

Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis

Rui Huang^{1,2*}† Shu Zhang^{1,2,3*} Tianyu Li^{1,2} Ran He^{1,2,3}

¹National Laboratory of Pattern Recognition, CASIA

²Center for Research on Intelligent Perception and Computing, CASIA

³University of Chinese Academy of Sciences, Beijing, China

huangrui@cmu.edu, tianyu.lizard@gmail.com, {shu.zhang, rhe}@nlpr.ia.ac.cn

Abstract

Photorealistic frontal view synthesis from a single face image has a wide range of applications in the field of face recognition. Although data-driven deep learning methods have been proposed to address this problem by seeking solutions from ample face data, this problem is still challenging because it is intrinsically ill-posed. This paper proposes a Two-Pathway Generative Adversarial Network (TP-GAN) for photorealistic frontal view synthesis by simultaneously perceiving global structures and local details. Four landmark located patch networks are proposed to attend to local textures in addition to the commonly used global encoder-decoder network. Except for the novel architecture, we make this ill-posed problem well constrained by introducing a combination of adversarial loss, symmetry loss and identity preserving loss. The combined loss function leverages both frontal face distribution and pre-trained discriminative deep face models to guide an identity preserving inference of frontal views from profiles. Different from previous deep learning methods that mainly rely on intermediate features for recognition, our method directly leverages the synthesized identity preserving image for downstream tasks like face recognition and attribution estimation. Experimental results demonstrate that our method not only presents compelling perceptual results but also outperforms state-of-the-art results on large pose face recognition.

1. Introduction

Benefiting from the rapid development of deep learning methods and the easy access to a large amount of annotated face images, unconstrained face recognition techniques [31, 32] have made significant advances in recent years. Although surpassing human performance has been

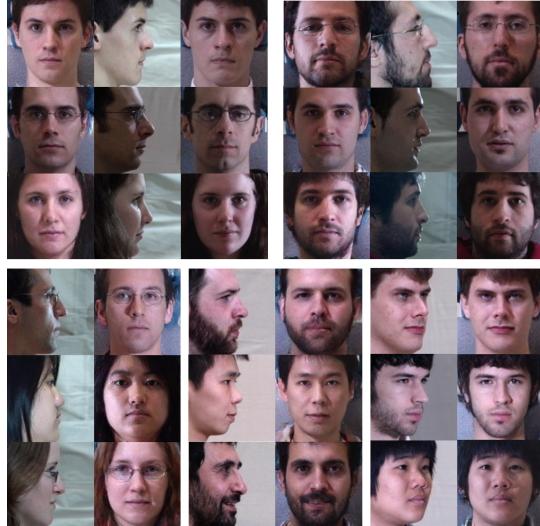


Figure 1. Frontal view synthesis by TP-GAN. The upper half shows the 90° profile image (middle) and its corresponding synthesized and ground truth frontal face. We invite the readers to guess which side is our synthesis results (please refer to Sec. 1 for the answer). The lower half shows the synthesized frontal view faces from profiles of 90° , 75° and 45° respectively.

achieved on several benchmark datasets [28], pose variations are still the bottleneck for many real-world application scenarios. Existing methods that address pose variations can be divided into two categories. One category tries to adopt hand-crafted or learned pose-invariant features [4, 28], while the other resorts to synthesis techniques to recover a frontal view image from a large pose face image and then use the recovered face images for face recognition [45, 46].

For the first category, traditional methods often make use of robust local descriptors such as Gabor [5], Haar [35] and LBP [2] to account for local distortions and then adopt metric learning [4, 36] techniques to achieve pose invariance. In contrast, deep learning methods often handle position variations with pooling operation and employ triplet loss [28] or

*These two authors contributed equally.

†Homepage <http://andrew.cmu.edu/user/ruih2/>

contrastive loss [31] to ensure invariance to very large intra-class variations. However, due to the tradeoff between invariance and discriminability, these approaches cannot deal with large pose cases effectively.

For the second category, earlier efforts on frontal view synthesis usually utilize 3D geometrical transformations to render a frontal view by first aligning the 2D image with either a general [12] or an identity specific [32,44] 3D model. These methods are good at normalizing small pose faces, but their performance decreases under large face poses due to severe texture loss. Recently, deep learning based methods are proposed to recover a frontal face in a data-driven way. For instance, Zhu *et al.* [46] propose to disentangle identity and pose representations while learning to estimate a frontal view. Although their results are encouraging, the synthesized image sometimes lacks fine details and tends to be blurry under a large pose so that they only use the intermediate features for face recognition. The synthesized image is still not good enough to perform other facial analysis tasks, such as forensics and attribute estimation.

Moreover, from an optimization point of view, recovering the frontal view from incompletely observed profile is an ill-posed or under-defined problem, and there exist multiple solutions to this problem if no prior knowledge or constraints are considered. Therefore, the quality of recovered results heavily relies on the prior or the constraints exploited in the training process. Previous work [17, 41, 45, 46] usually adopts pairwise supervision and seldom introduce constraints in the training process, so that they tend to produce blurry results.

When human try to conduct a view synthesis process, we firstly infer the global structure (or a sketch) of a frontal face based on both our prior knowledge and the observed profile. Then our attention moves to the local areas where all facial details will be filled out. Inspired by this process, we propose a deep architecture with two pathways (TP-GAN) for frontal view synthesis. These two pathways focus on the inference of global structure and the transformation of local texture respectively. Their corresponding feature maps are then fused for further process for the generation of the final synthesis. We also make the recovery process well constrained by incorporating prior knowledge of the frontal faces’ distribution with a Generative Adversarial Network (GAN) [9]. The outstanding capacity of GAN in modeling 2D data distribution has significantly advanced many ill-posed low level vision problems, such as super-resolution [19] and inpainting [24]. Particularly, drawing inspiration from the faces’ symmetric structure, a symmetry loss is proposed to fill out occluded parts. Moreover, to faithfully preserve the most prominent facial structure of an individual, we adopt a perceptual loss [16] in the compact feature space in addition to the pixel-wise L1 loss. Incorporating the identity preserving loss is critical for a faithful

synthesis and greatly improves its potential to be applied to face analysis tasks. We show some samples generated by TP-GAN in the upper half of Fig. 1 (the left side of each tuple).

The main contributions of our work lie in three folds: 1) We propose a human-like global and local aware GAN architecture for frontal view synthesis from a single image, which can synthesize photorealistic and identity preserving frontal view images even under a very large pose. 2) We combine prior knowledge from data distribution (adversarial training) and domain knowledge of faces (symmetry and identity preserving loss) to exactly recover the lost information inherent in projecting a 3D object into a 2D image space. 3) We demonstrate the possibility of a “recognition via generation” framework and outperform state-of-the-art recognition results under a large pose. Although some deep learning methods have been proposed for face synthesis, our method is the first attempt to be effective for the recognition task with synthesized faces.

2. Related Work

2.1. Frontal View Synthesis

Frontal view synthesis, or termed as face normalization, is a challenging task due to its ill-posed nature. Traditional methods address this problem either with 2D/3D local texture warping [12, 44] or statistical modeling [27]. For instance, Hassner *et al.* [12] employ a mean 3D model for face normalization. A joint frontal view synthesis and landmark localization method is proposed in [27] with a constrained low-rank minimization model. Recently, researchers employ Convolutional Neural Networks (CNN) for joint representation learning and view synthesis [17, 41, 45, 46]. Specifically, Yim *et al.* [41] propose a multi-task CNN to predict identity preserving rotated images. Zhu *et al.* [45, 46] develop novel architectures and learning objectives to disentangle the identity and pose representation while estimating the frontal view. Reed *et al.* [25] propose to use a Boltzmann machine to model factors of variation and generate rotated images via pose manifold traversal. Although it is much more convenient if the synthesized image can be directly used for facial analysis tasks, most of the previous methods mainly employ intermediate features for face recognition because they cannot faithfully produce an identity preserving synthesis.

2.2. Generative Adversarial Network (GAN)

As one of the most significant improvements on the research of deep generative models [18, 26], GAN [9] has drawn substantial attention from both the deep learning and computer vision society. The min-max two-player game provides a simple yet powerful way to estimate target distribution and generate novel image samples [6]. With its

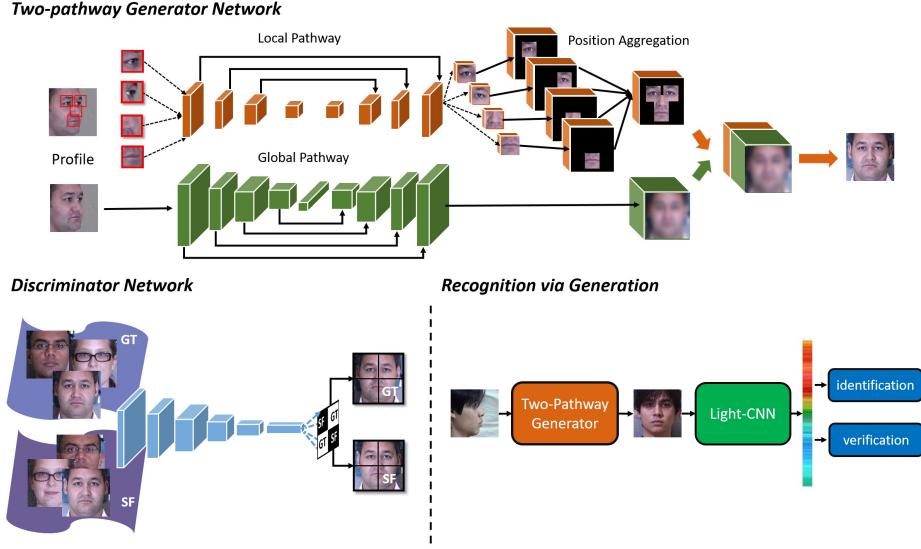


Figure 2. General framework of TP-GAN. The Generator contains two pathways with each processing global or local transformations. The Discriminator distinguishes between synthesized frontal (SF) views and ground-truth (GT) frontal views. Detailed network architectures can be found in the supplementary material.

power for distribution modeling, the GAN can encourage the generated images to move towards the true image manifold and thus generates photorealistic images with plausible high frequency details. Recently, modified GAN architectures, conditional GAN [21] in particular, have been successfully applied to vision tasks like image inpainting [24], super-resolution [19], style transfer [20], face attribute manipulation [29] and even data augmentation for boosting classification models [30,43]. These successful applications of GAN motivate us to develop frontal view synthesis methods based on GAN.

3. Approach

The aim of frontal view synthesis is to recover a photorealistic and identity preserving frontal view image I^F from a face image under a different pose, *i.e.* a profile image I^P . To train such a network, pairs of corresponding $\{I^F, I^P\}$ from multiple identities y are required during the training phase. Both the input I^P and output I^F come from a pixel space of size $W \times H \times C$ with C color channel.

It's our goal to learn a synthesis function that can infer the corresponding frontal view from any given profile images. Specifically, we model the synthesis function with a two-pathway CNN G_{θ_G} that is parametrized by θ_G . Each pathway contains an Encoder and a Decoder, denoted as $\{G_{\theta_E^g}, G_{\theta_D^g}\}$ and $\{G_{\theta_E^l}, G_{\theta_D^l}\}$, where g and l stand for the global structure pathway and the local texture pathway respectively. In the global pathway, the bottleneck layer, which is the output of $G_{\theta_E^g}$, is usually used for classification task [40] with the cross-entropy loss $L_{cross_entropy}$.

The network's parameters G_{θ_G} are optimized by mini-

mizing a specifically designed synthesis loss L_{syn} and the aforementioned $L_{cross_entropy}$. For a training set with N training pairs of $\{I_n^F, I_n^P\}$, the optimization problem can be formulated as follows:

$$\hat{\theta}_G = \frac{1}{N} \underset{\theta_G}{\operatorname{argmin}} \sum_{n=1}^N \{L_{syn}(G_{\theta_G}(I_n^P), I_n^F) + \alpha L_{cross_entropy}(G_{\theta_E^g}(I_n^P), y_n)\} \quad (1)$$

where α is a weighting parameter and L_{syn} is defined as a weighted sum of several losses that jointly constrain an image to reside in the desired manifold. We will postpone the detailed description of all the individual loss functions to Sec. 3.2.

3.1. Network Architecture

3.1.1 Two Pathway Generator

The general architecture of TP-GAN is shown in Fig. 2. Different from previous methods [17,41,45,46] that usually model the synthesis function with one single network, our proposed generator G_{θ_G} has two pathways, with one global network G_{θ_g} processing the global structure and four landmark located patch networks $G_{\theta_l^i}, i \in \{0, 1, 2, 3\}$ attending to local textures around four facial landmarks.

We are not the first to employ the two pathway modeling strategy. Actually, this is a quite popular routine for 2D/3D local texture warping [12,44] methods. Similar to the human cognition process, they usually divide the normalization of faces into two steps, with the first step to align the face globally with a 2D or 3D model and the second step to

warp or render local texture to the global structure. Moreover, Mohammed *et al.* [22] combines a global parametric model with a local non-parametric model for novel face synthesis.

Synthesizing a frontal face I^F from a profile image I^P is a highly non-linear transformation. Since the filters are shared across all the spatial locations of the face image, we argue that using merely a global network cannot learn filters that are suitable for both rotating a face and precisely recovering local details. Therefore, we transfer the success of the two pathway structure in traditional methods to a deep learning based framework and introduce the human-like two pathway generator for frontal view synthesis.

As shown in Fig. 2, G_{θ_g} is composed of a down-sampling Encoder $G_{\theta_E^g}$ and an up-sampling Decoder $G_{\theta_D^g}$, extra skip layers are introduced for multi-scale feature fusion. The bottleneck layer in the middle outputs a 256-dimension feature vector v_{id} , which is used for identity classification to allow for identity-preserving synthesis. At this bottleneck layer, as in [33], we concatenate a 100-dim Gaussian random noise to v_{id} to model variations other than pose and identity.

3.1.2 Landmark Located Patch Network

The four input patches of the landmark located patch network G_{θ_l} are center-cropped from four facial landmarks, *i.e.* left eye center, right eye center, nose tip and mouth center. Each $G_{\theta_l^i}, i \in \{0, 1, 2, 3\}$ learns a separate set of filters for rotating the center-cropped patch to its corresponding frontal view (after rotation, the facial landmarks are still in the center). The architecture of the landmark located patch network is also based on an encoder-decoder structure, but it has no fully connected bottleneck layer.

To effectively integrate the information from the global and local pathways, we adopt an intuitive method for feature map fusion. As shown in Fig. 2, we firstly fuse the output feature tensors (multiple feature maps) of four local pathways to one single feature tensor that is of the same spatial resolution as the global feature tensor. Specifically, we put each feature tensor at a “template landmark location”, and then a max-out fusing strategy is introduced to reduce the stitching artifacts on the overlapping areas. Then, we simply concatenate the feature tensor from each pathway to produce a fused feature tensor and then feed it to successive convolution layers to generate the final synthesis output.

3.1.3 Adversarial Networks

To incorporate prior knowledge of the frontal faces’ distribution into the training process, we further introduce an discriminator D_{θ_D} to distinguish real frontal face images I^F from synthesized frontal face images $G_{\theta_G}(I^P)$, following the work of Goodfellow *et al.* [9]. We train D_{θ_D} and G_{θ_G}

in an alternating way to optimize the following min-max problem:

$$\begin{aligned} & \min_{\theta_G} \max_{\theta_D} \mathbb{E}_{I^F \sim P(I^F)} \log D_{\theta_D}(I^F) + \\ & \mathbb{E}_{I^P \sim P(I^P)} \log(1 - D_{\theta_D}(G_{\theta_G}(I^P))) \end{aligned} \quad (2)$$

Solving this min-max problem will continually push the output of the generator to match the target distribution of the training frontal faces, thus it encourages the synthesized image to reside in the manifold of frontal faces, leading to photorealistic synthesis with appealing high frequency details. As in [30], our D_{θ_D} outputs a 2×2 probability map instead of one scalar value. Each probability value now corresponds to a certain region instead of the whole face, and D_{θ_D} can specifically focus on each semantic region.

3.2 Synthesis Loss Function

The synthesis loss function used in our work is a weighted sum of four individual loss functions, we will give a detailed description in the following sections.

3.2.1 Pixel-wise Loss

We adopt pixel-wise L1 loss at multiple locations to facilitate multi-scale image content consistency:

$$L_{pixel} = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H |I_{x,y}^{pred} - I_{x,y}^{gt}| \quad (3)$$

Specifically, the pixel wise loss is measured at the output of the global, the landmark located patch network and their final fused output. To facilitate a deep supervision, we also add the constraint on multi-scale outputs of the $G_{\theta_D^g}$. Although this loss will lead to overly smooth synthesis results, it is still an essential part for both accelerated optimization and superior performance.

3.2.2 Symmetry Loss

Symmetry is an inherent feature of human faces. Exploiting this domain knowledge as a prior and imposing a symmetric constraint on the synthesized images may effectively alleviate the self-occlusion problem and thus greatly improve performance for large pose cases. Specifically, we define a symmetry loss in two spaces, *i.e.* the original pixel space and the Laplacian image space, which is robust to illumination changes. The symmetry loss of a face image takes the form:

$$L_{sym} = \frac{1}{W/2 \times H} \sum_{x=1}^{W/2} \sum_{y=1}^H |I_{x,y}^{pred} - I_{W-(x-1),y}^{pred}| \quad (4)$$

For simplicity, we selectively flip the input so that the occluded part are all on the right side. Besides, only the occluded part (right side) of I^{pred} receives the symmetry



Figure 3. Comparison with state-of-the-art synthesis methods under the pose of 45° (first two rows) and 30° (last row).

loss, i.e. we explicitly pull the right side to be closer to the left. L_{sym} ’s contribution is twofold, generating realistic images by encouraging a symmetrical structure and accelerating the convergence of TP-GAN by providing additional back-propagation gradient to relieve self-occlusion for extreme poses. However, due to illumination changes or intrinsic texture difference, pixel values are not strictly symmetric most of the time. Fortunately, the pixel difference inside a local area is consistent, and the gradients of a point along all directions are largely reserved under different illuminations. Therefore, the Laplacian space is more robust to illumination changes and more indicative for face structure.

3.2.3 Adversarial Loss

The loss for distinguishing real frontal face images I^F from synthesized frontal face images $G_{\theta_G}(I^P)$ is calculated as follows:

$$L_{adv} = \frac{1}{N} \sum_{n=1}^N -\log D_{\theta_D}(G_{\theta_G}(I_n^P)) \quad (5)$$

L_{adv} serves as a supervision to push the synthesized image to reside in the manifold of frontal view images. It can prevent blur effect and produce visually pleasing results.

3.2.4 Identity Preserving Loss

Preserving the identity while synthesizing the frontal view image is the most critical part in developing the “recognition via generation” framework. In this work, we exploit the perceptual loss [16] that is originally proposed for maintaining perceptual similarity to help our model gain the identity preserving ability. Specifically, we define the identity preserving loss based on the activations of the last two layers of the Light CNN [38]:

$$L_{ip} = \sum_{i=1}^2 \frac{1}{W_i \times H_i} \sum_{x=1}^{W_i} \sum_{y=1}^{H_i} |F(I^P)_{x,y}^i - F(G(I^{pred}))_{x,y}^i| \quad (6)$$

where W_i, H_i denotes the spatial dimension of the last i th layer. The identity preserving loss enforces the prediction to have a small distance with the ground-truth in the compact deep feature space. Since the Light CNN is pre-trained to classify tens of thousands of identities, it can capture the most prominent feature or face structure for identity discrimination. Therefore, it is totally viable to leverage this loss to enforce an identity preserving frontal view synthesis.

L_{ip} has better performance when used with L_{adv} . Using L_{ip} alone makes the results prone to annoying artifacts, because the search for a local minimum of L_{ip} may go through a path that resides outside the manifold of natural face images. Using L_{adv} and L_{ip} together can ensure that the search resides in that manifold and produces photorealistic image.

3.2.5 Overall Objective Function

The final synthesis loss function is a weighted sum of all the losses defined above:

$$L_{syn} = L_{pixel} + \lambda_1 L_{sym} + \lambda_2 L_{adv} + \lambda_3 L_{ip} + \lambda_4 L_{tv} \quad (7)$$

We also impose a total variation regularization L_{tv} [16] on the synthesized result to reduce spike artifacts.

4. Experiments

Except for synthesizing natural looking frontal view images, the proposed TP-GAN also aims to generate identity preserving image for accurate face analysis with off-the-shelf deep features. Therefore, in this section, we demonstrate the merits of our model on qualitative synthesis results and quantitative recognition results in Sec. 4.1 and 4.2. Sec. 6.3 presents visualization of the final deep feature representations to illustrate the effectiveness of TP-GAN. Finally, in Sec. 4.4, we conduct detailed algorithmic evaluation to demonstrate the advantages of the proposed two-pathway architecture and synthesis loss function.

Implementation details We use colorful images of size $128 \times 128 \times 3$ in all our experiments for both the input



Figure 4. Synthesis results by TP-GAN under different poses. From left to right, the poses are 90° , 75° , 60° , 45° , 30° and 15° . The ground truth frontal images are provided at the last column.



Figure 5. Challenging situations. The facial attributes, e.g. beard, eyeglasses are preserved by TP-GAN. The occluded forehead and cheek are recovered.

Table 1. Rank-1 recognition rates (%) across views and illuminations under Setting 1. For all the remaining tables, only methods marked with * follow the “recognition via generation” procedure while others leverage intermediate features for face recognition.

| Method | $\pm 90^\circ$ | $\pm 75^\circ$ | $\pm 60^\circ$ | $\pm 45^\circ$ | $\pm 30^\circ$ | $\pm 15^\circ$ |
|------------------------------|----------------|----------------|----------------|----------------|----------------|----------------|
| CPF [41] | - | - | - | 71.65 | 81.05 | 89.45 |
| Hassner <i>et al.</i> * [12] | - | - | 44.81 | 74.68 | 89.59 | 96.78 |
| HPN [7] | 29.82 | 47.57 | 61.24 | 72.77 | 78.26 | 84.23 |
| FIP_40 [45] | 31.37 | 49.10 | 69.75 | 85.54 | 92.98 | 96.30 |
| c-CNN Forest [39] | 47.26 | 60.66 | 74.38 | 89.02 | 94.05 | 96.97 |
| Light CNN [38] | 9.00 | 32.35 | 73.30 | 97.45 | 99.80 | 99.78 |
| TP-GAN* | 64.03 | 84.10 | 92.93 | 98.58 | 99.85 | 99.78 |

I^P and the prediction $I^{pred} = G_{\theta_G}(I^P)$. Our method is evaluated on MultiPIE [10], a large dataset with 750,000+ images for face recognition under pose, illumination and expression changes. The feature extraction network, Light CNN, is trained on MS-Celeb-1M [11] and fine-tuned on the original images of MultiPIE. Our network is implemented with Tensorflow [1]. The training of TP-GAN lasts for one day with a batch size of 10 and a learning rate of 10^{-4} . In all our experiments, we empirically set $\alpha = 10^{-3}$, $\lambda_1 = 0.3$, $\lambda_2 = 10^{-3}$, $\lambda_3 = 3 \times 10^{-3}$ and $\lambda_4 = 10^{-4}$.

4.1. Face Synthesis

Most of the previous work on frontal view synthesis are dedicated to address that problem within a pose range of $\pm 60^\circ$. Because it is commonly believed that with a pose larger than 60° , it is difficult to faithfully recover a frontal view image. However, we will show that given enough training data and a proper architecture and loss design, it is in fact feasible to recover photorealistic frontal views from very large poses. Fig. 4 shows TP-GAN’s ability to recover compelling identity-preserving frontal faces from any pose



Figure 6. Mean faces from six images (within $\pm 45^\circ$) per identity.

and Fig. 3 illustrates a comparison with state-of-the-art face frontalization methods. Note that most of TP-GAN’s competitors cannot deal with poses larger than 45° , therefore, we only report their results under 30° and 45° .

Compared to competing methods, TP-GAN presents a good identity preserving quality while producing photo-realistic synthesis. Thanks to the data-driven modeling with prior knowledge from L_{adv} and L_{ip} , not only the overall face structure but also the occluded ears, cheeks and forehead can be hallucinated in an identity consistent way. Moreover, it also perfectly preserves observed face attributes in the original profile image, e.g. eyeglasses and hair style, as shown in Fig. 5.

To further demonstrate the stable geometry shape of the syntheses across multiple poses, we show the mean image of synthesized faces from different poses in Fig. 6. The mean faces from TP-GAN preserve more texture detail and contain less blur effect, showing a stable geometry shape across multiple syntheses. Note that our method does not rely on any 3D knowledge for geometry shape estimation, the inference is made through sheer data-driven learning.

As a demonstration of our model’s superior generalization ability to in the wild faces, we use images from LFW [14] dataset to test a TP-GAN model trained solely on Multi-PIE. As shown in Fig. 7, although the resultant color tone is similar to images from Multi-PIE, TP-GAN can faithfully synthesize frontal view images with both finer details and better global shapes for faces in LFW dataset compared to state-of-the-art methods like [12, 44].



(a) LFW (b) Ours (c) [44] (d) [12]

Figure 7. Synthesis results on the LFW dataset. Note that TP-GAN is trained on MultiPIE.

4.2. Identity Preserving Property

Face Recognition To quantitatively demonstrate our method’s identity preserving ability, we conduct face recognition on MultiPIE with two different settings. The experiments are conducted by firstly extracting deep features with Light-CNN [38] and then compare Rank-1 recognition accuracy with a cosine-distance metric. The results on the profile images I^P serve as our baseline and are marked by the notation Light-CNN in all tables. It should be noted that although many deep learning methods have been proposed for frontal view synthesis, none of their synthesized images proved to be effective for recognition tasks. In a recent study on face hallucination [37], the authors show that directly using a CNN synthesized high resolution face image for recognition will certainly degenerate the performance instead of improving it. Therefore, it is of great significance to validate whether our synthesis results can boost the recognition performance (whether the “recognition via generation” procedure works).

In Setting 1, we follow the protocol from [39], and only images from session one are used. We include images with neutral expression under 20 illuminations and 11 poses within $\pm 90^\circ$. One gallery image with frontal view and illumination is used for each testing subject. There is no overlap between training and testing sets. Table 1 shows our recognition performance and the comparison with the state-of-the-art. TP-GAN consistently achieves the best performance across all angles, and the larger the angle, the greater the improvement. When compared with c-CNN Forest [39], which is an ensemble of three models, we achieve a performance boost of about 20% on large pose cases.

In Setting 2, we follow the protocol from [41], where neural expression images from all four sessions are used. One gallery image is selected for each testing identity from their first appearance. All synthesized images of MultiPIE

Table 2. Rank-1 recognition rates (%) across views, illuminations and sessions under Setting 2.

| Method | $\pm 90^\circ$ | $\pm 75^\circ$ | $\pm 60^\circ$ | $\pm 45^\circ$ | $\pm 30^\circ$ | $\pm 15^\circ$ |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| FIP+LDA [45] | - | - | 45.9 | 64.1 | 80.7 | 90.7 |
| MVP+LDA [46] | - | - | 60.1 | 72.9 | 83.7 | 92.8 |
| CPF [41] | - | - | 61.9 | 79.9 | 88.5 | 95.0 |
| DR-GAN [33] | - | - | 83.2 | 86.2 | 90.1 | 94.0 |
| Light CNN [38] | 5.51 | 24.18 | 62.09 | 92.13 | 97.38 | 98.59 |
| TP-GAN* | 64.64 | 77.43 | 87.72 | 95.38 | 98.06 | 98.68 |

Table 3. Gender classification accuracy (%) across views and illuminations.

| Method | $\pm 45^\circ$ | $\pm 30^\circ$ | $\pm 15^\circ$ |
|------------------------------|----------------|----------------|----------------|
| I_{60}^P | 85.46 | 87.14 | 90.05 |
| CPI* [41] | 76.80 | 78.75 | 81.55 |
| Amir <i>et al.</i> * [8] | 77.65 | 79.70 | 82.05 |
| I_{128}^P | 86.22 | 87.70 | 90.46 |
| Hassner <i>et al.</i> * [12] | 83.83 | 84.74 | 87.15 |
| TP-GAN* | 90.71 | 89.90 | 91.22 |

in this paper are from the testing identities under Setting 2. The result is shown in Table 2. Note that all the compared CNN based methods achieve their best performances with learned intermediate features, whereas we directly use the synthesized images following a “recognition via generation” procedure.

Gender Classification To further demonstrate the potential of our synthesized images on other facial analysis tasks, we conduct an experiment on gender classification. All the compared methods in this part also follow the “recognition via generation” procedure, where we directly use their synthesis results for gender classification. The CNN for gender classification is of the same structure as the encoder $G_{\theta_E^g}$ and is trained on *batch1* of the UMD [3] dataset.

We report the testing performance on Multi-PIE (Setting-1) in Table 3. For fair comparison, we present the results on the unrotated original images in two resolutions, 128×128 (I_{128}^P) and 60×60 (I_{60}^P) respectively. TP-GAN’s synthesis achieves a better classification accuracy than the original profile images due to normalized views. It’s not surprising to see that all other compared models perform worse than the baseline, as their architectures are not designed for the gender classification task. Similar phenomenon is observed in [37] where synthesized high resolution face images severely degenerate the recognition performance instead of improving it. That indicates the high risk of losing prominent facial features of I^P when manipulating images in the pixel space.

4.3. Feature Visualization

We use t-SNE [34] to visualize the 256-dim deep feature on a two dimensional space. The left side of Fig. 8 illustrates the deep feature space of the original profile images. It’s clear that images with a large pose (90° in particular) are not separable in the deep feature space spanned by the Light-CNN. It reveals that even though the Light-CNN is trained with millions of images, it still cannot prop-

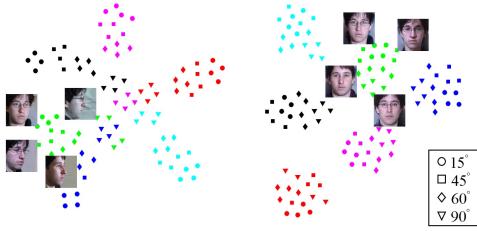


Figure 8. Feature space of the profile faces (left) and frontal view synthesized images (right). Each color represents a different identity. Each shape represent a view. The images for one identity are labeled.

Table 4. Model comparison: Rank-1 recognition rates (%) under Setting 2.

| Method | $\pm 90^\circ$ | $\pm 75^\circ$ | $\pm 60^\circ$ | $\pm 45^\circ$ | $\pm 30^\circ$ | $\pm 15^\circ$ |
|---------------|----------------|----------------|----------------|----------------|----------------|----------------|
| w/o P | 44.13 | 66.10 | 80.64 | 92.07 | 96.59 | 98.35 |
| w/o L_{ip} | 43.23 | 56.55 | 70.99 | 85.87 | 93.43 | 97.06 |
| w/o L_{adv} | 62.83 | 76.10 | 85.04 | 92.45 | 96.34 | 98.09 |
| w/o L_{sym} | 62.47 | 75.71 | 85.23 | 93.13 | 96.50 | 98.47 |
| TP-GAN | 64.64 | 77.43 | 87.72 | 95.38 | 98.06 | 98.68 |

erly deal with large pose face recognition problems. On the right side, after frontal view synthesis with our TP-GAN, the generated frontal view images can be easily classified into different groups according to their identities.

4.4. Algorithmic analysis

In this section, we go over different architectures and loss function combinations to gain insight into their respective roles in frontal view synthesis. Both qualitative visualization results and quantitative recognition results are reported for a comprehensive comparison.

We compare four variations of TP-GAN in this section, one for comparing the architectures and the other three for comparing the objective functions. Specifically, we train a network without the local pathway (denoted as P) as the first variant. With regards to the loss function, we keep the two-pathway architecture intact and remove one of the three losses, *i.e.* L_{ip} , L_{adv} and L_{sym} , in each case.

Detailed recognition performance is reported in Table 4. The two-pathway architecture and the identity preserving loss contribute the most for improving the recognition performance, especially on large pose cases. Although not as much apparent, both the symmetry loss and the adversarial loss help to improve the recognition performance. Fig. 9 illustrates the perceptual performance of these variants. As expected, inference results without the identity preserving loss or the local pathway deviate from the true appearance seriously. And the synthesis without adversarial loss tends to be very blurry, while the result without the symmetry loss sometimes shows unnatural asymmetry effect.

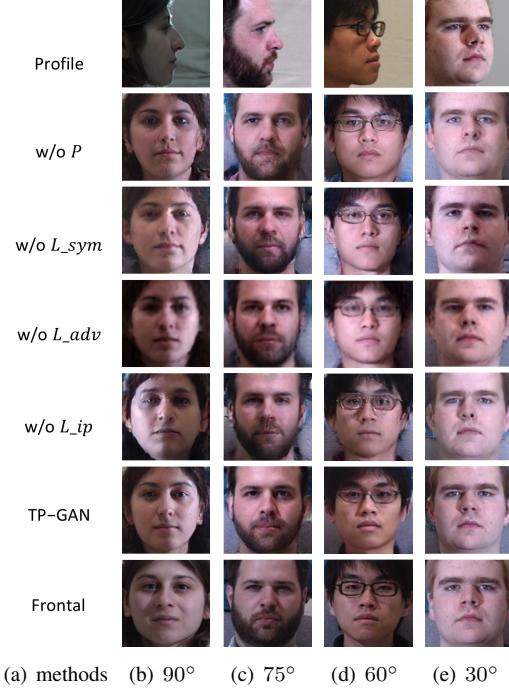


Figure 9. Model comparison: synthesis results of TP-GAN and its variants.

5. Conclusion

In this paper, we have presented a global and local perception GAN framework for frontal view synthesis from a single image. The framework contains two separate pathways, modeling the out-of-plane rotation of the global structure and the non-linear transformation of the local texture respectively. To make the ill-posed synthesis problem well constrained, we further introduce adversarial loss, symmetry loss and identity preserving loss in the training process. Adversarial loss can faithfully discover and guide the synthesis to reside in the data distribution of frontal faces. Symmetry loss can explicitly exploit the symmetry prior to ease the effect of self-occlusion in large pose cases. Moreover, identity preserving loss is incorporated into our framework, so that the synthesis results are not only visually appealing but also readily applicable to accurate face recognition. Experimental results demonstrate that our method not only presents compelling perceptual results but also outperforms state-of-the-art results on large pose face recognition.

6. Supplementary Material

6.1. Detailed Network Architecture

The detailed structures of the global pathway $G_{\theta_E^g}$ and $G_{\theta_D^g}$ are provided in Table 5 and Table 6. Each convolution layer of $G_{\theta_E^g}$ is followed by one residual block [13]. Particularly, the layer *conv4* is followed by four blocks. The



Figure 10. Our synthesized images present moderately better exposure in some cases. Each tuple consists of three images, with the input I^P on the left, the synthesized in the middle, the ground truth frontal face I^{gt} on the right. Each I^P and its corresponding I^{gt} are taken under a flash light from the same direction.



Figure 11. Synthesis results under various illuminations. The first row is the synthesized image, the second row is the input. Please to refer to the supplementary material for more results.

output of the layer $fc2$ (v_{id}) is obtained by selecting the maximum element from the two split halves of $fc1$.

The Decoder of the global pathway $G_{\theta_D^g}$ contains two parts. The first part is a simple deconvolution stack for up-sampling the concatenation of the feature vector v_{id} and the random noise vector z . The second part is the main deconvolution stack for reconstruction. Each layer takes the output of its previous layer as the regular input, which is omitted in the table for readability. Any extra inputs are specified in the *Input* column. Particularly, the layers *feat8* and *deconv0* have their complete inputs specified. Those extra inputs instantiate the skipping layers and the bridge between the two pathways. The fused feature tensor from the local pathway is denoted as *local* in Table 6. Tensor *local* is the fusion of the outputs of four $G_{\theta_D^l}$'s layer *conv4* (of Table 7). To mix the information of the various inputs, all extra inputs pass through one or two residual blocks before being concatenated for deconvolution. The profile image I^P is resized to the corresponding resolution and provides a shortcut access to the original texture for $G_{\theta_D^g}$.

Table 7 shows the structures of the local pathway $G_{\theta_E^l}$ and $G_{\theta_D^l}$. The local pathway contains three down-sampling and up-sampling processes respectively. The w and h denote the width and the height of the cropped patch. For the patches of the two eyes, we set w and h as 40; for the patch of the nose, we set w as 40 and h as 32; for the patch of the mouth, we set w and h as 48 and 32 respectively.

We use rectified linear units (ReLU) [23] as the non-linearity activation and adopt batch normalization [15] except for the last layer. In $G_{\theta_E^g}$ and $G_{\theta_E^l}$, the leaky ReLU is adopted.

Discussion: Our model is simple while achieving better performance in terms of the photorealism of synthesized

Table 5. Structure of the Encoder of the global pathway $G_{\theta_E^g}$

| Layer | Filter Size | Output Size |
|-------|----------------|----------------------------|
| conv0 | $7 \times 7/1$ | $128 \times 128 \times 64$ |
| conv1 | $5 \times 5/2$ | $64 \times 64 \times 64$ |
| conv2 | $3 \times 3/2$ | $32 \times 32 \times 128$ |
| conv3 | $3 \times 3/2$ | $16 \times 16 \times 256$ |
| conv4 | $3 \times 3/2$ | $8 \times 8 \times 512$ |
| fc1 | - | 512 |
| fc2 | - | 256 |

Table 6. Structure of the Decoder of the global pathway $G_{\theta_D^g}$. The *conv*s in *Input* column refer to those in Table 5.

| Layer | Input | Filter Size | Output Size |
|---------|------------------------------|----------------|----------------------------|
| feat8 | $fc2, z$ | - | $8 \times 8 \times 64$ |
| feat32 | - | $3 \times 3/4$ | $32 \times 32 \times 32$ |
| feat64 | - | $3 \times 3/2$ | $64 \times 64 \times 16$ |
| feat32 | - | $3 \times 3/2$ | $128 \times 128 \times 8$ |
| deconv0 | $feat8, conv4$ | $3 \times 3/2$ | $16 \times 16 \times 512$ |
| deconv1 | $conv3$ | $3 \times 3/2$ | $32 \times 32 \times 256$ |
| deconv2 | $feat32, conv2, I^P$ | $3 \times 3/2$ | $64 \times 64 \times 128$ |
| deconv3 | $feat4, conv1, I^P$ | $3 \times 3/2$ | $128 \times 128 \times 64$ |
| conv5 | $feat128, conv0, local, I^P$ | $5 \times 5/1$ | $128 \times 128 \times 64$ |
| conv6 | - | $3 \times 3/1$ | $128 \times 128 \times 32$ |
| conv7 | - | $3 \times 3/1$ | $128 \times 128 \times 3$ |

Table 7. Structure of the local pathway $G_{\theta_E^l}$ & $G_{\theta_D^l}$. The *conv*s in *Input* column refer to those in the same table.

| Layer | Input | Filter Size | Output Size |
|---------|---------|----------------|-----------------------------|
| conv0 | - | $3 \times 3/1$ | $w \times h \times 64$ |
| conv1 | - | $3 \times 3/2$ | $w/2 \times h/2 \times 128$ |
| conv2 | - | $3 \times 3/2$ | $w/4 \times h/4 \times 256$ |
| conv3 | - | $3 \times 3/2$ | $w/8 \times h/8 \times 512$ |
| deconv0 | $conv3$ | $3 \times 3/2$ | $w/4 \times h/4 \times 256$ |
| deconv1 | $conv2$ | $3 \times 3/2$ | $w/2 \times h/2 \times 128$ |
| deconv2 | $conv1$ | $3 \times 3/2$ | $w \times h \times 64$ |
| conv4 | $conv0$ | $3 \times 3/1$ | $w \times h \times 64$ |
| conv5 | - | $3 \times 3/1$ | $w \times h \times 3$ |

images. Yim *et al.* [41] and Zhu *et al.* [45] use locally connected convolutional layers for feature extraction and fully connected layer for synthesis. We use weight-sharing convolution in most cases. Our model reduces parameter num-

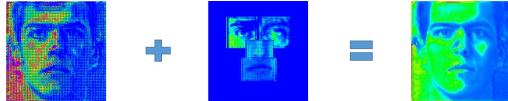


Figure 12. Synthesis process illustrated from the perspective of activation maps. The up-sampled feature map C_g is combined with the local pathway feature map C_l to produce feature maps with detailed texture.

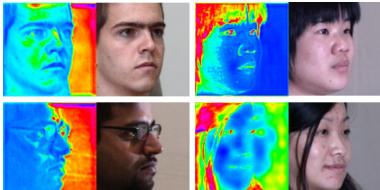


Figure 13. Automatic detection of certain semantic regions. Some skip layers’ activation maps are sensitive to certain semantic regions. One for detecting non-face region is shown on the left, another for detecting hair region is shown on the right. Note the delicate and complex region boundaries around the eyeglasses and the fringe.

bers to a large extent and avoids expensive computation for generating every pixel during synthesis. Yim *et al.* [41] and Amir *et al.* [8] add a second reconstruction branch or a refinement network. Our early supervised decoder achieves end-to-end generation of high-resolution image.

6.2. Additional Synthesis Results

Additional synthesized images I^{pred} are shown in Fig. 10 and Fig. 11. Under extreme illumination condition, the exposure of I^{pred} is consistent with or moderately better than that of its input I^P or its ground truth frontal face I^{gt} . Fig. 11 demonstrates TP-GAN’s robustness to illumination changes. Despite extreme illumination variations, the skin tone, global structure and local details are consistent across illuminations. Our method can automatically adjust I^P ’s exposure and white balance.

Additionally, we use a state-of-the-art face alignment method [42] to provide four landmarks for TP-GAN under extreme poses. The result is only slightly worse than that reported in Table 2 of the paper. Specifically, we achieve Rank-1 recognition rates of $87.63(\pm 60^\circ)$, $76.69(\pm 75^\circ)$, $62.43(\pm 90^\circ)$.

6.3. Activation Maps Visualization

In this part, we visualize the intermediate feature maps to gain some insights into the processing mechanism of the two-pathway network. Fig. 12 illustrates the fusion of global and local information before the final output. C_g contains the up-sampled outputs of the global pathway and C_l refers to the features maps fused from the four local pathways. Their information is concatenated and further integrated by the following convolutional layers.

We also discovered that TP-GAN can automatically detect certain semantic regions. Fig. 13 shows that certain skip layers have high activation for regions such as non-face region and hair region. The detection is learned by the network without supervision. Intuitively, dividing the input image into different semantic regions simplifies the following composition or synthesis of the frontal face.

Acknowledgement

This work is partially funded by National Natural Science Foundation of China (Grant No. 61622310, 61473289) and the State Key Development Program (Grant No. 2016YFB1001001). We thank Xiang Wu for useful discussion.

References

- [1] M. Abadi et al. Tensorflow: A system for large-scale machine learning. In *OSDI*, pages 265–283, 2016. 6
- [2] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *TPAMI*, 2006. 1
- [3] A. Bansal, A. Nanduri, R. Ranjan, C. D. Castillo, and R. Chellappa. Umdfaces: An annotated face dataset for training deep networks. *arXiv:1611.01484*, 2016. 7
- [4] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *CVPR*, 2013. 1
- [5] J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *JOSA*, 1985. 1
- [6] E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 2015. 2
- [7] C. Ding and D. Tao. Pose-invariant face recognition with homography-based normalization. *Pattern Recognition*, 66:144 – 152, 2017. 6
- [8] A. Ghodrati, X. Jia, M. Pedersoli, and T. Tuytelaars. Towards automatic image editing: Learning to see another you. In *BMVC*, 2016. 5, 6, 7, 10
- [9] 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. 2, 4
- [10] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image Vision Computing*, 2010. 6
- [11] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 6
- [12] T. Hassner, S. Harel, E. Paz, and R. Enbar. Effective face frontalization in unconstrained images. In *CVPR*, 2015. 2, 3, 5, 6, 7
- [13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. 8

- [14] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 6
- [15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*, 2015. 9
- [16] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2, 5
- [17] M. Kan, S. Shan, H. Chang, and X. Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. In *CVPR*, 2014. 2, 3
- [18] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2
- [19] C. Ledig et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 2, 3
- [20] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *CVPR*, 2016. 3
- [21] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014. 3
- [22] U. Mohammed, S. J. Prince, and J. Kautz. Visio-lization: generating novel facial images. In *TOG*, 2009. 4
- [23] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In J. Frnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814. Omnipress, 2010. 9
- [24] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2, 3
- [25] S. Reed, K. Sohn, Y. Zhang, and H. Lee. Learning to disentangle factors of variation with manifold interaction. In *ICML*, 2014. 2
- [26] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014. 2
- [27] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic. Robust statistical face frontalization. In *ICCV*, 2015. 2
- [28] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 1
- [29] W. Shen and R. Liu. Learning residual images for face attribute manipulation. In *CVPR*, 2017. 3
- [30] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb. Learning from simulated and unsupervised images through adversarial training. In *CVPR*, 2017. 3, 4
- [31] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *CVPR*, 2014. 1, 2
- [32] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1, 2
- [33] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017. 4, 5, 7
- [34] L. van der Maaten and G. E. Hinton. Visualizing high-dimensional data using t-sne. *JMLR*, 2008. 7
- [35] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 1
- [36] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 2009. 1
- [37] J. Wu, S. Ding, W. Xu, and H. Chao. Deep joint face hallucination and recognition. *arXiv:1611.08091*, 2016. 7
- [38] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *arXiv:1511.02683*, 2016. 5, 6, 7
- [39] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, and T. K. Kim. Conditional convolutional neural network for modality-aware face recognition. In *ICCV*, 2015. 6, 7
- [40] J. Yang, S. E. Reed, M.-H. Yang, and H. Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *NIPS*, 2015. 3
- [41] 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, 3, 5, 6, 7, 9, 10
- [42] H. Zhang, Q. Li, and Z. Sun. Combining data-driven and model-driven methods for robust facial landmark detection. *arXiv:1611.10152*, 2016. 10
- [43] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. *arXiv:1701.07717*, 2017. 3
- [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, 3, 5, 6, 7
- [45] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In *ICCV*, 2013. 1, 2, 3, 6, 7, 9
- [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