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

Peipei Li, Xiang Wu, Yibo Hu, Ran He and Zhenan Sun
National Laboratory of Pattern Recognition, CASIA, Beijing, China 100190
Center for Research on Intelligent Perception and Computing, CASIA, Beijing, China 100190
University of Chinese Academy of Sciences, Beijing, China 100049
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 viewpoint 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 [27], TP-GAN [12] and CAPG-GAN [10]. 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 [21, 5, 28] on some popular databases [11, 16]. However, in some real-world surveillance or mobile scenarios, the captured face images often contain extreme viewpoint variations so that face recognition performance is significantly affected. Recently, the great progress of face synthesis [10, 12, 33] has pushed forward the development of recognition via generation. TP-GAN [12] and CAPG-GAN [10] perform face frontalization to improve recognition accuracy under large poses. DA-GAN [33] 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 [11], IJB-A [17] and VGGFace2 [3], are collected from the Internet. Considering these databases, the pose variations follow a long-tailed distribution, that is there are few profile faces. Moreover, it is obvious that obtaining the accurate pose labels is difficult for these databases. The others, including CMU PIE [24], CAS-PEAL-R1 [6] and CMU Multi-PIE [8], are captured under the constrained environment across accurate poses. These databases often pay attention to yaw angles without considering pitch angles. However, in many surveillance or mobile scenarios, a facial image often has large yaw and pitch angles simultaneously. Current public face databases mainly consider the case of yaw variations. 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 Multi-yaw Multi-pitch high-quality database for Facial Pose Analysis (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 sub-
projects 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 $-90^\circ$ to $+90^\circ$), 5 pitch angles (ranging from $-30^\circ$ to $+45^\circ$) 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 ($1920 \times 1080$). (4) **Accessory variations.** We use several 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, 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. Further, in order to provide an effective benchmark for face frontalization and pose-invariant face recognition on the M^2FPA databases, we implement and evaluate several state-of-the-art methods, including DR-GAN\[27\], TP-GAN\[12\] and CAPG-GAN\[10\].

In addition, we propose a simple yet effective parsing guided discriminator, which introduces the parsing map \[19\] as a flexible attention to capture local consistency during GAN optimization. First, a pre-trained face 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 facial features (eyes, nose and mouth). Second, we treat these parsing masks as the soft attention, facilitating the facial features (eyes, nose and mouth). Second, we treat these parsing masks as the soft attention, facilitating the facial features (eyes, nose and mouth).

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\[27\], TP-GAN\[12\] and CAPG-GAN\[10\], on the M^2FPA database.

- We propose a simple yet effective parsing guided discriminator, which introduces parsing maps as 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. Some databases, including LFW \[11\], IJB-A \[17\], VGGFace2 \[3\], CelebA \[19\] and CelebA-HQ \[14\], are often collected from the Internet. Therefore, the pose variations in these databases follow a long-tailed distribution, which means 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. Others, such as CMU PIE \[24\], CMU Multi-PIE \[8\] and CAS-PEAL-R1 \[6\], are captured under constrained environment with precise controlling of angles. Both CUM PIE and CMU Multi-PIE vary only yaw angles ranging from $-90^\circ$ to $90^\circ$. CAS-PEAL-R1 contains 14 poses with pitch variations, but these pitch variations are captured by asking the subjects to look upward/downward, which leads to inaccurate pose labels. Moreover, CAS-PEAL-R1 contains accessory variations only with the frontal faces. Different from the 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) \[7\], face rotation has made great progress. Currently, state-of-the-art face rotation algorithms can be categorized into two aspects, including 2D \[27, 12, 10, 31, 23, 26\] and 3D \[30, 33, 9, 32, 2, 20\] based methods. For 2D based methods, Tran \[27\] proposes DR-GAN to disentangle pose variations from the facial images. TP-GAN \[12\] employs a two path model, including global and local generators, to synthesize photo-realistic frontal faces. Hu et. al \[10\] incorporate landmark heatmaps as a geometry guidance to synthesize face images with arbitrary poses. PIM \[31\] performs face frontalization in a mutual boosting way with a dual-path generator. FaceIDGAN \[23\] 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 \[30\] incorporates 3DMM into GAN to provide the shape and appearance prior. Dagan \[33\] 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 \[32\] incorporates a simulator with a 3D Morphable Model to obtain shape and appearance prior for face frontalization. Moreover, DepthNet \[20\] infers plausible 3D transformations from one face pose to another, to realize face frontalization.
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 $3072 \times 2048$ in size, but the most are $640 \times 480$ resolution. †Images have much background in IJB-A.

| Database          | View | Pitch | Yaw-Pitch | Attributes | Illuminations | Subjects | Images | Image Size (GB) | Controllable | Paired | Year |
|-------------------|------|-------|-----------|------------|---------------|----------|--------|-----------------|--------------|--------|------|
| PIE [24]          | 9    | 2     | 2         | 4          | 21            | 68       | 41,000+                        | ✓            | 40     | 2003 |
| LFW [11]          | No label | No label | No label | No label | No label | 5,749   | 13,233 | 250 x 250 | ×           | ×      | 2007 |
| CAS-PEAL-R1 [6]   | 7    | 2     | 12        | 5          | 15            | 1,040    | 30,563 | 640 x 480 | ✓            | 28.6   | 2008 |
| Multi-PIE [8]     | 13   | 0     | 2         | 6          | 19            | 337      | 755,370 | 640 x 480 | ✓            | 305    | 2009 |
| IJB-A [17]        | No label | No label | No label | No label | No label | 500     | 25,809 | 1026 x 698 | ✓            | 14.5   | 2015 |
| CelebA [19]       | No label | No label | No label | No label | No label | 10,177  | 202,599 | 505 x 606 | ✓            | 9.49   | 2016 |
| CelebA-HQ [14]    | No label | No label | No label | No label | No label | 70,000  | 1024   | 1024 x 1024 | ×            | 89.3   | 2017 |
| FF-HQ [15]        | No label | No label | No label | No label | No label | 70,000  | 1024   | 1024 x 1024 | ×            | 89.3   | 2018 |
| M$^2$FPA (Ours)   | 13   | 5     | 44        | 4          | 7             | 229      | 397,244 | 1920 x 1080 | ✓            | 421    | 2019 |

3. The M$^2$FPA Database

In this section we present an overview of the M$^2$FPA database, including how it was collected, cleaned, annotated and its statistics. To the best of our knowledge, M$^2$FPA is the first publicly available database that contains precise and multiple yaw and pitch variations. In the rest of this section, we first introduce the hardware configuration and data collection, then describe the cleaning and annotating procedure. Finally, we present the statistics of M$^2$FPA, including the yaw and pitch poses, 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 $15^\circ$. 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 $0^\circ$. As a result, from top to bottom, the intervals between the rest 5 camera layers and the Layer4 are $+45^\circ$, $+30^\circ$, $+15^\circ$, $-15^\circ$ and $-30^\circ$, 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 $-90^\circ$ to $+90^\circ$. 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 ensure the consistency. In addition, as described in Figure 3, there are 7 different directions of light source equipped on our acquisition system, including above, front, front-above, etc.
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^2-FPA 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 the attributes. Therefore, we totally capture 300 × 62 × 7 × 4 = 520,800 (participants × poses × illuminations × 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 M^2-FPA 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, 25] often fail to accurately detect facial 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 M^2-FPA.

### 3.3. The Statistics of M^2-FPA

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 M^2-FPA database. Compared with the existing facial pose analysis databases, as summarized in Table 1, the main advantages of M^2-FPA lie in four folds:

- **Large-scale.** M^2-FPA 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, M^2-FPA 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 variations.** In order to further increase the diversity of M^2-FPA, 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 map [19] 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 M^2-FPA 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 $G_{θ_2}$,

$$\hat{Y} = G_{θ_2}(X)$$

(1)
where $\theta_c$ is the parameter of $G_{\theta_c}$. 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 $D_{\theta_{D2}}$. Specialy, 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 image $Y$.

First, we utilize a pre-trained facial parser $f_p$ [19] to capture three local masks, including the hairstyle mask $M_h$, the skin mask $M_s$ and the facial feature mask $M_f$ from the real frontal image $Y$,

$$M_h, M_s, M_f = f_p(Y)$$ (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:

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

$$\hat{Y}_h = \hat{Y} \odot M_h, \hat{Y}_s = \hat{Y} \odot M_s, \hat{Y}_f = \hat{Y} \odot M_f$$ (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 $\hat{Y}$. And 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 [10], we employ a multi-scale pixel loss to enhance the content consistency between the synthesized image $\hat{Y}$ and the ground truth image $Y$.

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

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

**Global-Local Adversarial Loss.** We adopt a global-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 $Y$.

$$L_{adv1} = \min_{\theta_c} \max_{\theta_{D1}} E_{Y \sim P(Y)}[\log D_{\theta_{D1}}(Y)] + E_{\hat{Y} \sim P(\hat{Y})}[\log(1 - D_{\theta_{D1}}(\hat{Y}))]$$ (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 $Y_f$,

$$L_{adv2} = \min_{\theta_c} \max_{\theta_{D2}} E_{Y_h, Y_s, Y_f \sim P(Y_h, Y_s, Y_f)}[\log D_{\theta_{D2}}(Y_h, Y_s, Y_f)] + E_{\hat{Y}_h, \hat{Y}_s, \hat{Y}_f \sim P(\hat{Y}_h, \hat{Y}_s, \hat{Y}_f)}[\log(1 - D_{\theta_{D2}}(\hat{Y}_h, \hat{Y}_s, \hat{Y}_f))]$$ (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 [28] to extract the identity features of $\hat{Y}$ and $Y$. The identity preserving loss is as follows:

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

where $\varphi_f$ and $\varphi_p$ 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 [13] to remove the unfavorable artifacts.

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

where $C, W$ and $H$ are the channel, width and height of the synthesized image $\hat{Y}$. 

---

**Figure 6.** The overall framework of our method.
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:

\[ L = \lambda_1 L_{\text{pixel}} + \lambda_2 L_{\text{adv1}} + \lambda_3 L_{\text{adv2}} + \lambda_4 L_{\text{id}} + \lambda_5 L_{\text{tv}} \]  

(10)

where \( \lambda_1, \lambda_2, \lambda_3, \lambda_4 \) and \( \lambda_5 \) 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 [27], TP-GAN [12] and CAPG-GAN [10]. To further demonstrate the effective of the proposed method and assess the difficulty of \( M^2 \)FPA, we also conduct experiments on the Multi-PIE [8] 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 an 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 \(+45^\circ\) 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 [8] is a popular database for evaluating face synthesis and recognition across yaw angles. Following [10], 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 [27] [12] [10], we crop and align 128 \( \times \) 128 face images on \( M^2 \)FPA and Multi-PIE for experimental evaluation. Besides, we also conduct experiments on \( 256 \times 256 \) face images on the \( M^2 \)FPA database for high-resolution face frontalization under multiple yaw and pitch variations. A pre-trained LightCNN-29 [28] 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 \( \beta_1 \) of 0.5 and \( \beta_2 \) 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 \( 128 \times 128 \) resolution and 8 for \( 256 \times 256 \) 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 \( \lambda_5 \) to 0.2, 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 face images with large yaw and pitch variations. The synthesis results of \( +60^\circ \) \( \sim \) \( +90^\circ \) yaw angles and \( -30^\circ \) \( \sim \) \( +30^\circ \) 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. \( 128 \times 128 \). The collected \( M^2 \)FPA has higher quality and supports for face frontalization at \( 256 \times 256 \) resolution with multiple yaw and pitch angles. The frontalized \( 256 \times 256 \) results of our method on \( M^2 \)FPA are presented in Figure 8 where high quality and photorealistic frontal faces are obtained. More frontalized results are listed in supplementary materials due to the page limitation.

In addition, we provide several benchmark face frontalization results on \( M^2 \)FPA, including DR-GAN [27], TP-GAN [12], and CAPG-GAN [10]. 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 [13] and the other is the online demo [14]. Figure 9 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 poses.
Figure 7. The frontalized $128 \times 128$ results of our method under different poses on M$^2$FPA. From top to bottom, the yaw angles are $+90^\circ$, $+75^\circ$, and $+60^\circ$. 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 M$^2$FPA. For each subject, the first row shows the visualizations ($256 \times 256$) of our method. From left to right: our frontalized result, the input profile and the groundtruth. The second row shows the frontalized results ($128 \times 128$) of different benchmark methods. From left to right: CAPG-GAN [10], TP-GAN [12], DR-GAN [27], (96 $\times$ 96) and the online demo.

Table 3. Rank-1 recognition rates (%) across views at $0^\circ$, $15^\circ$, $30^\circ$, $45^\circ$, $60^\circ$, $75^\circ$, and $90^\circ$ 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 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 [27], TP-GAN [12], and CAPG-GAN [10]. We employ two open-source pre-trained recognition models, LightCNN-29 v2 [‡] and IR-50 [§] as the feature extractors and define the distance metric as the average distance of the original image pair and the generated image pair. Table 3 presents the Rank-1 accuracies of different methods on M$^2$FPA under $0^\circ$, $15^\circ$, and $30^\circ$ 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

[‡] https://github.com/AlfredXiangWu/LightCNN

[§] https://github.com/ZhaoJ9014/face.evoLVe._PyTorch
Table 4. Rank-1 recognition rates (%) across views at $\pm 15^\circ$ pitch angle on M$^2$FPA.

| Method      | Pitch  | $\pm 0^\circ$ | $\pm 15^\circ$ | $\pm 30^\circ$ | $\pm 45^\circ$ | $\pm 60^\circ$ | $\pm 75^\circ$ | $\pm 90^\circ$ |
|-------------|--------|---------------|----------------|----------------|----------------|----------------|----------------|---------------|
| Original    | +15$^\circ$ | 100           | 100            | 100            | 99.9           | 97.5           | 76.5           | 34.3          |
|             | -15$^\circ$ | 99.9          | 100            | 99.8           | 99.7           | 97.3           | 81.8           | 45.9          |
| DR-GAN[27]  | +15$^\circ$ | 99.1          | 98.8           | 98.0           | 94.8           | 85.6           | 61.1           | 20.8          |
|             | -15$^\circ$ | 98.1          | 98.2           | 96.5           | 93.3           | 83.1           | 62.7           | 31.0          |
| TP-GAN[12]  | +15$^\circ$ | 99.8          | 99.8           | 99.7           | 99.5           | 95.7           | 81.6           | 50.9          |
|             | -15$^\circ$ | 99.9          | 99.9           | 99.6           | 99.2           | 95.9           | 84.1           | 56.9          |
| CAPG-GAN[10]| +15$^\circ$ | 99.8          | 99.9           | 99.8           | 98.9           | 95.0           | 81.4           | 54.4          |
|             | -15$^\circ$ | 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$^\circ$ | 99.9          | 99.9           | 99.8           | 99.7           | 97.4           | 88.1           | 66.5          |

Table 5. Rank-1 recognition rates (%) across views at $\pm 30^\circ$ pitch angle on M$^2$FPA.

| Method      | Pitch  | $\pm 0^\circ$ | $\pm 22.5^\circ$ | $\pm 45^\circ$ | $\pm 67.5^\circ$ | $\pm 90^\circ$ |
|-------------|--------|---------------|-------------------|----------------|------------------|---------------|
| Original    | +30$^\circ$ | 99.7          | 99.2              | 96.5            | 71.6             | 24.5          |
|             | -30$^\circ$ | 98.6          | 98.2              | 93.6            | 69.2             | 22.1          |
| DR-GAN[27]  | +30$^\circ$ | 93.8          | 91.5              | 83.4            | 52.0             | 16.9          |
|             | -30$^\circ$ | 91.7          | 90.6              | 79.1            | 46.6             | 16.6          |
| TP-GAN[12]  | +30$^\circ$ | 99.7          | 98.8              | 95.8            | 77.2             | 43.4          |
|             | -30$^\circ$ | 98.2          | 97.6              | 93.4            | 75.7             | 38.9          |
| CAPG-GAN[10]| +30$^\circ$ | 98.8          | 98.4              | 94.1            | 79.5             | 48.0          |
|             | -30$^\circ$ | 98.9          | 98.3              | 93.8            | 75.3             | 49.3          |
| Ours        | +30$^\circ$ | 99.7          | 99.1              | 97.7            | 81.9             | 48.2          |
|             | -30$^\circ$ | 98.9          | 98.7              | 95.8            | 82.2             | 49.3          |

IR-50

| Method      | Pitch  | $\pm 0^\circ$ | $\pm 15^\circ$ | $\pm 30^\circ$ | $\pm 45^\circ$ | $\pm 60^\circ$ | $\pm 75^\circ$ | $\pm 90^\circ$ |
|-------------|--------|---------------|----------------|----------------|----------------|----------------|----------------|---------------|
| Original    | +15$^\circ$ | 99.8          | 99.9            | 99.6           | 98.7           | 95.7           | 77.1           | 23.4          |
|             | -15$^\circ$ | 98.7          | 99.4            | 99.2           | 98.1           | 95.7           | 78.8           | 27.9          |
| DR-GAN[27]  | +15$^\circ$ | 98.5          | 98.2            | 97.8           | 94.0           | 84.8           | 60.9           | 17.0          |
|             | -15$^\circ$ | 95.8          | 97.2            | 96.2           | 93.3           | 84.8           | 60.3           | 20.8          |
| TP-GAN[12]  | +15$^\circ$ | 99.0          | 99.6            | 99.1           | 98.5           | 94.7           | 79.1           | 40.6          |
|             | -15$^\circ$ | 98.2          | 98.9            | 98.1           | 97.2           | 94.8           | 80.9           | 43.5          |
| CAPG-GAN[10]| +15$^\circ$ | 98.9          | 99.0            | 98.5           | 95.8           | 91.5           | 75.7           | 40.7          |
|             | -15$^\circ$ | 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$^\circ$ | 98.6          | 99.1            | 98.7           | 98.8           | 96.5           | 83.9           | 49.7          |

5.3. Evaluation on Multi-PIE

In this section, we present the quantitative and qualitative evaluations on the popular Multi-PIE [8] database. Figure 10 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 [10], TP-GAN [12] and FF-GAN [30]. 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 FIP+LDA [35], MVP+LDA [36], CPF [29], DR-GAN [27], FF-GAN [30], TP-GAN [12] and CAPG-GAN [10].

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. We give the details in the Supplemental Materials, due to the page limitation.

6. Conclusion

This paper has introduced a new large-scale multi-yaw multi-pitch high-quality database for Facial Pose Analysis (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 and rotation.
pitch, attribute, illumination, accessory. We also provide an effective benchmark for face frontalization and pose-invariant face recognition on M²-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 M²-FPA. We believe that the new database and benchmark can significantly push forward the advance of facial pose analysis in community.

References

[1] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses an effective benchmark for face frontalization and pose-pitch, attribute, illumination, accessory. We also provide R. Huang, S. Zhang, T. Li, and R. He. Learning a high fidelity pose invariant model for high-resolution face frontalization. In NeurIPS, 2018.

[2] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. FG, 2018.

[3] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou. Uvgan: Adversarial facial uv map completion for pose-invariant face recognition. In CVPR, 2018.

[4] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo. Effective 3d based frontalization for unconstrained face recognition. In IJCV, 2016.

[5] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao. The cas-peak 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] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megface benchmark: 1 million faces for recognition at scale. CVPR, 2016.

[7] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. arXiv preprint arXiv:1812.04948, 2018.

[8] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.

[9] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.

[10] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.
[34] 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.

[35] Z. Zhu, P. Luo, X. Wang, and X. Tang. Deep learning identity-preserving face space. In ICCV, 2013.

[36] Z. Zhu, P. Luo, X. Wang, and X. Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In NeurIPS, 2014.
7. Supplementary Material

In this supplementary material, we first introduce the network architectures of the generator and discriminators in our method. Then we present the ablation study in Section 2. Additional in-the-wild experiments on LFW and CelebA-HQ are shown in Section 3 and 4, respectively. 256×256 frontalization results for all the 57 poses are given in Section 5. Furthermore, in Section 6, we conduct face frontalization of 512×512 resolution on the new M^2-FPA database, which reveals the superiority of M^2-FPA.

7.1. Network Architecture

Our generator $G_{θ_G}$ adopts an encoder-decoder architecture. Taking 256×256 resolution as an example, the detailed structure of $G_{θ_G}$ is listed in Table 7. In the encoder, each convolution layer is followed by one residual block. In the decoder, there are three parts. The first is a simple deconvolution structure to upsample the fc2 features. The second part contains stacked deconvolution layers for reconstruction and each of them is followed by two residual blocks. The third one involves some convolution layers for recovering different scales of face images.

The detailed structures of the global discriminator $D_{θ_D1}$ and the parsing guided local discriminator $D_{θ_D2}$ are shown in Tables 8 and 9 respectively. Each $convk$ in $D_{θ_D1}$ and $D_{θ_D2}$ contains a 3×3 convolution layer, an instance normalization layer and a leaky ReLU layer. The last layers in $D_{θ_D1}$ and $D_{θ_D2}$ produce probabilistic outputs by sigmoid functions.

Table 7. Structure of the generator $G_{θ_G}$.

| Layer | Input     | Filter Size       | Output Size       |
|-------|-----------|-------------------|-------------------|
| conv0 | X         | 7×7/1             | 256×256×64        |
| conv1 | conv0     | 5×5/2             | 128×128×64        |
| conv2 | conv1     | 3×3/2             | 64×64×128         |
| conv3 | conv2     | 3×3/2             | 32×32×256         |
| conv4 | conv3     | 3×3/2             | 16×16×512         |
| fcl   | conv4     | -                 | 512               |
| maxout| fcl       | -                 | 256               |
| fcl2  | n/a       | -                 | 16×16×64          |
| dec0_1| fcl2      | 4×4/4             | 64×64×32          |
| dec0_2| dec0_1    | 2×2/2             | 128×128×16        |
| dec0_3| dec0_2    | 2×2/2             | 256×256×8         |
| dec1  | fcl2, conv4| 2×2/2             | 32×32×512         |
| dec2  | dec1, conv3| 2×2/2             | 64×64×256         |
| dec3  | dec2, conv3, X, dec0_1| 2×2/2             | 128×128×128      |
| dec4  | dec3, conv1, X, dec0_2| 2×2/2             | 256×256×64       |
| conv5 | dec2      | 3×3/1             | 64×64×64          |
| conv6 | dec3      | 3×3/1             | 128×128×32        |
| conv7 | dec4, conv0, X, dec0_3| 5×5/1             | 256×256×3        |
| conv8 | conv7     | 3×3/1             | 256×256×3         |
| conv9 | conv8     | 3×3/1             | 256×256×3         |

Table 8. Structure of the discriminator $D_{θ_D1}$.

| Layer | Input     | Filter Size       | Output Size       |
|-------|-----------|-------------------|-------------------|
| conv1 | Yh/Yf    | 3×3/2             | 128×128×64        |
| conv2 | h_conv1  | 3×3/2             | 64×64×128         |
| conv3 | h_conv2  | 3×3/2             | 32×32×256         |
| conv4 | h_conv3  | 3×3/2             | 16×16×512         |
| conv5 | h_conv4  | 3×3/2             | 8×8×512           |
| s_conv1| Ys/Yf   | 3×3/2             | 128×128×64        |
| s_conv2| s_conv1  | 3×3/2             | 64×64×128         |
| s_conv3| s_conv2  | 3×3/2             | 32×32×256         |
| s_conv4| s_conv3  | 3×3/2             | 16×16×512         |
| s_conv5| s_conv4  | 3×3/2             | 8×8×512           |
| f_conv1| Yf/Yf   | 3×3/2             | 128×128×64        |
| f_conv2| f_conv1  | 3×3/2             | 64×64×128         |
| f_conv3| f_conv2  | 3×3/2             | 32×32×256         |
| f_conv4| f_conv3  | 3×3/2             | 16×16×512         |
| f_conv5| f_conv4  | 3×3/2             | 8×8×512           |

Table 9. Structure of the discriminator $D_{θ_D2}$.

| Layer | Input     | Filter Size       | Output Size       |
|-------|-----------|-------------------|-------------------|
| h_conv1| Yh/Yh   | 3×3/2             | 128×128×64        |
| F_conv1| h, s, f_conv5| 3×3/1             | 8×8×512           |
| F_conv2| F_conv1  | 3×3/2             | 4×4×512           |
| F_conv3| F_conv2  | 1×1/1             | 4×4×1             |

7.2. Ablation Study

In this section, we report both qualitative visualization results and quantitative recognition results for a comprehensive comparison as the ablation study. Figure 10 presents visual comparisons between our method and its four incomplete variants on the new M^2-FPA database. Without the $L_{adv1}$ loss, the synthesized faces are obviously blurry. Without the $L_{lip}$ loss, much identity information is lost during face frontalization. Without $L_{tv}$ loss, there are more artifacts on the synthesized faces. Specially, without the $L_{adv2}$ loss, we observe that the structures of facial features are quite different from the ground truth, where the eyes and mouth have deformations. These indicate that the parsing guided local discriminator can ensure the local consistency between real and synthesized frontal images.

Table 10 further presents the Rank-1 performance of different variants of our method on M^2-FPA. Similar to the visualization ablation study, we observe that the Rank-1 accuracy will decrease if one loss is removed. These phenomena indicate that each component in our method is essential for synthesizing photo-realistic frontal images.
Table 10. Model comparisons: Rank-1 recognition rates (%) on M2FFA.

| Method            | $\pm 15^\circ$ | $\pm 30^\circ$ | $\pm 45^\circ$ | $\pm 60^\circ$ | $\pm 75^\circ$ | $\pm 90^\circ$ |
|-------------------|----------------|----------------|----------------|----------------|----------------|----------------|
| LightCNN-29 v2 w/o $L_{adv,1,2}$ | 99.8           | 99.7           | 99.4           | 97.3           | 86.1           | 63.1           |
| w/o $L_{tv}$     | 99.8           | 99.6           | 99.5           | 97.9           | 88.6           | 67.1           |
| w/o $L_{lp}$     | 99.9           | 99.7           | 99.0           | 96.9           | 86.3           | 56.5           |
| w/o $L_{adv,2}$  | 100            | 100            | 99.7           | 98.4           | 89.3           | 63.5           |
| Ours             | 100            | 99.9           | 98.4           | 90.6           | 67.6           | 67.6           |

| Method            | $\pm 15^\circ$ | $\pm 30^\circ$ | $\pm 45^\circ$ | $\pm 60^\circ$ | $\pm 75^\circ$ | $\pm 90^\circ$ |
|-------------------|----------------|----------------|----------------|----------------|----------------|----------------|
| IR-50 w/o $L_{adv,1,2}$ | 99.7           | 99.3           | 98.3           | 94.9           | 82.1           | 44.9           |
| w/o $L_{tv}$     | 99.4           | 99.4           | 98.5           | 96.2           | 87.7           | 52.0           |
| w/o $L_{lp}$     | 99.2           | 99.0           | 98.3           | 95.3           | 83.8           | 43.4           |
| w/o $L_{adv,2}$  | 99.7           | 99.3           | 98.3           | 95.7           | 82.4           | 45.9           |
| Ours             | 99.5           | 99.5           | 99.0           | 97.3           | 89.6           | 55.8           |

7.3. Additional Results on LFW

Additional frontalization results and comparisons with the previous methods on LFW are shown in Figure 11 and Figure 12, respectively. Same as TP-GAN [12] and CAPG-GAN [10], our model is only trained on Multi-PIE and tested on LFW. In Figure 11 for each subject, the input image is on the left and the frontalized result is on the right. We can observe that both the visual realism and the identity information are well preserved during frontalization. In addition, as shown in Figure 12 our method obtains good visualization results that are comparable to or better than the previous methods, including LFW-3D [9], LFW-HPEN [34], TP-GAN [12], CAPG-GAN [10], our method and the input image.

![Figure 11: Visualization results on LFW. For each subject, the left is the input and the right is the frontalized result.](image1)

![Figure 12: Visualization comparisons on LFW. For each subject, from left to right is the synthesized result of LFW-3D [9], HPEN [34], TP-GAN [12], CAPG-GAN [10], our method and the input image.](image2)

Table 11. Face verification accuracy (ACC) and area-under-curve (AUC) results on LFW.

| Method           | ACC(%) | AUC(%) |
|------------------|--------|--------|
| Ferrari et al. [5] | -      | 94.29  |
| LFW-3D [9]      | 93.62  | 88.36  |
| LFW-HPEN [34]   | 96.25  | 99.39  |
| FF-GAN [30]     | 96.42  | 99.45  |
| CAPG-GAN [10]   | 99.37  | 99.90  |
| Ours             | 99.41  | 99.92  |

7.4. Additional Results on CelebA-HQ

CelebA-HQ [30] is a newly proposed high-quality database with small pose variations for face synthesis. We conduct additional experiments on CelebA-HQ to demonstrate the effectiveness of our method under such in-the-
wild settings. We observe that the images in CelebA-HQ are almost frontal view. In order to take advantage of the high-quality images, following [2], we utilize a 3DMM model [22] to produce the paired profile images for each frontal image. We random choose 3,451 images as the testing set and the frontalization results of our method are presented in Figure 13. Note that there are no overlap subjects between the training and testing sets.

7.5. Additional 256 × 256 Results on M²FPA

Additional 256 × 256 frontalization results under 57 poses on M²FPA are shown in Figure 14. For each subject, the top is the input with different poses and the bottom is the synthesized result. As expected, our method can frontalize the faces with sunglasses. In addition, we also observe that most frontalization results preserve the visual realism and the identity information well, even under extreme yaw and pitch poses.

7.6. Additional 512 × 512 Results on M²FPA

Generating high-resolution results is significant to enlarge the application field of face rotation. However, the current facial pose analysis databases, which are collected in the constrained environment, only provide 128 × 128 images. Our proposed M²FPA supports higher resolution up to 512 × 512 and contains various yaw and pitches angels. Additional 512 × 512 frontalization results of our method on M²FPA are shown in Figure 15. We observe that our high resolution results have richer textures and look more plausible. We believe that the high-resolution M²FPA can push forward the advance of facial pose analysis in mobile or surveillance applications.
Figure 13. High-quality frontalization results on CelebA-HQ. For each subject, the left is the input and the right is the synthesized result.
Figure 14. The $256 \times 256$ frontalization results of our method under 57 poses on M²FPA. From top to bottom, the pitch angles of the Layer 2-6 are $+30^\circ$, $+15^\circ$, $0^\circ$, $-15^\circ$ and $-30^\circ$, respectively. From left to right, the yaw angles are from $-90^\circ$ to $+90^\circ$. For each subject, the top is the input and the bottom is the synthesized result.
Figure 15. The 512×512 frontalization results of our method under extreme poses on M²FPA. For each subject, the bottom left corner is the input image.