M²FPA: A Multi-Yaw Multi-Pitch High-Quality Dataset and Benchmark for Facial Pose Analysis

Peipei Li[1,2], Xiang Wu[1], Yibo Hu[1], Ran He[1,2*], Zhenan Sun[1,2]

[1]CRIPAC & NLPR & CEBSIT, CASIA   [2]University of Chinese Academy of Sciences
Email: {peipei.li, yibo.hu}@cripac.ia.ac.cn, alfredxiangwu@gmail.com, {rhe, znsun}@nlpr.ia.ac.cn

Abstract

Facial images in surveillance or mobile scenarios often have large view-point variations in terms of pitch and yaw angles. These jointly occurred angle variations make face recognition challenging. Current public face databases mainly consider the case of yaw variations. In this paper, a new large-scale Multi-yaw Multi-pitch high-quality database is proposed for Facial Pose Analysis (M²FPA), including face frontalization, face rotation, facial pose estimation and pose-invariant face recognition. It contains 397,544 images of 229 subjects with yaw, pitch, attribute, illumination and accessory. M²FPA is the most comprehensive multi-view face database for facial pose analysis. Further, we provide an effective benchmark for face frontalization and pose-invariant face recognition on M²FPA with several state-of-the-art methods, including DR-GAN[24], TP-GAN[10] and CAPG-GAN[8]. We believe that the new database and benchmark can significantly push forward the advance of facial pose analysis in real-world applications. Moreover, a simple yet effective parsing guided discriminator is introduced to capture the local consistency during GAN optimization. Extensive quantitative and qualitative results on M²FPA and Multi-PIE demonstrate the superiority of our face frontalization method. Baseline results for both face synthesis and face recognition from state-of-the-art methods demonstrate the challenge offered by this new database.

1. Introduction

With the development of deep learning, face recognition systems have achieved 99% accuracy [19, 3, 25] on some popular databases [9, 14]. However, in some real-world surveillance or mobile scenarios, the captured face images often contain extreme view-point variations so that

---

*corresponding author

face recognition performance is significantly affected. Recently, the great progress of face synthesis [8, 10, 30] has pushed forward the development of recognition via generation. TP-GAN [10] and CAPG-GAN [8] perform face frontalization to improve recognition accuracy under large poses. DA-GAN [30] is proposed to simulate profile face images, facilitating pose-invariant face recognition. However, their performance often depends on the diversity of pose variations in the training databases.

The existing face databases with pose variations can be categorized into two classes. The ones, such as LFW [9], IJB-A [15] and VGGFace2 [3], are collected from the Internet, whose pose variations follow a long-tailed distribution. Moreover, it is obvious that obtaining the accurate pose labels is difficult for these databases. The others, including CMU PIE [21], CAS-PEAL-R1 [5] and CMU Multi-PIE [7], are captured under the constrained environment across accurate poses. These databases often pay attention to yaw angles without considering pitch angles. However, facial images captured in surveillance or mobile scenarios often have large yaw and pitch variations simultaneously. Such the face recognition across both yaw and pitch angles needs to be extensively evaluated in order to ensure the robustness of recognition system. Therefore, it is crucial to provide researchers with a multi-yaw multi-pitch high-quality face database for facial pose analysis, including face frontalization, face rotation, facial pose estimation and pose-invariant face recognition.

In this paper, a **M**ulti-yaw **M**ulti-pitch high-quality database for **F**acial **P**ose **A**nalysis (M²FPA) is proposed to address this issue. The comparisons with the existing facial pose analysis databases are summarized in Table 1. The main advantages lie in the following aspects: (1) **Large-scale.** M²FPA includes totally 397,544 images of 229 subjects with 62 poses, 4 attributes and 7 illuminations. (2) **Accurate and diverse poses.** We design an acquisition system to simultaneously capture 62 poses, including 13

yaw angles (ranging from $-90°$ to $+90°$), 5 pitch angles (ranging from $-30°$ to $+45°$) and 44 yaw-pitch angles. (3) **High-resolutions.** All the images are captured by the SHL-200WS (2.0-megapixel CMOS camera), which leads to high-quality resolutions (1920×1080). (4) **Accessory.** We use five types of glasses as accessories to further increase the diversity of our database with occlusions.

To the best of our knowledge, M^2FPA is the most comprehensive multi-view face database which covers variations in yaw, pitch, attribute, illumination and accessory. M^2FPA will provide researchers developing and evaluating the new algorithms for facial pose analysis, including face frontalization, face rotation, facial pose estimation and pose-invariant face recognition. Furthermore, in order to provide an effective benchmark for face frontalization and pose-invariant face recognition on M^2FPA, we implement and evaluate several state-of-the-art methods, including DR-GAN[24], TP-GAN[10] and CAPG-GAN[8].

In addition, we propose a simple yet effective parsing guided discriminator, which introduces parsing maps [17] as a flexible attention to capture the local consistency during GAN optimization. First, a pre-trained facial parser captures the three local masks, including hairstyle, skin and facial features (eyes, nose and mouth). Second, we treat these parsing masks as the soft attention, facilitating the synthesized frontal images and the ground truth. Then, these local features are fed into a discriminator, called parsing guided discriminator, to ensure the local consistency of the synthesized frontal images. In this way, we can synthesize photorealistic frontal images with extreme yaw and pitch variations on M^2FPA and Multi-PIE databases.

The main contributions of this paper are as follows:

- We introduce a Multi-yaw Multi-pitch high-quality database for Facial Pose Analysis (M^2FPA). It contains 397,544 images of 229 subjects with yaw, pitch, attribute, illumination and accessory.

- We provide a comprehensive qualitative and quantitative benchmark of several state-of-the-art methods for face frontalization and pose-invariant face recognition, including DR-GAN[24], TP-GAN[10] and CAPG-GAN[8], on M^2FPA.

- We propose a simple yet effective parsing guided discriminator, which introduces parsing maps as a soft attention to capture the local consistency during GAN optimization. In this way, we can synthesize photorealistic frontal images on M^2FPA and Multi-PIE.

2. Related Work

2.1. Databases

The existing face databases with pose variations can be categorized into two classes. The ones, including LFW [9],

IJB-A [15], VGGFace2 [3], CelebA [17] and CelebA-HQ [12], are often collected from the Internet. Therefore, the pose variations in these databases follow a long-tailed distribution, that is there are lots of nearly frontal faces but few profile ones. In addition, it is expensive to obtain the precious pose labels for these facial images, which leads to difficulties for face frontalization, face rotation and facial pose estimation. The others, such as CMU PIE [21], CMU Multi-PIE [7] and CAS-PEAL-R1 [5], are captured under constrained environment with precise controlling of angles. CUM PIE and CMU Multi-PIE have only yaw angles ranging from $-90°$ to $90°$. CAS-PEAL-R1 contains 14 yaw-pitch angles, but these pitch variations are captured by asking the subjects to look upward/downward, which leads to the inaccurate pose labels. Moreover, in CAS-PEAL-R1, only frontal facial images contain accessory variations. Different from these existing databases, M^2FPA contains variations including attribute, illumination, accessory across precious yaw and pitch angles.

2.2. Face Rotation

Face rotation is an extremely challenging ill-posed task in computer vision. In recent years, benefiting from Generative Adversarial Network (GAN) [6], face rotation has made great progress. Currently, state-of-the-art face rotation algorithms can be categorized into two aspects, including 2D [24, 10, 8, 28, 20, 23] and 3D [27, 30, 4, 29, 2, 18] based methods. For 2D based methods, Tran et.al [24] propose DR-GAN to disentangle pose variations from the facial images. TP-GAN [10] employs a two path model, including global and local generators, to synthesize photorealistic frontal faces. Hu et. al [8] incorporate landmark heatmaps as a geometry guidance to synthesize face images with arbitrary poses. PIM [28] performs face frontalization in a mutual boosting way with a dual-path generator. FaceID-GAN [20] extends the conventional two-player GAN to three players, competing with the generator by disentangling the identities of real and synthesized faces. Considering 3D-based methods, FF-GAN [27] incorporates 3D-MM into GAN to provide the shape and appearance prior. DA-GAN [30] employs a dual architecture to refine a 3D simulated profile face. UV-GAN [4] considers face rotation as a UV map completion task. 3D-PIM [29] incorporates a simulator with a 3D Morphable Model to obtain shape and appearance priors for face frontalization. Moreover, Depth-Net [18] infers plausible 3D transformations from one face pose to another, to realize face frontalization.

3. The M^2FPA Database

In this section, we present an overview of the M^2FPA database, including how it was collected, cleaned, annotated and its statistics. To the best of our knowledge, M^2FPA is the first publicly available database that contains precise and

Table 1. Comparisons of existing facial pose analysis databases. Image Size is the average size across all the images in the database. [*]In Multi-PIE, part of frontal images are 3072×2048 in size, but the most are 640×480 resolution. [+]Images have much background in IJB-A.

Database	Yaw	Pitch	Yaw-Pitch	Attributes	Illuminations	Subjects	Images	Image Size	Controllabled	Size[GB]	Paired	Year
PIE [21]	9	2	2	4	21	68	41,000+	640×486	✓	40	✓	2003
LFW [9]	No label	No label	No label	No label	No label	5,749	13,233	250×250	✗	0.17	✗	2007
CAS-PEAL-R1 [5]	7	2	12	5	15	1,040	30,863	640×480	✓	26.6	✓	2008
Multi-PIE [7]	13	0	2	6	19	337	755,370	640×480[*]	✓	305	✓	2009
IJB-A [15]	No label	No label	No label	No label	No label	500	25,809	1026×698[+]	✗	14.5	✗	2015
CelebA [17]	No label	No label	No label	No label	No label	10,177	202,599	505×606	✗	9.49	✗	2016
CelebA-HQ [12]	No label	No label	No label	No label	No label	No label	30,000	1024×1024	✗	27.5	✗	2017
FF-HQ [13]	No label	No label	No label	No label	No label	No label	70,000	1024×1024	✗	89.3	✗	2018
M^2FPA (Ours)	13	5	44	4	7	229	397,544	1920×1080	✓	421	✓	2019

Figure 1. An example of the yaw and pitch variations in our M^2FPA database. From top to bottom, the pitch angles of the 6 camera layers are $+45°$, $+30°$, $+15°$, $0°$, $-15°$ and $-30°$, respectively. The yaw pose of each image is shown in the green box.

multiple yaw and pitch variations. In the rest of this section, we first introduce the hardware configuration and the data collection. Then we describe the cleaning and annotating procedure. Finally, we present the statistics of M^2FPA, including the yaw and pitch variations, the types of attributes and the positions of illuminations.

3.1. Data Acquisition

We design a flexible multi-camera acquisition system to capture faces with multiple yaw and pitch angles. Figure 2 shows an overview of the acquisition system. It is built by many removable brackets, forming an approximate hemisphere with a diameter of 3 meters. As shown in Figure 3, the acquisition system contains 7 horizontal layers, where the first six (Layer1∼Layer6) are the camera layers and the last one is the balance layer. The interval between two adjacent layers is $15°$. The Layer4 has the same height with the center of hemisphere (red circle in Figure 3). Therefore, we set the pitch angle of Layer4 to $0°$. As a result, from top to bottom, the intervals between the rest 5 camera layers and the Layer4 are $+45°$, $+30°$, $+15°$, $-15°$ and $-30°$, respectively.

A total of 62 SHL-200WSs (2.0-megapixel CMOS camera with 12mm prime lens) are located on these 6 camera layers. As shown in Figure 3, there are 5, 9, 13, 13, 13 and 9 cameras on the Layer1, 2, 3, 4, 5 and 6, respectively. For each layer, the cameras are evenly located from $-90°$ to $+90°$. The detailed yaw and pitch angles of each camera can be found in Figure 1 and Table 2. All the 62 cameras are connected to 6 computers through USB interfaces and a master computer synchronously dominates these computers. We develop a software to simultaneously control the 62 cameras and collect all the 62 images in one shot to en-
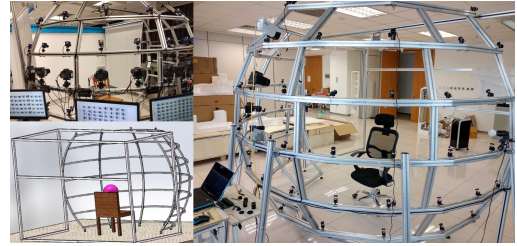

Figure 2. Overview of the acquisition system. It contains total 7 horizontal layers. The bottom is the balanced layer and the rest are the camera layers.
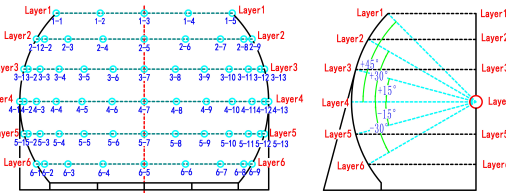

Figure 3. The diagram of camera positions. The left and right are the cutaways of frontal and side views, respectively.

sure the consistency. In addition, as described in Figure 4, there are 7 different directions of light source equipped on our acquisition system, including above, front, front-above, front-below, behind, left and right. In order to maintain the consistency of the background, we construct some brackets and white canvas behind the acquisition system, as shown in the upper left corner in Figure 2.

A total of 300 volunteers are chosen to create the M^2FPA and all the participants have signed a license. During the collection procedure, we fix a chair and provide a headrest to ensure position of face is at the center of hemisphere. Each participant has 4 attributes, including neutral, wearing glass, smile and surprise. Figure 5 shows some examples of
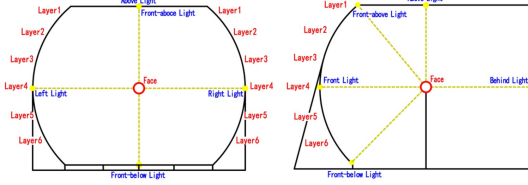
Figure 4. The diagram of illumination positions. The left and right are the cutaways of frontal and side views, respectively.
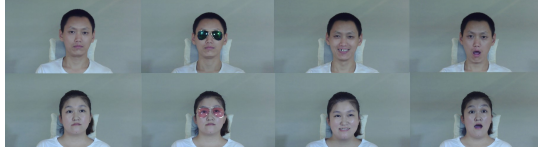

Figure 5. Examples of four attributes in M²FPA.

the attributes. Therefore, we totally capture $300 \times 62 \times 7 \times 4 = 520,800$ ($participants \times poses \times illuminations \times attributes$) facial images.

3.2. Data Cleaning and Annotating

After collection, we manually check all the facial images and remove those participants whose entire head is not captured by one or more cameras. In the end, we eliminate 71 participants with information missing, and the remaining 229 participants form our final M²FPA database. Facial landmark detection is an essential preprocessing in facial pose analysis, such as face rotation and pose-invariant face recognition. However, current methods [1, 22] often fail to accurately detect facials landmarks with extreme yaw and pitch angles. In order to ease the utilization of our database, we manually mark the five facial landmarks of each image in M²FPA.

3.3. The Statistics of M²FPA

Table 2. The poses, attributes and illuminations in M²FPA.

Poses	Pitch = $+45°$	Yaw = $-90°, -45°, 0°, +45°, +90°$
	Pitch = $+30°$	Yaw = $-90°, -67.5°, -45°, -22.5°$ $0°, +22.5°, +45°, +67.5°, +90°$
	Pitch = $+15°$	Yaw = $-90°, -75°, -60°, -45°, -30°, -15°$ $0°, +15°, +30°, +45°, +60°, +75°, +90°$
	Pitch = $0°$	Yaw = $-90°, -75°, -60°, -45°, -30°, -15°$ $0°, +15°, +30°, +45°, +60°, +75°, +90°$
	Pitch = $-15°$	Yaw = $-90°, -75°, -60°, -45°, -30°, -15°$ $0°, +15°, +30°, +45°, +60°, +75°, +90°$
	Pitch = $-30°$	Yaw = $-90°, -67.5°, -45°, -22.5°$ $0°, +22.5°, +45°, +67.5°, +90°$
Attributes		Happy, Normal, Wear glasses, Surprise
Illuminations		Above, Front, Front-above, Behind Front-below, Left, Right

After manually cleaning, we retain 397,544 facial images of 229 subjects, covering 62 poses, 4 attributes and 7 illuminations. Table 2 presents the poses, attributes and illuminations of our M²FPA database. Compared with the existing facial pose analysis databases, as summarized in Table 1, the main advantages of M²FPA lie in four-folds:

- **Large-scale.** M²FPA contains total 397,544 facial images of 229 subjects with 62 poses, 4 attributes and 7 illuminations. It spends almost one year to establish the multi-camera acquisition system and collect such a number of images.

- **Accurate and diverse poses.** Our acquisition system can simultaneously capture 62 poses in one shot, including 13 yaw angles (ranging from $-90°$ to $+90°$), 5 pitch angles (ranging from $-30°$ to $+45°$) and 44 yaw-pitch angles. To the best of our knowledge, M²FPA is the first publicly available database that contains precise and multiple yaw and pitch angles.

- **High-resolution.** All the images are captured by the SHL-200WS (2.0-megapixel CMOS camera), leading to high resolution (1920×1080).

- **Accessory.** In order to further increase the diversity of M²FPA, we add five types of glasses as the accessories, including dark sunglasses, pink sunglasses, round glasses, librarian glasses and rimless glasses.

4. Approach

In this section, we propose a parsing guided local discriminator into GAN training, as is shown in Figure 6. We introduce parsing maps [17] as a flexible attention to capture the local consistency of the real and synthesized frontal images. In this way, our method can effectively frontalize a face with yaw-pitch variations and accessory occlusions on the new M²FPA database.

4.1. Network Architecture

Given a profile facial image X and its corresponding frontal face Y, we can obtain the synthesized frontal image $\hat{Y}$ by a generator G_{θ_G},

$$\hat{Y} = G_{\theta_G}(X) \tag{1}$$

where θ_G is the parameter of G_{θ_G}. The architecture of generator is detailed in Supplementary Materials.

As shown in Figure 6, we introduce two discriminators during GAN optimization, including a global discriminator $D_{\theta_{D1}}$ and a parsing guided local discriminator $D_{\theta_{D2}}$. Specially, the discriminator $D_{\theta_{D1}}$ aims to distinguish the real image Y and the synthesized frontal image $\hat{Y}$ from a global view. Considering photo-realistic visualizations, especially for faces with extreme yaw-pitch angles or accessory, it is crucial to ensure the local consistency between the synthesized frontal image and the ground truth. First, we utilize a pre-trained facial parser f_P [16] to capture three local
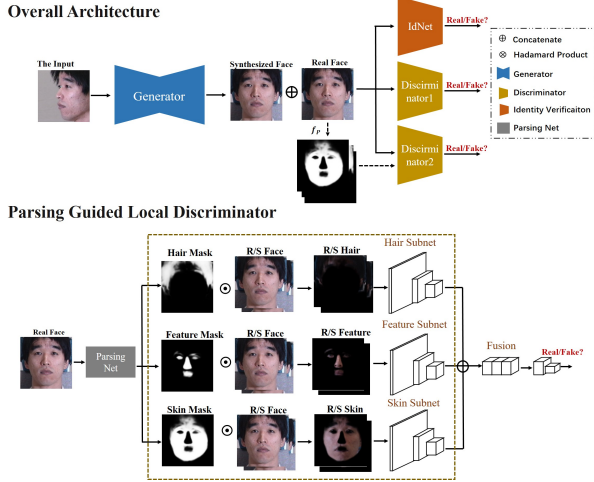
Overall Architecture

Parsing Guided Local Discriminator

Figure 6. The overall framework of our method.

masks, including the hairstyle mask M_h, the skin mask M_s and the facial feature mask M_f from the real frontal image Y,

$$M_h, M_s, M_f = f_P(Y) \quad (2)$$

where the values of three masks are ranged from 0 to 1. Second, we treat these masks as the soft attention, facilitating the synthesized frontal image $\hat{Y}$ and the ground truth Y as follows:

$$Y_h = Y \odot M_h, Y_s = Y \odot M_s, Y_f = Y \odot M_f \quad (3)$$

$$\hat{Y}_h = \hat{Y} \odot M_h, \hat{Y}_s = \hat{Y} \odot M_s, \hat{Y}_f = \hat{Y} \odot M_f \quad (4)$$

where $\odot$ denotes the hadamard product. Y_h, Y_s and Y_f denote the hairstyle, skin and facial feature information from Y, while $\hat{Y}_h$, $\hat{Y}_s$ and $\hat{Y}_f$ are from $\hat{Y}$. Then these local features are fed into the parsing guided local discriminator $D_{\theta_{D2}}$. As shown in Figure 6, three subnets are used to encode the output feature maps of the hairstyle, skin and facial features, respectively. Finally, we concatenate the three encoded feature maps and feed it with binary cross entropy loss to distinguish that the input of local features is real or fake. The parsing guided local discriminator can efficiently ensure whether the local consistency of the synthesized frontal images is similar with the ground truth or not.

4.2. Training Losses

Multi-Scale Pixel Loss. Following [8], we employ a multi-scale pixel loss to enhance the content consistency between the synthesized $\hat{Y}$ and the ground truth Y.

$$L_{pixel} = \frac{1}{3} \sum_{i=1}^{3} \frac{1}{W_i H_i C} \sum_{w,h,c=1}^{W_i, H_i, C} \left| \hat{Y}_{i,w,h,c} - Y_{i,w,h,c} \right| \quad (5)$$

where C is the channel number, i is the i-th image scale, $i \in \{1, 2, 3\}$. W_i and H_i represent the width and height of the i-th image scale, respectively.

Global-Local Adversarial Loss. We adopt a global-local adversarial loss, aiming at synthesizing photo-realistic frontal face images. Specifically, the global discriminator $D_{\theta_{D1}}$ distinguishes the synthesized face image $\hat{Y}$ from real image Y.

$$L_{adv1} = \min_{\theta_G} \max_{\theta_{D1}} E_{Y \sim P(Y)}[\log D_{\theta_{D1}}(Y)] \\ + E_{Y \sim P(Y)}[\log(1 - D_{\theta_{D1}}(Y))] \quad (6)$$

The parsing guided local discriminator $D_{\theta_{D2}}$ aims to make the synthesized local facial details $\hat{Y}_h$, $\hat{Y}_s$ and $\hat{Y}_f$ close to the real Y_h, Y_s and Y_f,

$$L_{adv2} = \min_{\theta_G} \max_{\theta_{D2}} E_{Y_h, Y_s, Y_f \sim P(Y_h, Y_s, Y_f)}[\log D_{\theta_{D2}}(Y_h, Y_s, Y_f)] \\ + E_{Y_h, Y_s, Y_f \sim P(\hat{Y}_h, \hat{Y}_s, \hat{Y}_f)}[\log(1 - D_{\theta_{D2}}(\hat{Y}_h, \hat{Y}_s, \hat{Y}_f))] \quad (7)$$

Identity Preserving Loss. An identity preserving loss is employed to constrain the identity consistency between $\hat{Y}$ and Y. We utilize a pre-trained LightCNN-29 [25] to extract the identity features from $\hat{Y}$ and Y. The identity preserving loss is as follows:

$$L_{id} = ||\varphi_f(Y) - \varphi_f(\hat{Y})||_2^2 \\ + ||\varphi_p(Y) - \varphi_p(\hat{Y})||_F^2 \quad (8)$$

where φ_f and φ_p denote the fully connected layer and the last pooling layer of the pre-trained LightCNN, respectively. $||\cdot||_2$ and $||\cdot||_F$ represent the vector 2-norm and matrix F-norm, respectively.

Total Variation Regularization. We introduce a total variation regularization term [11] to remove the unfavorable artifacts.

$$L_{tv} = \sum_{c=1}^{C} \sum_{w,h=1}^{W,H} \left| \hat{Y}_{w+1,h,c} - \hat{Y}_{w,h,c} \right| + \left| \hat{Y}_{w,h+1,c}^b - \hat{Y}_{w,h,c} \right| \quad (9)$$

where C, W and H are the channel, width and height of the synthesized image Y, respectively.

Overall Loss. Finally, the total supervised loss is a weighted sum of the above losses. The generator and two discriminators, including a global discriminator and a parsing guided local discriminator, are trained alternately to play a min-max problem. The overall loss is written as:

$$L = \lambda_1 L_{pixel} + \lambda_2 L_{adv1} + \lambda_3 L_{adv2} + \lambda_4 L_{id} + \lambda_5 L_{tv} \quad (10)$$

where λ_1, λ_2, λ_3, λ_4 and λ_5 are the trade-off parameters.

5. Experiments

We evaluate our method qualitatively and quantitatively on the proposed M^{2}FPA database. For qualitative evaluation, we show the results of face frontalization on several

yaw and pitch faces. For quantitative evaluation, we perform pose-invariant face recognition based on both the original and synthesized face images. We also provide three face frontalization benchmarks on M^{2}FPA, including DR-GAN [24], TP-GAN [10] and CAPG-GAN [8]. To further demonstrate the effectiveness of the proposed method and assess the difficulty of M^{2}FPA, we also conduct experiments on Multi-PIE [7] database, which is widely used in facial pose analysis. In the following subsections, we begin with an introduction of databases and settings, especially the training and testing protocols of M^{2}FPA. Then we present the qualitative frontalization results and quantitative recognition results on M^{2}FPA and Multi-PIE. Lastly, we conduct ablation study to demonstrate the effect of each part in our method.

5.1. Databases and Settings

Databases. The M^{2}FPA database totally contains 397,544 images of 229 subjects under 62 poses, 4 attributes and 7 illuminations. 57 of 62 poses are chosen in our experiments, except for $+45°$ pitch angles. We randomly select 162 subjects as the training set, i.e., $162 \times 57 \times 4 \times 7 = 258,552$ images in total. The remaining 67 subjects form the testing set. For testing, one gallery image with frontal view, neutral attribute and above illumination is employed for each of the 67 subjects. The remaining yaw and pitch face images are treated as probes. The number of the probe and gallery images are 105,056 and 67 respectively. We will release the original M^{2}FPA database together with the annotated five facial landmarks and the training and testing protocols.

The Multi-PIE database [7] is a popular database for evaluating face synthesis and recognition across yaw angles. Following [8], we use Setting 2 protocol in our experiments. There are 161,460, 72,000, 137 images in the training, probe and gallery sets, respectively.

Implementation Details. Following the previous methods [24, 10, 8], we crop and align 128×128 face images on M^{2}FPA and Multi-PIE for experimental evaluation. Besides, we also conduct experiments on 256×256 face images on M^{2}FPA for high-resolution face frontalization under multiple yaw and pitch variations. A pre-trained LightCNN-29 [25] is chosen for calculating the identity preserving loss and is fixed during training. Our model is implemented with Pytorch. We choose Adam optimizer with the β_1 of 0.5 and β_2 of 0.99. The learning rate is initialized by $2e-4$ and linearly decayed by $2e-5$ after each epoch until 0. The batch size is 16 for 128×128 resolution and 8 for 256×256 resolution on a single NVIDIA TITAN Xp GPU with 12G memory. In all experiments, we empirically set the trade-off parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 to 20, 1, 1, 0.08 and $1e-4$, respectively.

5.2. Evaluation on M^{2}FPA

5.2.1 Face Frontalization

The collected M^{2}FPA database provides a possibility for face frontalization under various yaw and pitch angles. Benefiting from the global-local adversary, our method can frontalize facial images with large yaw and pitch angles. The synthesis results of $+60°\sim+90°$ yaw angles and $-30°\sim+30°$ pitch angles are shown in Figure 7. We observe that not only the global facial structure but also the local texture details are recovered in an identity consistent way. Surprisingly, the sunglasses under extreme poses can also be well preserved. Besides, the current databases for large pose face frontalization are limited to yaw angles and a low resolution, i.e. 128×128. The collected M^{2}FPA has higher quality and supports for face frontalization at 256×256 resolution with multiple yaw and pitch angles. The frontalized 256×256 results of our method on M^{2}FPA are presented in Figure 8, where high quality and photo-realistic frontal faces are obtained. More frontalized results are listed in supplementary materials.

In addition, we provide several benchmark face frontalization results on M^{2}FPA, including DR-GAN[24], TP-GAN[10], and CAPG-GAN[8]. We re-implement CAPG-GAN and TP-GAN according to the original papers. For DR-GAN, we provide two results: one is the re-implemented version[1] and the other is the online demo[2]. Figure 8 presents the comparison results. We observe that our method, CAPG-GAN and TP-GAN achieve good visualizations, while DR-GAN fails to preserve the attributes and the facial structures due to its unsupervised learning procedure. However, there are also some unsatisfactory synthesized details among most of the methods, such as the hair, the face shape. These demonstrate the difficulties of synthesizing photorealistic frontal faces from extreme yaw and pitch angles. Therefore, we expect that collected M^{2}FPA pushes forward the advance in multiple yaw and pitch face synthesis.

5.2.2 Pose-invariant Face Recognition

Face recognition accuracy is a commonly used metric to evaluate the identity preserving ability of different frontalization methods. The better the recognition accuracy, the more identity information is preserved during the synthesis process. Hence, we quantitatively evaluate our method and compare it with several state-of-the-art frontalization methods on M^{2}FPA, including DR-GAN[24], TP-GAN[10], and CAPG-GAN[8]. We employ two open-source pre-trained
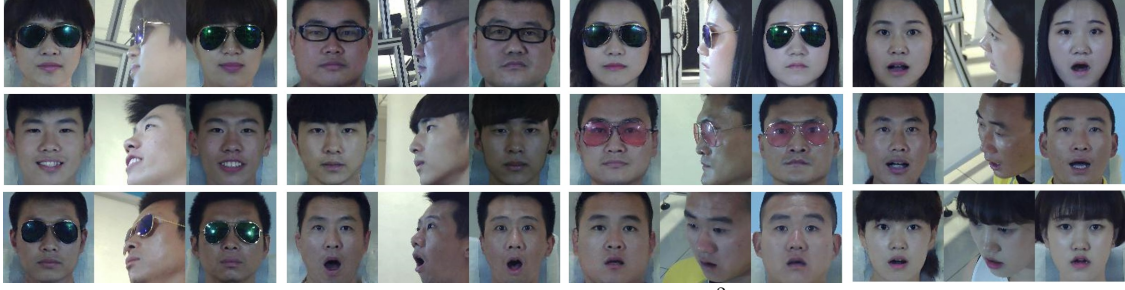
---

[1]`https://github.com/zhangjunh/DR-GAN-by-pytorch`
[2]`http://cvlab.cse.msu.edu/cvl-demo/DR-GAN-DEMO/index.html`

Figure 7. The frontalized 128×128 results of our method under different poses on M²FPA. From top to bottom, the yaw angles are $+90°$, $+75°$, and $+60°$. For each subject, the first column is the generated frontal image, the second column is the input profile, and the last column is the ground-truth frontal image.

Figure 8. Frontalized results of different methods under extreme poses on M²FPA. For each subject, the first row shows the visualizations (256×256) of our method. From left to right: our frontalized result, the input profile and the groundtruth. The second row shows the frontalized results (128×128) of different benchmark methods. From left to right: CAPG-GAN [8], TP-GAN [10], DR-GAN [24] (96×96) and the online demo.

Table 3. Rank-1 recognition rates (%) across views at $0°$ pitch angle on M²FPA.

Method	$\pm15°$	$\pm30°$	$\pm45°$	$\pm60°$	$\pm75°$	$\pm90°$
LightCNN-29 v2						
Original	100	100	99.8	98.6	86.9	51.7
DR-GAN[24]	98.9	97.9	95.7	89.5	70.3	35.5
TP-GAN[10]	99.9	99.8	99.4	97.3	87.6	62.1
CAPG-GAN[8]	99.9	99.7	99.4	96.4	87.2	63.9
Ours	**100**	**100**	**99.9**	98.4	**90.6**	67.6
IR-50						
Original	99.7	99.7	99.2	97.2	87.2	35.3
DR-GAN[24]	97.8	97.6	95.6	89.9	70.6	26.5
TP-GAN[10]	99.7	99.2	98.2	96.3	86.6	48.0
CAPG-GAN[8]	98.8	98.5	97.0	93.4	81.9	50.1
Ours	**99.5**	**99.5**	**99.0**	**97.3**	**89.6**	**55.8**

recognition models, LightCNN-29 v2[3] and IR-50[4], as the feature extractors and define the distance metric as the average distance between the original image pair and the generated image pair. Tables 3, 4 and 5 present the Rank-1 accuracies of different methods on M²FPA under $0°$, $\pm15°$

---

and $\pm30°$ pitch angles, respectively. When keeping the yaw angle consistent, we observe that the larger the pitch angle, the lower the accuracy is obtained, suggesting the great challenge in pitch variations. Besides, by recognition via generation, TP-GAN, CAPG-GAN and our method achieve better recognition performance than the original data under the large poses, such as $\pm90°$ yaw and $\pm30°$ pitch angles. We further observe that the accuracy of DR-GAN is inferior to the original data. The reason may be that DR-GAN is trained in an unsupervised way and there are too many pose variations in M²FPA.

5.3. Evaluation on Multi-PIE

In this section, we present the quantitative and qualitative evaluations on the popular Multi-PIE [7] database. Figure 9 shows the frontalized image of our method. We observe that our method can achieve photo-realistic visualizations against other state-of-the-art methods, including CAPG-GAN [8], TP-GAN [10] and FF-GAN [27]. Table 6 further tabulates the Rank-1 performance of different methods under the Setting 2 for Multi-PIE. It is obvious that our method outperforms its competitors, including

Table 4. Rank-1 recognition rates (%) across views at $\pm15^\circ$ pitch angle on M^{2}FPA.

Method	Pitch	$\pm0^\circ$	$\pm15^\circ$	$\pm30^\circ$	$\pm45^\circ$	$\pm60^\circ$	$\pm75^\circ$	$\pm90^\circ$
LightCNN-29 v2								
Original	$+15^\circ$	100	100	100	99.8	97.5	76.5	34.3
	-15°	99.9	100	99.8	99.7	97.3	81.8	45.9
DR-GAN[24]	$+15^\circ$	99.1	98.8	98.0	94.8	85.6	61.1	20.8
	-15°	98.1	98.2	96.5	93.3	83.1	62.7	31.0
TP-GAN[10]	$+15^\circ$	99.8	99.8	99.7	99.5	95.7	81.6	50.9
	-15°	99.9	99.9	99.6	99.2	95.9	84.1	56.9
CAPG-GAN [8]	$+15^\circ$	99.8	99.9	99.8	98.9	95.0	81.4	54.4
	-15°	99.8	99.9	99.7	98.7	95.1	85.5	65.6
Ours	$+15^\circ$	**99.9**	**99.9**	**99.8**	**99.7**	**97.5**	**86.2**	**56.2**
	-15°	**99.9**	**99.9**	**99.8**	**99.7**	**97.4**	**88.1**	**66.5**
IR-50								
Original	$+15^\circ$	99.8	99.9	99.6	98.7	95.7	77.1	23.4
	-15°	98.7	99.4	99.2	98.1	95.7	78.8	27.9
DR-GAN[24]	$+15^\circ$	98.5	98.2	97.8	94.0	84.8	60.9	17.0
	-15°	95.8	97.2	96.2	93.3	84.8	60.3	20.8
TP-GAN[10]	$+15^\circ$	99.0	99.6	99.1	98.5	94.7	79.1	40.6
	-15°	98.2	98.9	98.1	97.2	94.8	80.9	43.5
CAPG-GAN [8]	$+15^\circ$	98.9	99.0	98.5	95.8	91.5	75.7	40.7
	-15°	98.5	98.5	97.9	95.3	90.3	76.0	47.8
Ours	$+15^\circ$	**99.7**	**99.6**	**99.4**	**98.7**	**96.1**	**84.5**	**43.6**
	-15°	**98.6**	**99.1**	**98.7**	**98.8**	**96.5**	**83.9**	**49.7**

Table 5. Rank-1 recognition rates (%) across views at $\pm30^\circ$ pitch angle on M^{2}FPA.

Method	Pitch	$\pm0^\circ$	$\pm22.5^\circ$	$\pm45^\circ$	$\pm67.5^\circ$	$\pm90^\circ$
LightCNN-29 v2						
Original	$+30^\circ$	99.7	99.2	96.5	71.6	24.5
	-30°	98.6	98.2	93.6	69.9	22.1
DR-GAN[24]	$+30^\circ$	93.8	91.5	83.4	52.0	16.9
	-30°	91.7	90.6	79.1	46.6	16.6
TP-GAN[10]	$+30^\circ$	99.7	98.8	95.8	77.2	43.4
	-30°	98.2	97.6	93.4	75.7	38.9
CAPG-GAN[8]	$+30^\circ$	98.8	98.4	94.1	79.5	48.0
	-30°	98.9	98.3	93.8	75.3	49.3
Ours	$+30^\circ$	**99.7**	**99.1**	**97.7**	**81.9**	**48.2**
	-30°	**98.9**	**98.7**	**95.8**	**82.2**	**49.3**
IR-50						
Original	$+30^\circ$	99.2	98.1	94.7	73.5	17.6
	-30°	97.1	97.3	93.0	67.2	9.0
DR-GAN[24]	$+30^\circ$	92.9	92.3	83.8	56.4	13.9
	-30°	93.0	92.0	82.1	50.3	7.5
TP-GAN[10]	$+30^\circ$	98.1	97.3	94.4	76.8	34.5
	-30°	95.7	96.1	92.2	71.6	27.5
CAPG-GAN[8]	$+30^\circ$	97.1	96.2	90.5	73.1	34.5
	-30°	95.8	95.4	89.2	67.6	33.0
Ours	$+30^\circ$	**98.6**	**97.8**	**96.0**	**79.6**	**36.4**
	-30°	**97.2**	**97.4**	**95.1**	**76.7**	**33.1**

FIP+LDA[31], MVP+LDA[32], CPF[26], DR-GAN[24], FF-GAN[27], TP-GAN[10] and CAPG-GAN[8].

5.4. Ablation Study

We report both quantitative recognition results and qualitative visualization results of our method and its four variants for a comprehensive comparison as the ablation study.
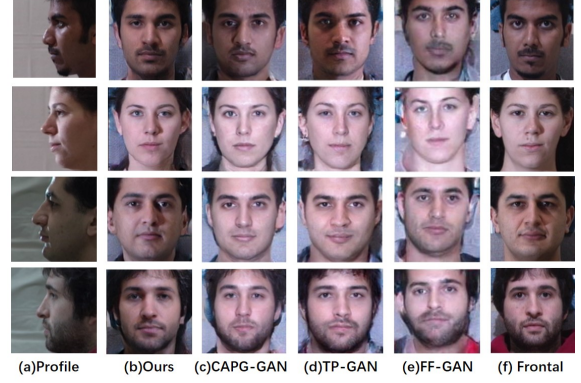

Figure 9. Comparisons with different methods under the pose of 75° (first two rows) and 90° (last two rows) on Multi-PIE.

Table 6. Rank-1 recognition rates (%) across views under Setting 2 on Multi-PIE.

Method	$\pm15^\circ$	$\pm30^\circ$	$\pm45^\circ$	$\pm60^\circ$	$\pm75^\circ$	$\pm90^\circ$
FIP+LDA[31]	90.7	80.7	64.1	45.9	-	-
MVP+LDA[32]	92.8	83.7	72.9	60.1	-	-
CPF[26]	95.0	88.5	79.9	61.9	-	-
DR-GAN[24]	94.0	90.1	86.2	83.2	-	-
FF-GAN[27]	94.6	92.5	89.7	85.2	77.2	61.2
TP-GAN[10]	98.68	98.06	95.38	87.72	77.43	64.64
CAPG-GAN[8]	99.82	99.56	97.33	90.63	83.05	66.05
Ours	99.96	99.78	99.53	96.18	88.74	75.33

We give the details in the Supplemental Materials.

6. Conclusion

This paper has introduced a new large-scale **M**ulti-yaw **M**ulti-pitch high-quality database for **F**acial **P**ose **A**nalysis (M^{2}FPA), including face frontalization, face rotation, facial pose estimation and pose-invariant face recognition. To the best of our knowledge, M^{2}FPA is the most comprehensive multi-view face database that covers variations in yaw, pitch, attribute, illumination, accessory. We also provide an effective benchmark for face frontalization and pose-invariant face recognition on M^{2}FPA. Several state-of-the-art methods, such as DR-GAN, TP-GAN and CAPG-GAN, are implemented and evaluated. Moreover, we propose a simple yet effective parsing guided local discriminator to capture the local consistency during GAN optimization. In this way, we can synthesize photo-realistic frontal images with extreme yaw and pitch variations on Multi-PIE and M^{2}FPA. We believe that the new database and benchmark can significantly push forward the advance of facial pose analysis in community.

7. Acknowledgement

References

[1] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017.

[2] Jie Cao, Yibo Hu, Hongwen Zhang, Ran He, and Zhenan Sun. Learning a high fidelity pose invariant model for high-resolution face frontalization. In *NeurIPS*, 2018.

[3] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 2018.

[4] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *CVPR*, 2018.

[5] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. The cas-peal large-scale chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(1):149–161, 2008.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[7] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[8] Yibo Hu, Xiang Wu, Bing Yu, Ran He, and Zhenan Sun. Pose-guided photorealistic face rotation. In *CVPR*, 2018.

[9] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[10] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *ICCV*, 2017.

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018.

[14] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, 2016.

[15] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *CVPR*, 2015.

[16] Sifei Liu, Jimei Yang, Chang Huang, and Ming-Hsuan Yang. Multi-objective convolutional learning for face labeling. In *CVPR*, 2015.

[17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[18] Joel Ruben Antony Moniz, Christopher Beckham, Simon Rajotte, Sina Honari, and Chris Pal. Unsupervised depth estimation, 3d face rotation and replacement. In *NeurIPS*, 2018.

[19] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, 2015.

[20] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *CVPR*, 2018.

[21] Terence Sim, Simon Baker, and Maan Bsat. The cmu pose, illumination, and expression (pie) database. In *FG*, 2002.

[22] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013.

[23] Yu Tian, Xi Peng, Long Zhao, Shaoting Zhang, and Dimitris N Metaxas. Cr-gan: learning complete representations for multi-view generation. *arXiv preprint arXiv:1806.11191*, 2018.

[24] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017.

[25] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[26] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. Rotating your face using multi-task deep neural network. In *CVPR*, 2015.

[27] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *ICCV*, 2017.

[28] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, et al. Towards pose invariant face recognition in the wild. In *CVPR*, 2018.

[29] Jian Zhao, Lin Xiong, Yu Cheng, Yi Cheng, Jianshu Li, Li Zhou, Yan Xu, Jayashree Karlekar, Sugiri Pranata, Shengmei Shen, et al. 3d-aided deep pose-invariant face recognition. In *IJCAI*, 2018.

[30] Jian Zhao, Lin Xiong, Panasonic Karlekar Jayashree, Jianshu Li, Fang Zhao, Zhecan Wang, Panasonic Sugiri Pranata, Panasonic Shengmei Shen, Shuicheng Yan, and Jiashi Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *NeurIPS*, 2017.

[31] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning identity-preserving face space. In *ICCV*, 2013.

[32] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *NeurIPS*, 2014.