level [18]. Our method is mostly related to identity preserving methods, in which we apply a face recognition engine as an additional discriminator to guide the generator for better synthesis, while it updates itself towards better discriminative capability for identity classification.

3. Proposed Approach

The mainstay of FF-GAN is a generative adversarial network that consists of a generator G and a discriminator D . G takes a non-frontal face as input to generate a frontal output, while D attempts to classify it as a real frontal image or a generated one. Additionally, we include a face recognition engine C that regularizes the generator output to preserve identity features. A key component is a deep 3DMM model R that provides shape and appearance priors to the GAN that play a crucial role in alleviating the difficulty of large pose face frontalization. Figure 2 summarizes our framework.

Let $\mathbb{D} = \{\mathbf{x}_i, \mathbf{x}_i^g, \mathbf{p}_i^g, y_i\}_{i=1}^N$ be the training set with N samples, with each sample consisting of an input image \mathbf{x}_i with arbitrary pose, a corresponding ground truth frontal face \mathbf{x}_i^g , the ground truth 3DMM coefficients \mathbf{p}_i^g and the identity label y_i . We henceforth omit the sample index i for clarity.

3.1. Reconstruction Module

Frontalization from extreme pose variation is a challenging problem. While a purely data-driven approach might be possible given sufficient data and an appropriate training regimen, however it is non-trivial. Therefore, we propose to impose a prior on the generation process, in the form of a 3D Morphable Model (3DMM) [1]. This reduces the training complexity and leads to better empirical performance with limited data.

Recall that 3DMM represents faces in the PCA space:

$$\begin{aligned} \mathbf{S} &= \bar{\mathbf{S}} + \mathbf{A}_{id}\alpha_{id} + \mathbf{A}_{exp}\alpha_{exp}, \\ \mathbf{T} &= \bar{\mathbf{T}} + \mathbf{A}_{tex}\alpha_{tex}, \end{aligned} \quad (1)$$

where \mathbf{S} are the 3D shape coordinates computed as the linear combination of the mean shape $\bar{\mathbf{S}}$, the shape basis \mathbf{A}_{id} , and

the expression basis \mathbf{A}_{exp} , while \mathbf{T} is the texture that is the linear combination of the mean texture $\bar{\mathbf{T}}$ and the texture basis \mathbf{A}_{tex} . The coefficients $\{\alpha_{id}, \alpha_{exp}, \alpha_{tex}\}$ define a unique 3D face.

Previous work [14, 43] applies 3DMM for face alignment where a weak perspective projection model is used to project the 3D shape into 2D space. Similar to [14], we optimize a projection matrix $\mathbf{m} \in \mathbb{R}^{2 \times 4}$ based on pitch, yaw, roll, scale and 2D translations to represent the pose of an input face image. Let $\mathbf{p} = \{\mathbf{m}, \alpha_{id}, \alpha_{exp}, \alpha_{tex}\}$ denotes the 3DMM coefficients. The target of the reconstruction module R is to estimate $\mathbf{p} = R(\mathbf{x})$, given an input image \mathbf{x} . Since the intent is for R to also be trainable with the rest of the framework, we use a CNN model based on CASIA-Net [37] for this regression task. We apply z -score normalization to each dimension of the parameters before training. A weighted parameter distance cost similar to [43] is used:

$$\min_{\mathbf{p}} L_R = (\mathbf{p} - \mathbf{p}^g)^\top \mathbf{W} (\mathbf{p} - \mathbf{p}^g), \quad (2)$$

where \mathbf{W} is the importance matrix whose diagonal is the weight of each parameter.

3.2. Generation Module

The pose estimate obtained from 3DMM is quite accurate, however, frontalization using it leads to loss of high frequency details present in the original image. This is understandable, since a low-dimensional PCA representation can preserve most of the energy in lower frequency components. Thus, we use a generative model that relies on 3DMM coefficients \mathbf{p} and the input image \mathbf{x} to recover a frontal face that preserves both the low and high frequency components.

In Figure 2, features from the two inputs to the generator G are fused through an encoder-decoder network to synthesize a frontal face $\mathbf{x}^f = G(\mathbf{x}, \mathbf{p})$. To penalize the output from ground truth \mathbf{x}^g , the straight-forward objective is the reconstruction loss that aims at reconstructing the ground truth with minimal error:

$$L_{G_{rec}} = \|G(\mathbf{x}, \mathbf{p}) - \mathbf{x}^g\|_1. \quad (3)$$

Since an L_2 loss empirically leads to blurry output, we use an L_1 loss to better preserve high frequency signals. At the beginning of training, the reconstruction loss harms the overall process since the generation is far from frontalized, so the reconstruction loss operates on a poor set of correspondences. Thus, the weight for reconstruction loss should be set in accordance to the training stage. The details of tuning the weight are discussed in Section 4.

To reduce block artifacts, we use a spatial total variation loss to encourage smoothness in the generator output:

$$L_{G_{tv}} = \frac{1}{|\Omega|} \int_{\Omega} |\nabla G(\mathbf{x}, \mathbf{p})| du, \quad (4)$$

where $|\nabla G|$ is the image gradient, $u \in \mathbb{R}^2$ is the two dimensional coordinate increment, Ω is the image region and $|\Omega|$ is the area normalization factor.

Based on the observation that human faces share self-similarity across left and right halves, we explicitly impose a symmetry loss. We recover a frontalized 2D projected silhouette, \mathcal{M} , from the frontalized 3DMM model indicating the visible parts of the face. The mask \mathcal{M} is binary, with nonzero values indicating visible regions and zero otherwise. Similar masking constraint is shown effective in recent work [19]. By horizontally flipping the face, we achieve another mask \mathcal{M}_{flip} . We demand that generated frontal images for the original input image and its flipped version should be similar within their respective masks:

$$L_{G_{sym}} = \|\mathcal{M} \odot G(\mathbf{x}, \mathbf{p}) - \mathcal{M} \odot G(\mathbf{x}_{flip}, \mathbf{p}_{flip})\|_2 + \|\mathcal{M}_{flip} \odot G(\mathbf{x}, \mathbf{p}) - \mathcal{M}_{flip} \odot G(\mathbf{x}_{flip}, \mathbf{p}_{flip})\|_2. \quad (5)$$

Here, \mathbf{x}_{flip} is the horizontal flipped image for input \mathbf{x} , \mathbf{p}_{flip} are the 3DMM coefficients for \mathbf{x}_{flip} and \odot denotes element-wise multiplication. Note that the role of the mask is to focus on the visible face portion, rather than invisible face portion and background.

3.3. Discrimination Module

Generative Adversarial Networks [10], formulated as a two-player game between a generator and a discriminator, have been widely used for image generation. In this work, the generator G synthesizes a frontal face image \mathbf{x}^f and the discriminator D distinguishes between the generated face from the frontal face \mathbf{x}^g . Note that in a conventional GAN, all images used for training are considered as real samples. However, we limit “real” faces to frontal views only. Thus, G must generate both realistic and frontal face images.

Our D consists of five convolutional layers and one linear layer that generates a 2D vector with each dimension representing the probability of the input to be real or generated. During training, D is updated with two batches of samples in each iteration. The following objective is minimized:

$$\min_D L_D = -\mathbb{E}_{\mathbf{x}^g \in \mathcal{R}} \log(D(\mathbf{x}^g)) - \mathbb{E}_{\mathbf{x} \in \mathcal{K}} \log(1 - D(G(\mathbf{x}, \mathbf{p}))), \quad (6)$$

where \mathcal{R} and \mathcal{K} are the real and generated image sets respectively.

On the other hand, G aims to fool D to classify the generated image to be real with the following loss:

$$L_{G_{gan}} = -\mathbb{E}_{\mathbf{x} \in \mathcal{K}} \log(D(G(\mathbf{x}, \mathbf{p}))). \quad (7)$$

The competition between G and D improves both modules. In the early stages when face images are not fully frontalized, D focuses on the pose of the face to make the real or generated decision, which in turn helps G to generate a frontal face. In the later stages when face images are frontalized, D focuses on subtle details of frontal faces, which guides G to generate a realistic frontal face that is difficult to achieve with the supervisions of (3) and (5) alone.

3.4. Recognition Module

A key challenge in large-pose face frontalization is to preserve the original identity. This is a difficult task due to

self-occlusion in profile faces. The discriminator above can only determine whether the generated image is realistic and in frontal pose, but does not consider whether identity of the input image is retained. Although we have L1, total variation and masked symmetry losses for face generation, they treat each pixel equally that result in the loss of discriminative power for identity features. Therefore, we use a recognition module C to impart correct identity to the generated images.

We use a CASIA-Net structure for the recognition engine C , with a cross-entropy loss for training C to classify image \mathbf{x} with the ground truth identity y :

$$\min_C L_C = \sum_j -y_j \log(C_j(\mathbf{x})), \quad (8)$$

where j is the index of the identity classes. Now, our generator G is regularized by the signal from C to preserve the same identity as the input image. If the identity label of the input image is not available, we regularize the extracted identity features \mathbf{h}^f of the generated image to be similar to those of the input image, denoted as \mathbf{h} . During training, C is updated with real input images to retain discriminative power and the loss from the generated images is back-propagated to update the generator G :

$$L_{G_{id}} = \begin{cases} \sum_j -y_j \log(C_j(G(\mathbf{x}, \mathbf{p}))), & \exists y \\ \|\mathbf{h}^f - \mathbf{h}\|_2^2, & \nexists y. \end{cases} \quad (9)$$

To summarize the framework, the reconstruction module R provides guidance to the frontalization through (2), the discriminator does so through (6) and the recognition engine through (8). Thus, the generator combines all these sources of information to optimize an overall objective function:

$$\min_G L_G = \lambda_{rec} L_{G_{rec}} + \lambda_{tv} L_{G_{tv}} + \lambda_{sym} L_{G_{sym}} + \lambda_{gan} L_{G_{gan}} + \lambda_{id} L_{G_{id}}. \quad (10)$$

The weights above are discussed in Section 4 to illustrate how each component contributes to joint optimization of G .

4. Implementation Details

Our framework consists of mainly four parts as shown in Figure 2, the deep 3DMM reconstructor R , a two-way fused encoder-decoder generator, the real/generated discriminator and a face recognition engine jointly trained for the identity regularization. It is difficult to initialize the overall network from scratch. The generator G is expected to receive the correct 3DMM coefficients, whereas the reconstructor needs to be pre-trained. Our identity regularization also requires correct identity information from the recognizer. Thus, the reconstructor R is pre-trained until we achieve comparable performance for face alignment compared to previous work [43] using 300W-LP [43]. The recognizer is pre-trained using CASIA-WebFace [37] to achieve good verification accuracy on LFW.

The end-to-end joint training is conducted after R and C are ready. Notice that we leave generator G and discriminator D training from scratch simultaneously because

we believe pre-trained G and D do not contribute much to the adversarial training. Good G with poor D will quickly pull G to be poor again and vice versa. Further these two components should also match with each other. Good G may be evaluated poor by a good D as the discriminator may be trained from other sources. As shown in Equation 10, we set up five balance factors to control the loss contribution to the overall loss. The end-to-end training can be divided into three steps. For the first step, λ_{rec} is set to 0 and λ_{id} is set to 0.01, since these two parts are highly related with the mapping from the generated output to the reference input. Typical values for λ_{tv} , λ_{sym} and λ_{gan} are all 1.0. Once the training error of G and D strikes a balance within usually 20 epochs, we change λ_{rec} and λ_{id} to be 1.0 while tuning down λ_{tv} to be 0.5, λ_{sym} to be 0.8, respectively. It takes another 20 epochs to strike a new balance. Note that we fix model C for training these two stages. After that, we relax model C and further fine-tune all the modules jointly. Network structures and further details are referred to: <http://cvlab.cse.msu.edu/project-face-frontalization.html>.

5. Experiments

5.1. Settings and Databases

We evaluate our proposed FF-GAN on a variety of tasks including face frontalization, 3D face reconstruction and face recognition. Frontalization and 3D reconstruction are evaluated qualitatively by comparing the visual quality of the generated images to ground truth. We also report quantitative results on sparse 2D landmark localization accuracy. Face recognition is evaluated quantitatively over several challenging databases. We pre-process the images by applying some state-of-the-art face detection [35] and face alignment [42] algorithms and crop to 100×100 image size across all the databases. The face databases used for training and testing are introduced below.

300W-LP consists of 122, 450 images that are augmented from 300W [25] by the face profiling approach [44], which is designed to generate images with yaw angles ranging from -90° to 90° . We use 300W-LP as our training set by forming image pairs of pose-variant and frontal-view images of the same identity. The estimated 3DMM coefficients provided with dataset are treated as ground truth to train module R .

AFLW2000 is constructed for 3D face alignment evaluation by the same face profiling method applied in 300W-LP. The database includes 3DMM coefficients and augmented 68 landmarks for the first 2,000 images in AFLW. We use this database to evaluate module R for reconstruction.

Multi-PIE consists of 754, 200 images from 337 subjects with large variations in pose, illumination and expression. We select a subset of 301, 600 images with 13 poses, 20 illuminations, neutral expression from all four sessions. The first 200 subjects are used for training and the remaining

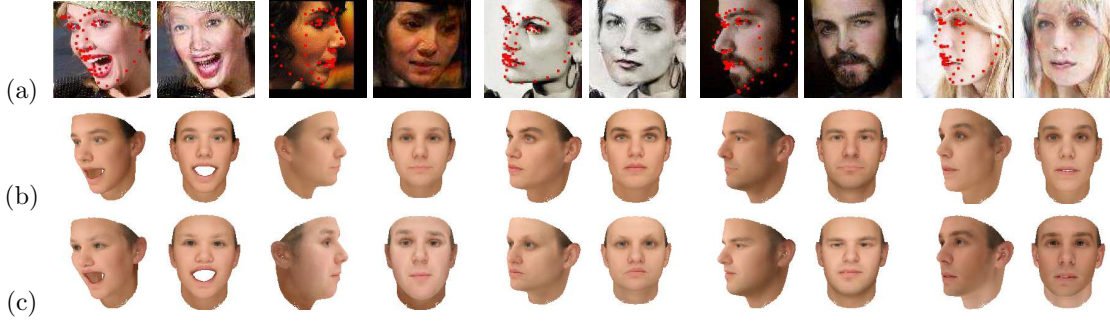


Figure 3: (a) our landmark localization and face frontalization results; (b) our 3DMM estimation; (c) ground truth from [43].

137 subjects for testing, in the setting of [30]. We randomly choose one image for each subject with frontal pose and neutral illumination as gallery and all the rest as probe images.

CASIA consists of 494,414 images of 10,575 subjects where the images of 26 overlapping subjects with IJB-A are removed. It is a widely applied large-scale database for face recognition. We apply it to pre-train and finetune module C .

LFW contains 13,233 images collected from Internet. The verification set consists of 10 folders, each with 300 same-person pairs and 300 different-person pairs. We evaluate face verification performance on frontalized images and compare with previous frontalization algorithms.

IJB-A includes 5,396 images and 20,412 video frames for 500 subjects, which is a challenging with uncontrolled pose variations. Different from previous databases, IJB-A defines face template matching where each template contains a variant amount of images. It consists of 10 folders, each of which being a different partition of the full set. We finetune model C on the training set of each folder and evaluate on the testing set for face verification and identification.

5.2. 3D Reconstruction

FF-GAN borrows prior shape and appearance information from 3DMM to serve as the reference for frontalization. Though we do not specifically optimize for the reconstruction task, it is interesting to see whether our reconstructor is doing a fair job in the 3D reconstruction task.

Figure 3 (a) shows five examples on AFLW2000 for landmark localization and frontalization. Our method localizes the key points correctly and generates realistic frontal faces even for extreme profile inputs. We also quantitatively evaluated the landmark localization performance using the normalized mean square error. Our model R achieves 6.01 normalized mean square error, compared to 5.42 for 3DDFA [43] and 6.12 for SDM [34]. Note that our method achieves competitive performance compared to 3DDFA and SDM, even though those methods are tuned specifically for the localization task. This indicates that our reconstruction module performs well in providing correct geometric information.

Given the input images in (a), we compute the 3DMM coefficients with our model R and generate the 3D geometry and texture using (1), as shown in Figure 3 (b). We observe

that our method effectively preserves shape and identity information in the estimated 3D models, which can even outperform the ground truth provided by 3DDFA. For example, the shape and texture estimations in the last example is more similar to the input while the ground truth clearly shows a male subject rather than a female. Given that the 3DMM coefficients cannot preserve local appearance, we obtain such high frequency information from the input image. Thus, the choice of fusing 3DMM coefficients with the original input is shown to be a reasonable one empirically.

5.3. Face Recognition

One motivation for face frontalization is to see, whether the frontalized images bring in the correct identity information for the self-occluded missing part, and thus boost the performance in face recognition. To verify this, we evaluate our model C on LFW [13], MultiPIE [11], and IJB-A [17] for verification and identification tasks. Features are extracted from model C across all the experiments. Euclidean distance is used as the metric for face matching.

Evaluation on LFW We evaluate the face verification performance on our frontalized images of LFW, compared to previous face frontalization methods. LFW-FF-GAN represents our method to generate frontalized images, LFW-3D is from [12], and LFW-HPEN from [43]. Those collected databases are pre-processed in *the same way* as ours. As shown in Table 1, our method achieves strong results compared to the state-of-the-art methods, which verifies that our frontalization technique better preserves identity information. Figure 4 shows some visual examples. Compared to the state-of-the-art, our method can generate realistic and identity-preserving faces especially for large poses. The facial detail filling technique in [43] relies on a symmetry assumption and may lead to inferior results (1st row, 4th column). In contrast, we introduce a symmetry loss in the training process that generalizes to test images without the need to impose symmetry as a hard constraint for post-processing.

Evaluation on IJB-A We further evaluate our algorithm on IJB-A database. Following prior work [30], we select a subset of well aligned images in each template for face matching. We define our distance metric as the original image pair distance plus the weighted generated image pair

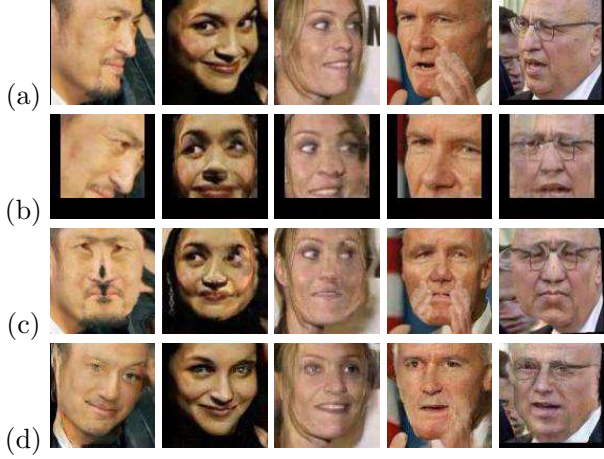


Figure 4: Face frontalization results on LFW. (a) Input; (b) LFW-3D [12]; (c) HPEN [44]; (d) FF-GAN (ours).

Database	ACC(%)	AUC(%)
Ferrari et al. [9]	-	94.29
LFW-3D [12]	93.62 \pm 1.17	98.36 \pm 0.06
LFW-HPEN [44]	96.25 \pm 0.76	99.39 \pm 0.02
FF-GAN (syn.)	96.42 \pm 0.89	99.45 \pm 0.03

Table 1: Face verification accuracy (ACC) and area-under-curve (AUC) results on LFW.

distance. The weights are the confidence score provided by our model D , i.e. $D(G(\mathbf{x}, \mathbf{p}))$. Recall that model D is trained for real or generated classification task, which reflects the quality of the generated images. Obviously, the poorer quality of the generated image, the less likely we take the generated image pair for distance metric. With the fused metric distance, we expect the generated images to provide complimentary information to boost the recognition performance.

Table 2 shows the verification and identification performance. On verification, our method achieves consistently better accuracy compared to the baseline methods. The gap is 6.46% at FAR 0.01 and 11.13% at FAR 0.001, which is a significant improvement. On identification, our fused metric also achieves consistently better result, 4.95% improvement at Rank-1 and 1.66% at Rank-5. As a challenging face database in the wild, large pose variations, complex background and the uncontrolled illumination prevent the compared methods to perform well. Closing one of those variation gaps would lead to large improvement, as evidenced by our face frontalization method in rectifying pose variation.

Evaluation on Multi-PIE Multi-PIE allows for a graded evaluation with respect to pose, illumination, and expression variations. Thus, it is an important database to validate the performance of our methods with respect to prior works on frontalization. The rank-1 identification rate is reported in Table 3. Note that previous works only consider poses within 60°, while our method can handle all pose variation

Method ↓ Metric (%) →	Verification		Identification	
	@FAR=0.01	@FAR=0.001	@Rank-1	@Rank-5
OpenBR [17]	23.6 \pm 0.9	10.4 \pm 1.4	24.6 \pm 1.1	37.5 \pm 0.8
GOTS [17]	40.6 \pm 1.4	19.8 \pm 0.8	44.3 \pm 2.1	59.5 \pm 2.0
Wang et al. [32]	72.9 \pm 3.5	51.0 \pm 6.1	82.2 \pm 2.3	93.1 \pm 1.4
PAM [18]	73.3 \pm 1.8	55.2 \pm 3.2	77.1 \pm 1.6	88.7 \pm 0.9
DCNN [3]	78.7 \pm 4.3	-	85.2 \pm 1.8	93.7 \pm 1.0
DR-GAN [30]	77.4 \pm 2.7	53.9 \pm 4.3	85.5 \pm 1.5	94.7 \pm 1.1
FF-GAN (fuse)	85.2 \pm 1.0	66.3 \pm 3.3	90.2 \pm 0.6	95.4 \pm 0.5

Table 2: Performance comparison on IJB-A database.

	0°	15°	30°	45°	60°	75°	90°	avg1	avg2
Zhu et al. [45]	94.3	90.7	80.7	64.1	45.9	-	-	72.9	-
Zhu et al. [46]	95.7	92.8	83.7	72.9	60.1	-	-	79.3	-
Yim et al. [38]	99.5	95.0	88.5	79.9	61.9	-	-	83.3	-
DR-GAN [30]	97.0	94.0	90.1	86.2	83.2	-	-	89.2	-
FF-GAN (ori.)	95.6	95.0	92.6	89.8	84.3	75.9	58.2	91.5	84.5
FF-GAN (syn.)	94.9	86.1	82.0	77.0	70.8	62.1	46.2	82.1	74.2
FF-GAN (fuse)	95.7	94.6	92.5	89.7	85.2	77.2	61.2	91.6	85.2

Table 3: Performance comparison on Multi-PIE database. Avg1 and avg2 are the average accuracy in 0-60° and 0-90° respectively.

including profile views at 90°. The results suggest that when pose variation is within 15°, which is near frontal, our method is competitive to state-of-the-art methods. But when the pose is 30° or larger, our method demonstrates significant advantages over all the other methods. Although the recognition rate of the synthetic images performs worse than the original images, the fused results perform better than the original images, especially on large-pose face images.

5.4. Qualitative Results

Figure 5 shows some visual results for unseen images on Multi-PIE, AFLW, and IJB-A. The input images are of medium to large pose and under a large variation of race, age, expression, and lighting conditions. However, FF-GAN can still generate realistic and identity-preserved frontal faces.

5.5. Ablation Study on Multi-PIE

FF-GAN consists of four modules $\mathbb{M} = \{R, G, C, D\}$. Our generator G is the key component for image synthesis, which cannot be removed. We train three partial variants by removing each of the remaining modules, which results in $\mathbb{M} \setminus \{C\}$, $\mathbb{M} \setminus \{D\}$, and $\mathbb{M} \setminus \{R\}$. Further, we train another three variants by removing each of the three loss functions (including G_{id} , G_{tv} , G_{sym}) applied on the generated images, resulting in $\mathbb{M} \setminus \{G_{id}\}$, $\mathbb{M} \setminus \{G_{tv}\}$, and $\mathbb{M} \setminus \{G_{sym}\}$. We keep the training process and all hyper-parameters the same and explore how the performances of those models differ.

Figure 6 shows visual comparisons between the proposed framework and its incomplete variants. Our method is visually better than those variants, across all different poses, which suggests that each component in our model is essential for face frontalization. Without the recognizer C , it is hard to preserve identity especially on large poses. Without the discriminator D , the generated images are blurry without much high-frequency identity information. Without the reconstructor R , there are artifacts on the generated faces,

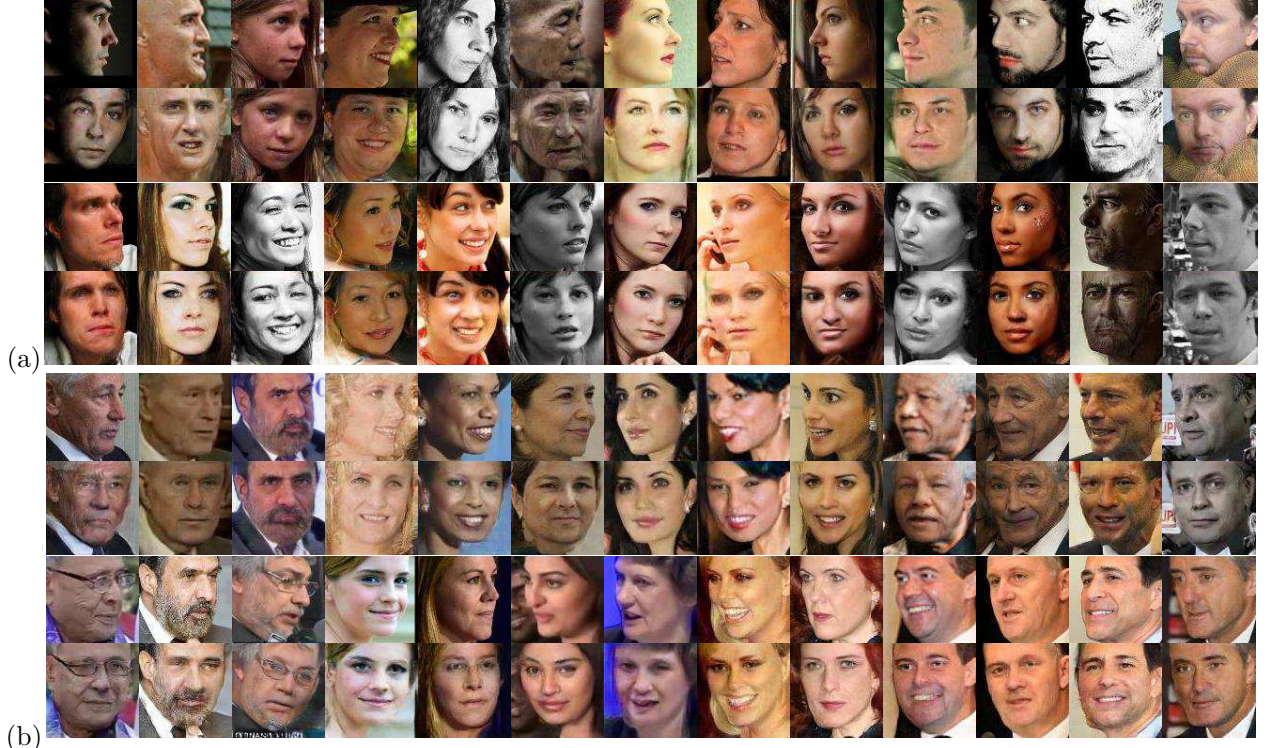


Figure 5: Visual results of our method on large-pose cases from different databases. (a) AFLW2000, (b) IJB-A.

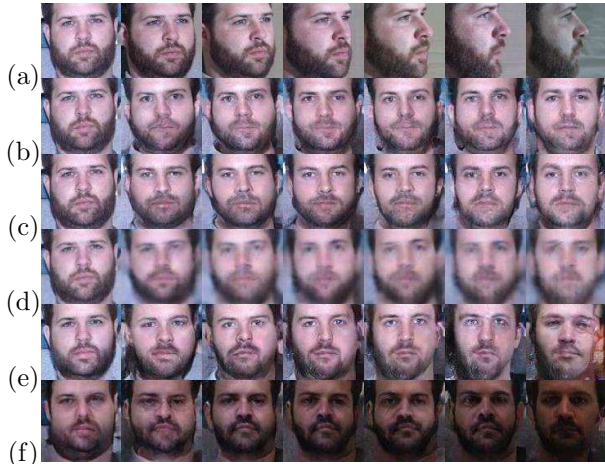


Figure 6: Ablation study results. (a) input images. (b) \mathbb{M} (ours). (c) $\mathbb{M} \setminus \{C\}$. (d) $\mathbb{M} \setminus \{D\}$. (e) $\mathbb{M} \setminus \{R\}$. (f) $\mathbb{M} \setminus \{G_{id}\}$.

which highlights the effectiveness of 3DMM in frontalization. Without the reconstruction loss G_{id} , the identity can be preserved to some extent, however the overall image quality is low and the lighting condition is not preserved. Therefore, we do not include the visual results considering space constraint.

Table 4 shows the quantitative results of the ablation study models by evaluating the recognition rate of the synthetic images generated from each model. Our FF-GAN with all modules and all loss functions performs the best among

all other variants, which suggests the effectiveness of each part of our framework. For example, the performance drops dramatically without the recognition engine regularization. The 3DMM module also performs a significant role in face frontalization.

removed module	—	C	D	R	G_{id}	G_{tv}	G_{sym}
performance (syn.)	74.2	59.2	73.4	68.5	69.3	72.9	73.1

Table 4: Quantitative results of ablation study.

6. Conclusions

In this work, we propose a 3DMM-conditioned GAN framework to frontalize faces under all pose ranges including extreme profile views. To the best of our knowledge, this is the first work to expand pose ranges to 90° in challenging large-scale in-the-wild databases. The 3DMM reconstruction module provides an important shape and appearance prior to guide the generator to rotate faces. The recognition engine regularizes the generated image to preserve identity. We propose new losses and carefully design the training procedure to obtain high-quality frontalized images. Extensive experiments suggest that FF-GAN can boost face recognition performance and be applied for 3D face reconstruction. Large-pose face frontalization in the wild is a challenging and ill-posed problem; we believe this work has made convincing progress towards a viable solution.

References

- [1] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. *SIGGRAPH*, 1999. 2, 3
- [2] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *CVPR*, 2017. 2
- [3] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep CNN features. In *WACV*, 2016. 3, 7
- [4] X. Chen, Y. Duan, R. Houthoof, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016. 3
- [5] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. Synthesizing normalized faces from facial identity features. In *CVPR*, 2017. 2
- [6] E. Denton, S. Chintala, and R. Fergus. Deep generative image models using a Laplacian pyramid of adversarial networks. In *NIPS*, 2015. 3
- [7] A. Dosovitskiy, J. Springenberg, and T. Brox. Learning to generate chairs with convolutional neural networks. In *CVPR*, 2015. 3
- [8] J. Fagan. Infants' recognition of invariant features of faces. *Child Development*, 1976. 1
- [9] C. Ferrari, G. Lisanti, S. Berretti, and A. Bimbo. Effective 3D based frontalization for unconstrained face recognition. In *ICPR*, 2016. 2, 7
- [10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 3, 4
- [11] R. Gross, I. Matthew, J. Cohn, T. Kanade, and S. Baker. MultiPIE. *Image and Vision Computing*, 2009. 6
- [12] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2014. 2, 6, 7
- [13] G. 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, University of Massachusetts, Amherst, 2007. 6
- [14] A. Jourabloo and X. Liu. Pose-invariant 3D face alignment. In *ICCV*, 2015. 2, 4
- [15] A. Jourabloo and X. Liu. Pose-invariant face alignment via CNN-based dense 3D model fitting. *IJCV*, 2017. 2
- [16] M. Kan, S. Shan, and X. Chen. Multi-view deep network for cross-view classification. In *CVPR*, 2016. 1, 2, 3
- [17] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *CVPR*, 2015. 6, 7
- [18] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *CVPR*, 2016. 1, 3, 7
- [19] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *CVPR*, 2017. 4
- [20] X. Peng, X. Yu, K. Sohn, D. N. Metaxas, and M. Chandraker. Reconstruction-based disentanglement for pose-invariant face recognition. In *ICCV*, 2017. 3
- [21] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, 2015. 3
- [22] J. Roth, Y. Tong, and X. Liu. Unconstrained 3D face reconstruction. In *CVPR*, 2015. 2
- [23] J. Roth, Y. Tong, and X. Liu. Adaptive 3D face reconstruction from unconstrained photo collections. *TPAMI*, 2016. 2
- [24] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Robust statistical face frontalization. In *ICCV*, 2015. 2
- [25] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV*, 2013. 5
- [26] S. Sengupta, J. C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *WACV*, 2016. 1
- [27] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, 2014. 1
- [28] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, 2014. 1
- [29] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. In *ICLR*, 2017. 3
- [30] L. Tran, X. Yin, and X. Liu. Disentangled representation learning GAN for pose-invariant face recognition. In *CVPR*, 2017. 3, 6, 7
- [31] C. Turati, H. Bulf, and F. Simion. Newborns' face recognition over changes in viewpoint. *Cognition*, 2008. 1
- [32] D. Wang, C. Otto, and A. K. Jain. Face search at scale. *TPAMI*, 2016. 7
- [33] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *NIPS*, 2016. 3
- [34] X. Xiong and F. D. la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013. 6
- [35] F. Yang, W. Choi, and Y. Lin. Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In *CVPR*, 2016. 5
- [36] J. Yang, S. Reed, M.-H. Yang, and H. Lee. Weakly supervised disentangling with recurrent transformations for 3D view synthesis. In *NIPS*, 2015. 2
- [37] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint:1411.7923*, 2014. 4, 5
- [38] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim. Rotating your face using multi-task deep neural network. In *CVPR*, 2015. 2, 7
- [39] X. Yin and X. Liu. Multi-task convolutional neural network for pose-invariant face recognition. *arXiv preprint arXiv:1702.04710*, 2017. 3
- [40] X. Yu, J. Huang, S. Zhang, and D. N. Metaxas. Face landmark fitting via optimized part mixtures and cascaded deformable model. *TPAMI*, 2015. 2

- [41] X. Yu, Z. Lin, J. Brandt, and D. N. Metaxas. Consensus of regression for occlusion-robust facial feature localization. In *ECCV*, 2014. [2](#)
- [42] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. In *ECCV*, 2016. [5](#)
- [43] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3D solution. In *CVPR*, 2016. [2](#), [4](#), [5](#), [6](#)
- [44] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, 2015. [2](#), [5](#), [7](#)
- [45] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *ICCV*, 2013. [3](#), [7](#)
- [46] Z. Zhu, P. Luo, X. Wang, and X. Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *NIPS*, 2014. [1](#), [2](#), [3](#), [7](#)