Multi-View Face Synthesis via Progressive Face Flow

Yangyang Xu, Xuemiao Xu, Jianbo Jiao, *Member, IEEE*, Keke Li, Cheng Xu, and Shengfeng He, *Senior Member, IEEE*

***Abstract*— Existing GAN-based multi-view face synthesis methods rely heavily on "creating" faces, and thus they struggle in reproducing the faithful facial texture and fail to preserve identity when undergoing a large angle rotation. In this paper, we combat this problem by dividing the challenging large-angle face synthesis into a series of easy small-angle rotations, and each of them is guided by a face flow to maintain faithful facial details. In particular, we propose a Face Flow-guided Generative Adversarial Network (FFlowGAN) that is specifically trained for small-angle synthesis. The proposed network consists of two modules, a face flow module that aims to compute a dense correspondence between the input and target faces. It provides strong guidance to the second module, face synthesis module, for emphasizing salient facial texture. We apply FFlowGAN multiple times to progressively synthesize different views, and therefore facial features can be propagated to the target view from the very beginning. All these multiple executions are cascaded and trained end-to-end with a unified back-propagation, and thus we ensure each intermediate step contributes to the final result. Extensive experiments demonstrate the proposed divide-and-conquer strategy is effective, and our method outperforms the state-of-the-art on four benchmark datasets qualitatively and quantitatively.**

***Index Terms*— Multi-view face synthesis, pose-invariant face recognition, face reconstruction.**

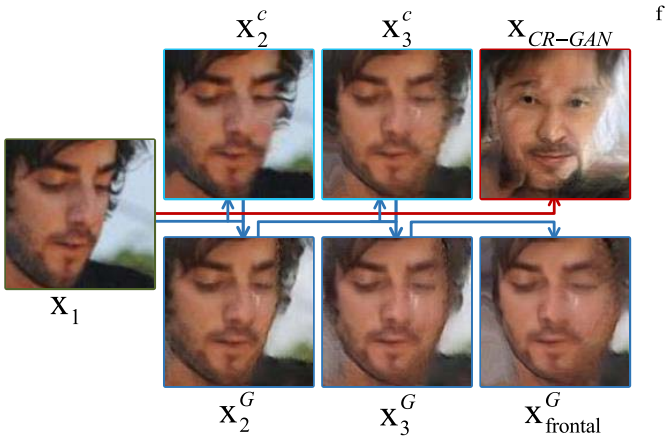


Fig. 1. Directly synthesizing a novel view with large pose variations suffers from erroneous facial details (CR-GAN [40] at the top-right corner). We break down this challenging problem into several small-angle synthesis tasks ($x_\cdot^G$), each of which is guided by a face flow ($x_\cdot^c$ indicates the warped result). A frontal face $x_{frontal}^G$ can be obtained in a progressive manner with faithful facial features.

I. Introduction

FACE recognition is a fundamental problem in computer vision [1], [31], [42]. Although significant progress has been made because of deep learning and large-scale datasets [36], [37], [45], face recognition under unconstrained environments is still a challenging problem [13]–[15], especially with large pose variations [4], [8]. As a result, many works rotate profile faces to the frontal ones, and then the ordinary face recognition methods can be applied.

Generative Adversarial Network (GAN) is arguably the most commonly used technique for synthesizing novel face views [2], [16], [21], [27], [54]. Notwithstanding the demonstrated success, GAN-based methods focus on generating faces that conform to the global distribution, while neglecting local facial details. This problem is minor for small-angle synthesis, but critical for large pose variations. As the input face is heavily occluded, GAN-based methods tend to "*imagine*" the target face rather than *leveraging* input facial features (see the top-right corner of Fig. 1). On the other hand, flow-based warping is an opposite of GAN that reconstructs the target view according to the estimated dense correspondence between input and output faces. However, it is impossible to compute an accurate face flow due to the large pose difference between input and output faces, and most pixels cannot find the correct corresponding points as they are occluded in the input face.

We argue that GAN and flow-based methods are complementary to each other, as they respectively synthesize faces from global and local perspectives. The only issue is that they cannot cooperate in the scenarios with large pose differences. As a consequence, instead of directly generating

Manuscript received June 23, 2020; revised January 14, 2021; accepted June 9, 2021. Date of publication June 28, 2021; date of current version July 2, 2021. This work was supported in part by the Key-Area Research and Development Program of Guangdong Province, China, under Grant 2020B010166003, Grant 2020B010165004, and Grant 2018B010107003; in part by the National Natural Science Foundation of China under Grant 61772206, Grant U1611461, Grant 61472145, and Grant 61972162; in part by the Guangdong International Science and Technology Cooperation Project under Grant 2021A0505030009; in part by the Guangdong Natural Science Foundation under Grant 2021A1515012625; in part by the Guangzhou Basic and Applied Research Project under Grant 202102021074; and in part by the China Computer Federation (CCF)-Tencent Open Research Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christophoros Nikou. *(Corresponding authors: Xuemiao Xu; Shengfeng He.)*

Yangyang Xu, Keke Li, Cheng Xu, and Shengfeng He are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: cnnlstm@gmail.com; cslikeke@mail.scut.edu.cn; cschengxu@gmail.com; hesfe@scut.edu.cn).

Xuemiao Xu is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China, and also with the State Key Laboratory of Subtropical Building Science, Ministry of Education Key Laboratory of Big Data and Intelligent Robot, Guangdong Provincial Key Laboratory of Computational Intelligence and Cyberspace Information, Guangzhou 510006, China (e-mail: xuemx@scut.edu.cn).

Jianbo Jiao is with the Department of Engineering Science, University of Oxford, Oxford OX1 2JD, U.K. (e-mail: jiaojianbo.i@gmail.com).

Digital Object Identifier 10.1109/TIP.2021.3090658

the target view, we break down the challenging large-angle face synthesis problem into a series of easy small-angle face rotation operations. In this way, both of two strategies can maximize their advantages. In particular, we propose a Face Flow-guided Generative Adversarial Network (FFlowGAN) that emphasized on small-angle rotations. It first estimates the face flow by warping the input face to the target one (see the 2nd and 3rd images in the first row of Fig. 1). Because of the small difference between input and output faces, the face flow can faithfully capture facial texture. The obtained face flow is then used to guide the face synthesis module in a multi-scale manner. As a consequence, given a face with an arbitrary view, the proposed FFlowGAN can rotate the input face by a small angle (see the bottom row of Fig. 1 and 15° in our paper), and an arbitrary target view ($0° - 90°$) can be obtained by applying the proposed network. It is extremely flexible, as we, unlike previous works, do not require additional input such as input/output landmarks or the target label vector. More importantly, the progressive process shares the same FFlowGAN, but it is trained end-to-end for the entire rotation sequence with a unified back-propagation. This enforces all the intermediate steps contribute to the final result. We conduct extensive experiments on four benchmark datasets, both qualitative and quantitative results show that our FFlowGAN outperforms state-of-the-art methods.

The contributions of our paper are summarized as follows:

- We delve into the facial displacement consistency of small rotation angles, and propose a unified, end-to-end framework that combats the problem of large pose variations by dividing it into a series of small-angle face matching and synthesis sub-tasks.
- We propose a Face Flow-guided Generative Adversarial Network. It maximizes the advantages of GAN and flow-based strategies. By applying this network, identity-preserved faces can be generated and faithful facial details can be propagated to the target view from the input.
- Experiment results on four datasets demonstrate the effectiveness of our divide-and-conquer solution. Qualitative and quantitative results show superior performances than state-of-the-art methods both under constrained and unconstrained environments.

II. Related Work

A. Pose-Invariant Face Recognition

Large pose variations influence the face recognition performance significantly. Existing methods address pose-invariant face recognition mainly in two ways: the first one aims to learn the pose-invariant features [5], [37], while the other one ensures the recognition accuracy under large rotation angles of profile faces by synthesizing their frontal face [7], [27], [54]. Our method belongs to the latter one to produce identity-preserved faces.

B. Dense Correspondence for Face Synthesis

Dense correspondence describes the matching relationship between image pairs at each pixel. Once the correspondence is acquired, the target image can be reconstructed by warping the source image [39], [56]. There are many works measure the correspondence at image or feature level, such as HOG and SIFT [6], [33]. Meanwhile, many deep learning based methods are proposed to learn a more robust correspondence across images [10], [24], [39], [56]. Dense correspondence has also been introduced to face synthesis. Deng *et al*. [7] present a UV-GAN to complete the facial UV map for real-world faces. Recently, Cao *et al*. [2] propose an AD-GAN for multi-view face image synthesis by face normalizer and editor while estimating the dense UV correspondence field. They also propose a HF-PIM [3] for producing the frontal faces through texture warping. Zhang *et al*. [53] use an appearance-flow-based convolutional neural network for face frontalization, in which the flows are calculated by the alignment of the profile and frontal faces directly. However, for a large-angle profile, the gap between profile and frontal is infeasible to bridge by simply warping the pixels from one to another. In this work, we address this problem by a newly proposed FFlowGAN that learns the unsupervised face flow in a progressive manner, resulting in more accurate facial texture matching. Besides, existing methods directly apply optical flow to produce the warped face as the final result, which introduces the optical flow error to the final prediction. Instead, we treat the face flows as the guidance for multi-scale feature representations, which eliminates flow artifacts through the convolution operations meanwhile preserving the local details.

C. Multi-View Face Synthesis

Early works in the literature only focus on synthesizing the frontal face, which can be considered as single view face synthesis. Traditional works synthesize frontal face by adopting 2D/3D local texture warping [12], [19], [57]. For instance, Hassner *et al*. [19] employ a 3D approximation for face frontalization. However, those methods suffer from severe texture loss and artifacts. Recently, deep learning based methods demonstrate a promising progress. Huang *et al*. [23] propose a TP-GAN for face frontalization by preserving the global and local texture information with complex architecture. Zhao *et al*. [54] propose to learn the pose-invariant features during face frontalization. On the other hand, multi-view face synthesis is much more challenging and more appealing in real-world scenarios. Yim *et al*. [51] utilize multi-task learning for multi-view face synthesis. Tran *et al*. [41] propose to simultaneously recognize pose-invariant face and synthesize multi-view faces. Similar to the methods discussed above, these works struggle in frontalization of large-angle profiles. To tackle this problem, Shen *et al*. [38] edit the faces by interpreting the latent codes of a pre-trained StyleGAN [28], but the results are limited to the same domain of the pretrained generator. On the other hand, Hu *et al*. [21] propose to rotate the input face to multiple views by introducing the guidance of source and target pose landmarks. Xu *et al*. [49] propose a gated deformable network to model the deformation of faces in the rotation process. HoloGAN [35] controls the pose of faces through the transformations of the learned 3D features. Although these methods can handle large pose variations

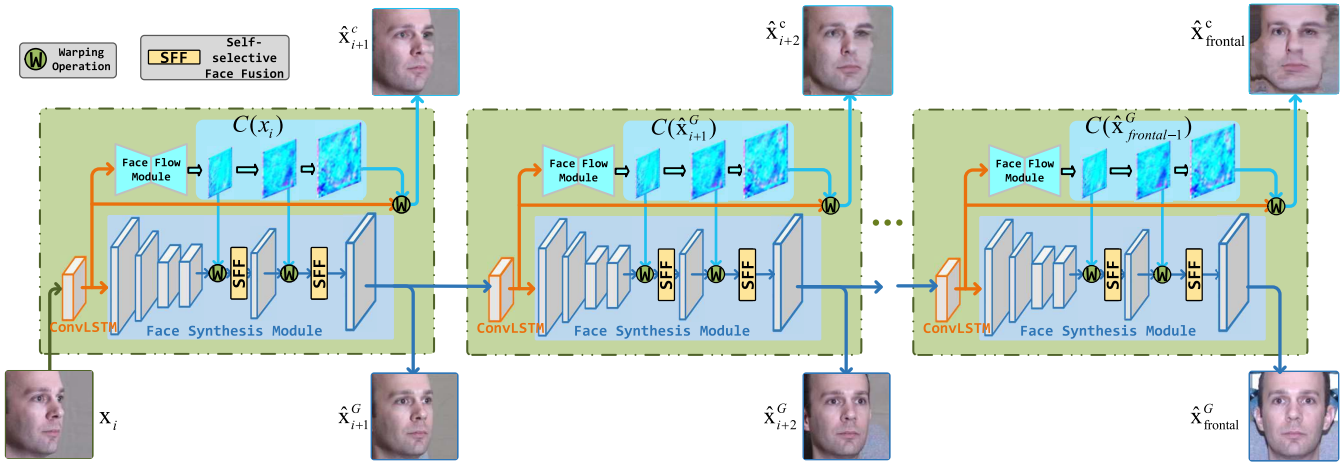


Fig. 2. Overview of our multi-view face synthesis process. The face flow-guided GAN consists of a face flow module and a face synthesis module, in which the flow module produces multi-scale face flows, while the face synthesis module warps the decoder features according to the face flows. The warped features are further processed by a self-selective face fusion module to eliminate artifacts, meanwhile preserving facial details. Note that the parameters of the network are shared across each step (green block) and the discriminator is not shown here for simplicity. x_i is the input face, x^c, x^G denote the faces synthesized by the face flow module and face synthesis module respectively.

to some extent, they formulate multi-view synthesis as a direct mapping problem, which is indeed an intractable and ambiguous problem due to the missing facial details, especially for large rotation angles.

The most related work to our FFlowGAN is SPAE [27], which converts non-frontal faces to frontal ones by stacking shallow auto-encoders. However, SPAE relies only on the observation that each small rotation step should be separately handled, it neglects a critical fact that the displacements between small rotation angles are reliable and consistent across different faces. This crucial finding enables us to employ the unsupervised face flow, and combats the main problem of unreliable matching with large displacements. Enabling face flow is also the key to maintain accurate facial details. Compared with SPAE, our face flow is essential to the success of the progressive rotation strategy, as it injects additional prior knowledge to ensure important facial features can be correctly propagated. On the other hand, progressive strategy is more feasible for integrating the flow-based method on large-angle face synthesis. As a result, SPAE trains several auto-encoders for different rotation steps that can only produce fixed one-to-one face mappings, while our method is a unified, end-to-end framework that can synthesize new faces without knowing the input face angle.

III. Approach

Given a face image x_i at viewpoint i, we aim to synthesize its corresponding faces at different viewpoints progressively under the guidance of face flow. To achieve this goal, we propose a Face Flow-guided Generative Adversarial Network (FFlowGAN). As illustrated in Fig. 2, the proposed FFlowGAN iists of a face flow module and a face synthesis module.

A. Face Flow Module

Previous flow estimation methods [10], [24] take both source and target images as input and be trained under the supervision of ground truth flow. However, the flow ground truth between two faces is unavailable in our case. Instead, we predict the face flow by one source image in an weakly supervised manner. The face flow is learned by warping the source face to the target face. Our face flow module takes one face image x_i as input and predict the corresponding face flow $C(x_i)$ with different scales to the next step, which can be defined as:

$$C(x_i) = \{c^1(x_i), c^2(x_i), \ldots, c^K(x_i)\}, \tag{1}$$

where $C(x_i)$ is the set of face flows with K scales ($K = 3$ in this paper). Each finest face flow $c^K(x_i)$ represents the dense correspondence between the source face and the target one.

The target face x_{i+1} at the current step can be synthesized by warping the source image x_i according to the face flow $c^K(x_i)$ (bilinear sampling layer [25] is used as the warping operation in this paper). Since the angle of each step is very small, the source face x_i is similar to the target face x_{i+1}. As a result, the predicted face flow is accurate for small-angle face reconstruction.

Although the face flow module is able to synthesis the target face $\hat{x}^c_{i+1}$, as we discussed in Sec. I, flow-based method has its own limitations. Thus instead of directly using $\hat{x}^c_{i+1}$ as the final result, we utilize $C(x_i)$ as a guidance for the face synthesis module.

B. Face Synthesis Module

Our face synthesis module considers face flows in a multi-scale manner, to simultaneously generate multi-scale facial features.

1) Face Features Extraction: For a profile face image with large-angle variation, our FFlowGAN synthesizes its frontal face in a progressive way. To capture the correlation of faces between different views and propagate information along the progressive process, we incorporate a ConvLSTM layer [47] at the beginning of the face synthesis module. It receives information from all previous steps such that salient facial features

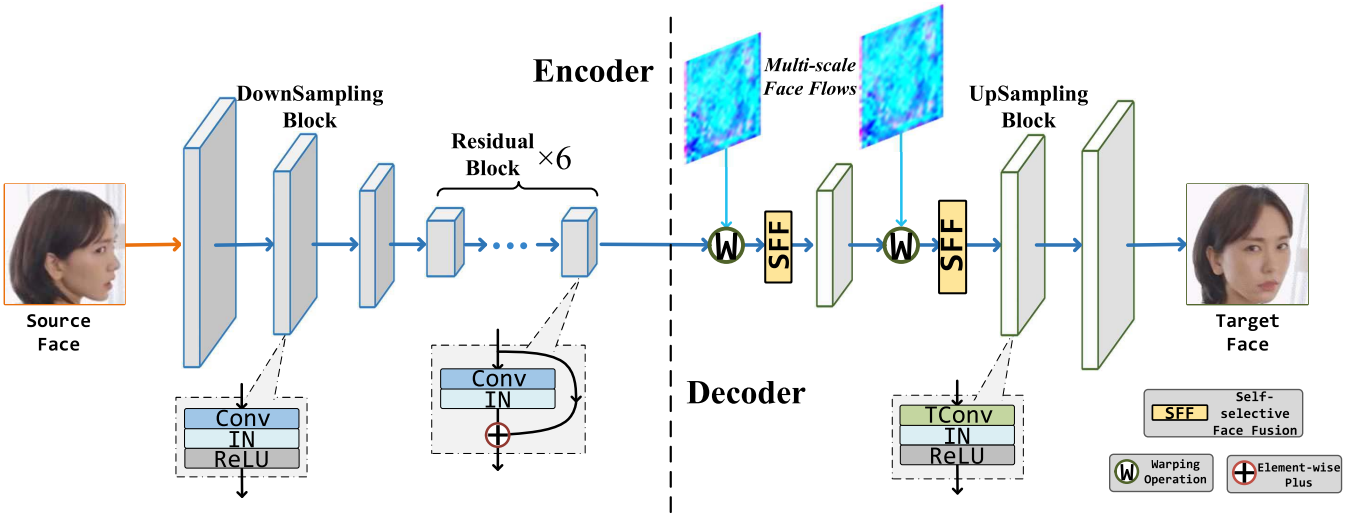


Fig. 3. Architecture of our generator, which has an encoder-decoder architecture. The encoder has 3 down-sampling block, each down-sampling block consists of a convolution layer (**Conv**) followed by an Instance Normalization layer [43] (**IN**) and **ReLU** activation layer [34]). The last part of the encoder contains 6 residual blocks as the bottleneck. The decoder also has 3 up-sampling blocks. Each up-sampling block has a transposed convolutional layers (**TConv**) followed by an Instance Normalization layer and a **ReLU** activation layer, except the last one that we use the **Tanh** as the activation layer. Meanwhile, multi-scale face flows provide guidance in the decoder. A self-selective face fusion module is attached after the face flow-based warping operation.

will not be eliminated during the propagation. Meanwhile, the ConvLSTM module could control the input information to the face synthesis module, and preserves more discriminative identity features of the input face.

Our face synthesis module consists of generator G and discriminator D. As in illustrated in Fig. 3, the generator G is a CNN model with encoder-decoder architecture. It takes a face image processed by the ConvLSTM as input (note that the input face image could be a real image x_i, or a fake face image $\hat{x}^G_{i+1}$ generated by the last step). The encoder has 3 down-sampling blocks (A convolutional layer followed with a Instance Normalization layer [43] and ReLU activation [34] layer) and 6 residual blocks as the bottleneck.

The discriminator D tries to distinguish the real face x_{i+1} from a synthesized one $G(x_i)$. The discriminator consists of 3 down-sampling convolutional layers followed by 6 residual blocks.

2) Face Flow-Based Warping: The decoder of the generator G also has 3 up-sampling blocks. Each block has a transposed convolutional layer followed with Instance Normalization layer [43] and ReLU activation [34] (Tanh as the activation in last convolutional layer). The output features of encoder and 1_{st} layer of decoder can be presented as f^k [1] ($k = 1$ for encoder and $k = 2$ for 1_{st} layer of decoder). The last transposed convolutional layer of the generator produces the final face image.

To leverage the predicted multi-scale face flows, one solution is to construct the target face according to the face flow $c^K(x)$ on source face image x. However, it introduces extra artifacts due to the estimation errors (see the face images x^c_i on the top of Fig. 2). As a consequence, we instead provide face flow guidance to the feature representations, but warping source facial features to follow the target facial patterns.

[1]The specific view i is omitted for simplicity.

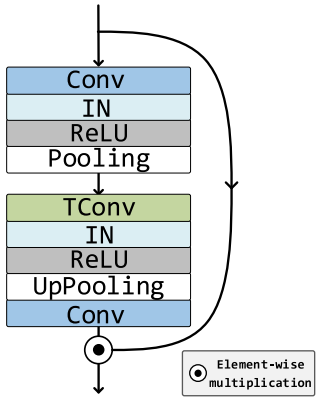


Fig. 4. Architecture of our self-selective face fusion model. It has an encoder-decoder architecture which takes the warped feature f^k_{warped} as input and produces the spatial weighted mask $M^k_{spatial}$. The final output $f^k_{updated}$ is obtained by the element-wise multiplication between f^k_{warped} and $M^k_{spatial}$.

The target-view face images are synthesized by warping the reference-view according to the face flow between adjacent views. The warping operation on feature f^k with its corresponding face flow $c^k(x)$ can be represented as:

$$f^k_w = warp(f^k, c^k(x)), \tag{2}$$

where $warp(\cdot)$ is the warping operation. Face flows are leveraged in each upsampling step, and thus multi-scale face flows are deeply incorporated into the generator.

3) Self-Selective Face Fusion: As mentioned above, flow-based methods can preserve texture details efficiently, whereas often introduce unnatural artifacts, even though we perform warping in feature-level. To further eliminate this issue, we propose a self-selective face fusion in the generator. It aims to adaptively filter out irrelevant artifacts while reinforcing salient facial details. As shown in Fig. 4, the self-selective face fusion module with an encoder-decoder

architecture, the encoder consists of a convolutional layer and a Maxpooling layer, and the decoder consists of a transposed convolutional layer (with Instance Normalization layer [43], ReLU activation [34] and a Uppooling layer in between) and a convolutional layer). The self-selective face fusion takes the warped feature f_w^k as input and produces the spatial weighted mask M_s^k, which has the same height and width with f_w^k with one channel to assign weight for the warped feature f_w^k. Then we update the warped feature by the weighted combination of the corresponding spatial weighted mask M_s^k:

$$M_s^k = conv(Tconv(conv(f_w^k))), \tag{3}$$

$$f_{up}^k = M_s^k \odot f_w^k, \tag{4}$$

where $conv$ and $Tconv$ represent the convolutional and transposed convolutional layers respectively, f_{up}^k denotes the updated features, and $\odot$ denotes element-wise multiplication. In this way, self-selective face fusion is performed for each scale and therefore producing artifacts-eliminated features at multiple levels.

C. Loss Function

Our FFlowGAN model is trained end-to-end by a combined loss function, including the warping-consistency loss, pixel-wise loss, adversarial loss, identity-preserving loss, and a total variation loss.

1) Warping-Consistency Loss: The face flow module is trained under the supervision of target face by a warping-consistency loss. It constrains the module to learn the dense correspondence from the source face to the target one. The ℓ_1 norm is taken as the loss function due to its ability for preserving high frequency information. The corresponding loss can be presented as:

$$\mathcal{L}_{wc} = \frac{1}{n}\|x_{i+1} - warp(x_i, c^K(x_i))\|_1, \tag{5}$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm.

2) Pixel Loss: We first use a ℓ_1 loss function to minimize the pixel-wise difference between generated face image and the ground truth for facilitating the image content consistency, which can be represented as:

$$\mathcal{L}_{pix} = \frac{1}{n}\|x_{i+1} - G(x_i)\|_1. \tag{6}$$

3) Adversarial Loss: For capturing the target distribution in the training process, we train the generator G and its discriminator D in an adversarial manner, where the discriminator tries to distinguish the real face x_{i+1} from a synthesized one $G(x_i)$. In contrast, the generator G tries to fool the discriminator into believing that the synthesized faces are sampled from the real data. Concretely, G and D are trained to alternatively optimize in the following objectives:

$$\mathcal{L}_G = \mathbb{E}_{\boldsymbol{x_i}\sim p_{\text{real}}}[-\log(D(G(\boldsymbol{x_i}))], \tag{7}$$

$$\mathcal{L}_D = \mathbb{E}_{\boldsymbol{x_i}\sim p_{\text{syn}}}[-\log(1 - D(G(\boldsymbol{x_i}))] + \mathbb{E}_{\boldsymbol{x_{i+1}}\sim p_{real}}[-\log(D(\boldsymbol{x_{i+1}}))], \tag{8}$$

where p_{real} and p_{syn} denote the distribution of the real training faces and synthesized faces respectively. Optimizing the abovementioned objectives will push $G(\boldsymbol{x_i})$ to match the data distributions and lead to a photo-realistic result with high frequency details.

4) Identity-Preserving Loss: For multi-view face synthesis, the preservation of identity information of the source face is important. However, the adversarial loss prone to violate the identity guarantee. As a result, the identity-preserving loss is further included for this purpose, which can be represented as:

$$\mathcal{L}_{idt} = \frac{1}{n}\|\phi(x_{i+1}) - \phi(G(x_i))\|_2^2, \tag{9}$$

where $\phi(\cdot)$ denotes the identity feature extracted from the last pooling layer in a pre-trained Light CNN [46] and $\|\cdot\|_2$ is the vector 2-norm. The identity-preserving loss constrains the synthesized face image to be with a small distance to the ground truth image in the feature space, which guarantees the identity between the inputs.

5) Total Variation Loss: To reduce spike artifacts and synthesize a smooth face image, we also introduce the total variation loss $\mathcal{L}_{tv}$ [26] in our training process.

6) Final Objective: The final loss function for training the whole model is a weighted sum of the above losses:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{wc} + \lambda_2\mathcal{L}_{pix} + \lambda_3\mathcal{L}_G + \lambda_4\mathcal{L}_{idt} + \lambda_5\mathcal{L}_{tv}, \tag{10}$$

where λ_1, λ_2, λ_3, λ_4, λ_5 are weighted parameters to balance the five loss terms.

D. Unified Back-Propagation

Our FFlowGAN can handle different small rotation angles. One training strategy is to train the network with arbitrary face pairs (x_i and its next target face x_{i+1}), *i.e.*, no progressive synthesis is involved in training. However, when we apply the model more than once in a long rotation process, the model is not robust for the synthesized faces since all the inputs for training are the real one. Furthermore, each individual rotation step looks at the next target face only, it cannot preserve importance features for long-term propagation. As a result, we propose a unified training strategy for a robust model. Specifically, we train the network in a chain manner. Given an input face x_i, the network synthesizes its target face x_{i+1}^G. Then this output serves as the input x_{i+1}^G of the next step. We iteratively do so until we reach the target view. For each synthesis process, we update the parameters only after the last step. As a consequence, the entire process consists of losses from different steps, and they are combined together for updating the network. In addition, we train our FFlowGAN with various target views (not necessarily 90°), such that the model can be well-trained for synthesizing short- or long-term rotation.

Our model does not require extra information as input (*e.g.*, input face angle or landmarks) and can generate multi-view faces. We apply multiple iterations (at most 6) during training and testing, and the user can pick one according to different purposes (*e.g.*, pick the frontalized result for recognition). In case the user cannot determine the best one, we can estimate the input face angle in advance using PRNet [11] for the determination of iterations numbers.

E. Implementation Details

Our face flow module shares the similar architecture with FlowNet-SD in [24]. We modify the input channel number of the first convolutional layer to 3 while that is 6 in the original FlowNet-SD since it takes two concatenated images as input. Besides, the flow produced by the original FlowNet-SD has a quarter resolution of the input image, we extend the FlowNet-SD with 2 transposed convolutional layers for the final face flow has the same resolution with the input face meanwhile a more accurate dense correspondence. The image size of the source and target faces for training is 128×128. The proposed face flow module could produce 3 face flow images with different scales, from which we utilize the flows with size of 32×32 and 64×64 as the guidance for the face synthesis module.

All the convolution layers in the generator of the face synthesis module have filters with size 3×3 except the filter size in the first layer is 7×7. Besides the residual blocks and the first layer, the stride is set to 2 for all convolution layers in the encoder for halving the height and width of the input and double the feature channels. Meanwhile, the function decoder is a mirror structure with the decoder. We set the channel number of the first convolution layer to 16. In the residual blocks, the stride is set to 1 and keep the feature channel number unchanged. In the self-selective face fusion module, we set the filter size to 2×2 and strides to 2 for pooling layers and 1 for channel number of the last convolution layer. Also, the discriminator has the same architecture with the encoder of generator.

The Light CNN [46] is pre-trained on the MS-Celeb-1M dataset [18] and serves as our feature extractor. We also follow the work [23] that flip the profile faces selectively so that the occluded part are all on the right side for simplicity. Our network is implemented using PyTorch and we train two models by feeding the face view data clockwise and counterclockwise. We train our FFlowGAN with various number of steps using Adam optimizer [29]. The learning rate is set to 0.0001 and weight decay to 0.0005. The batch size is set to 32. We empirically set the weighting parameters as $\lambda_1 = 10$, $\lambda_2 = 10$, $\lambda_3 = 5$, $\lambda_4 = 0.002$ and $\lambda_5 = 0.0001$.

IV. Experiments

Our FFlowGAN has a light-weight network structure, the number of parameters of the generator is 16.39M and 1.1M for the discriminator. Thanks to the light-weighed structure, it takes about 4ms to rotate a given image to the next angle on a single GTX 1080Ti GPU.

A. Datasets and Settings

We evaluate the proposed network on four benchmark datasets:

(1) **Multi-PIE** [17] is the largest multi-view face recognition dataset under controlled environment. It contains 754,204 images of 337 subjects in 15 poses and 20 illumination conditions and is divided into four sessions. Following previous works [21], [23], [55], we use images from 13 poses and 20 illuminations within $\pm 90°$ yaw angles under two different settings. The first setting only uses session-1 (249 subjects) in which the first 149 subjects constitute the training set and the rest 100 subjects for testing. The second setting takes all the subjects (four sessions) into consideration, with the first 200 subjects for training and the rest 137 subjects for testing. During the test phase, the face with neutral expression and illumination of each subject is taken as the gallery while all the other faces as probes.

(2) **IJB-A** [30] is captured in uncontrolled environment, which consists of 5,396 images and 20,412 video frames from 500 subjects with large pose variations. We use the standard evaluation protocol provided by IJB-A, with 10 folders. IJB-A is leveraged to evaluate the performance of our model in the wild with large pose variations when the model is trained on Multi-PIE and CelebA datasets.

(3) **CelebA** [32] is a large dataset in the wild, which consists of 202,599 face images from 10,177 identities with various poses, expressions, and occlusions. Since CelebA does not provide identity labels, we mainly use it for qualitative evaluation. Meanwhile, we also include CelebA into our training set for the evaluation in uncontrolled environments. Furthermore, since CelebA does not include multi-view samples, we follow [58] to generate pair faces with different views for training.

(4) **LFW** [22] is a widely used dataset for face recognition in uncontrolled environment. It contains 13,223 face images from 5,729 subjects with a huge variety of expression, pose, and occlusions. We use LFW to evaluate the performance of our model trained on Multi-PIE and CelebA.

B. Qualitative Evaluation

Fig. 5 shows the qualitative results on Multi-PIE with faces synthesized from 0° to 90° using six input faces at different angles (*i.e.*, 0°, 15°, 45°, 60°, 90°). It shows that our model can rotate the input face to its target face gradually even with a large pose variation. Meanwhile, fine details (*e.g.*, glasses and hairs) and global structure can be well preserved.

Fig. 6 also presents the qualitative comparisons with SPAE [27] for face frontalization on Multi-PIE dataset. Note that the original SPAE only conduct experiments on the profile faces with the maximum angle of 45°, and it was performed on the gray images with the resolution of 40×32. For a fair comparison, we re-implemented their method with the same RGB input as ours and examine larger rotation angles. As we can see that the faces rotated by SPAE are blurry with ghosting artifacts. This is because their progressive-only strategy cannot deal with the highly ambiguous direct face mapping, and severe artifacts will be accumulated in large-angle rotation. This drawback is addressed by introducing our unsupervised face flow.

Fig. 7 shows the frontalization results on IJB-A dataset. Note that our synthesized frontal faces have the same subtle expressions with the input faces, which shows that our FFlowGAN is able to preserve fine facial details.

We also present the frontalization process on the CelebA dataset. As can be seen in Fig. 8, our model could synthesize

Fig. 5. Multi-view face synthesis from 0° to 90° on Multi-PIE. Odd columns are the synthesized results, and even columns are their corresponding ground truth images (inputs are marked in red).

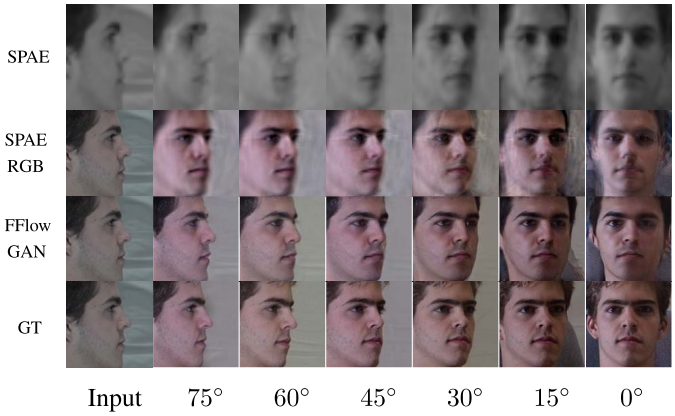


Fig. 6. Qualitative results of FFlowGAN and SPAE [27] on the frontlization with large pose variations (90°) on Multi-PIE.

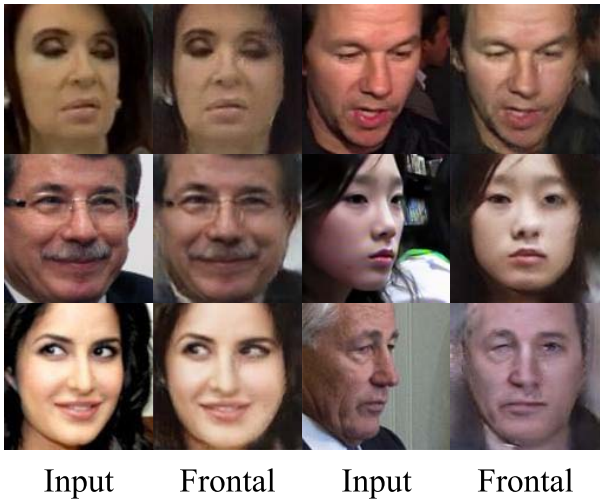


Fig. 7. Frontalization results on the IJB-A dataset. Our synthesized frontal face has the same subtle expressions with its corresponding input face.

the frontal faces gradually with visual-pleasing performance and the identity is well-preserved at the same time, which demonstrates the generalizability of our model in uncontrolled environment.

To further compare with state-of-the-art methods on face frontalization, we perform a qualitative evaluation on the LFW dataset to generate the frontal face for a given profile. As shown in Fig. 9, the frontal faces generated by the competitors show worse performance than ours. The faces rotated by SPAE cannot capture the original identity while being blurry. In addition, some of them also fail to recover clear global structures and vivid details. On the contrary, our method produces more realistic faces with identity well-preserved.

Fig. 8. Our frontalization process on the CelebA dataset. The 1_{st}, 2_{nd} and 3_{rd} raw is the input, intermediate results and the final frontal respectively.

(a) Input (b) Ours (c) SPAE(RGB) (d) DR-GAN (e) CR-GAN (f) CAPG-GAN

Fig. 9. Face frontalization comparison on the LFW dataset.

TABLE I

COMPARISONS WITH STATE-OF-THE-ART FACE SYNTHESIZE METHODS ON RANK-1 RECOGNITION RATES (%) ACROSS VIEWS AND ILLUMINATIONS UNDER SETTING-1 OF MULTI-PIE

Methods	±90°	±75°	±60°	±45°	±30°	±15°
CPF [51]	-	-	-	71.65	81.05	89.45
SPAE [27]	-	-	-	89.40	95.05	97.89
SPAE (RGB) [27]	17.05	35.57	81.75	91.65	97.75	98.95
Hassner *et al.* [19]	-	-	44.81	74.68	89.59	96.78
HPN [9]	29.82	47.57	61.24	72.77	78.26	84.23
FIP [59]	31.37	49.10	69.75	85.54	92.98	96.30
c-CNN [48]	47.26	60.66	74.38	89.02	94.05	96.97
LightCNN [46]	9.00	32.35	73.30	97.45	99.80	99.78
TP-GAN [23]	64.03	84.10	92.93	98.58	99.85	99.78
PIM [54]	75.00	91.20	97.70	98.30	99.40	99.80
CAPG-GAN [21]	77.10	87.40	93.74	98.28	99.37	99.95
Ours	**91.62**	**96.62**	**98.23**	**99.57**	**99.87**	**100**

TABLE II

COMPARISONS WITH STATE-OF-THE-ART FACE SYNTHESIZE METHODS ON RANK-1 RECOGNITION RATES (%) ACROSS VIEWS AND ILLUMINATIONS UNDER SETTING-2 OF MULTI-PIE

Methods	±90°	±75°	±60°	±45°	±30°	±15°
FIP [59]	-	-	45.90	64.10	80.70	90.70
MVP [59]	-	-	45.90	64.10	80.70	90.70
CPF [51]	-	-	61.90	79.90	88.50	95.00
SPAE [27]	-	-	-	84.65	93.45	96.00
SPAE (RGB) [27]	15.23	30.48	71.40	86.60	94.75	97.50
DR-GAN [41]	-	-	83.20	86.20	90.10	94.00
LightCNN [46]	5.51	24.18	62.09	92.13	97.38	98.59
FF-GAN [52]	61.20	77.20	85.20	89.7	92.50	94.60
TP-GAN [23]	64.64	77.43	87.72	95.38	98.06	98.68
CAPG-GAN [21]	66.05	83.05	90.63	97.33	99.56	99.82
PIM [54]	86.50	95.00	98.10	98.50	99.00	99.30
AD-GAN [2]	89.70	95.30	98.80	99.50	99.70	99.80
HF-PIM [3]	92.32	96.40	99.14	**99.88**	99.98	**99.99**
Ours	**93.01**	**97.00**	**99.17**	99.83	**99.99**	**99.99**

C. Quantitative Evaluation

In this section, we quantitatively perform face recognition and verification experiments on three datasets to evaluate the identity-preserving capability of our model. Specifically, we first synthesize the frontal face of all profiles, and then evaluate the face recognition and verification performance accordingly. A pre-trained Light CNN model [46] is employed to extract the features of the input faces, and the cosine similarity between two face feature vectors is taken as the similarity metric.

We evaluate the face recognition performance on Multi-PIE under two experimental settings. The experiment result for setting-1 is shown in Table I. Similar to qualitative evaluation, we show both results of SPAE using grayscale and RGB inputs (denoted with SPAE and SPAE (RGB)). We can see that our FFlowGAN outperforms SPAE and SPAE (RGB) by a large margin for all the angles. Furthermore, we can see that our FFlowGAN outperforms all other competitors for angles of ±90°. Meanwhile, for angles of ±75° and ±60°, our model also shows a strong significance.

The experimental results demonstrate that our model can learn pose-invariant features even though large angles exist. The result on setting-2 is shown in Table II, in which the proposed approach also consistently surpasses its competitors, especially on views with large angle. The above quantitative results validate that our method is able to preserve the identity features effectively.

Table III shows the verification and recognition accuracy performances on the uncontrolled setting of IJB-A dataset. Although our performance is not the best in recognition, but we largely outperform the others in the task of verification.

In Table IV we show the verification accuracy and AUC (area under the curve) results on uncontrolled LFW dataset. Compared with other methods under in-the-wild setting, our model shows a better face verification performance on both ACC and AUC, which demonstrates that our model is robust to uncontrolled faces.

D. Ablation Study

In this section, we analyze the efficacy of the main components in the network and loss functions of the proposed method. We compare seven variants: 1) without face flow module and without the subsequent self-selective face fusion; 2) without the face flow module but with the subsequent

(a) Input (b) GT (c) w/o$\mathcal{L}_{idt}$ (d) w/o$\mathcal{L}_G$ (e) w/o$\mathcal{L}_{pix}$ (f) w/o$\mathcal{L}_{tv}$ (g) w/oF,w/oS (h) w/oF,w/iS (i) w/iF,w/oS (j) FFlowGAN

Fig. 10. Qualitative results of FFlowGAN and its variants on Multi-PIE, From (a) to (j) denote the results of 'Input', 'GT', variant 'w/o $\mathcal{L}_{idt}$', variant 'w/o $\mathcal{L}_G$', variant 'w/o $\mathcal{L}_{pix}$', variant 'w/o $\mathcal{L}_{tv}$', variant 'w/o F, w/o S', variant 'w/o F, w/i S', variant 'w/i F, w/o S' and the final FFlowGAN respectively.

TABLE III

COMPARISONS WITH STATE-OF-THE-ART FACE SYNTHESIZE METHODS ON FACE RECOGNITION AND VERIFICATION ON THE IJB-A DATASET

Methods	Verification		Recognition	
	FAR=0.01	FAR=0.001	Rank-1	Rank-5
OpenBR [50]	23.6 ± 0.9	10.4 ± 1.4	24.6 ± 1.1	37.5 ± 0.8
SPAE [27]	23.3 ± 2.5	12.6 ± 0.6	27.9 ± 1.2	40.9 ± 0.3
SPAE (RGB) [27]	25.6 ± 1.4	12.4 ± 1.4	29.8 ± 1.7	41.2 ± 0.5
TP-GAN [23]	31.5 ± 1.8	9.2 ± 1.1	48.6 ± 5.0	59.3 ± 5.6
GOTS [50]	40.6 ± 1.4	19.8 ± 0.8	44.3 ± 2.1	59.5 ± 2.0
Pooling faces [20]	30.9	-	84.6	-
PAM [44]	73.3 ± 1.8	55.2 ± 3.2	77.1 ± 1.6	88.7 ± 0.9
DR-GAN [41]	77.4 ± 2.7	53.9 ± 4.3	85.5 ± 1.5	94.7 ± 1.1
FF-GAN [52]	85.2 ± 1.0	66.3 ± 3.3	90.2 ± 0.6	95.4 ± 0.5
Light CNN [46]	93.3 ± 1.0	91.5 ± 1.4	91.8 ± 1.7	93.5 ± 1.3
AD-GAN [2]	94.6 ± 0.8	88.9 ± 1.4	$\mathbf{95.9 \pm 0.5}$	$\mathbf{97.0 \pm 0.3}$
Ours	$\mathbf{96.9 \pm 0.8}$	$\mathbf{93.4 \pm 0.8}$	92.0 ± 1.6	95.0 ± 1.4

TABLE IV

COMPARISONS WITH STATE-OF-THE-ART FACE SYNTHESIZE METHODS ON THE FACE VERIFICATION TASK ON THE LFW DATASET. RESULTS FOR ACCURACY (ACC) AND AREA UNDER THE CURVE (AUC) ARE PRESENTED

Methods	ACC(%)	AUC(%)
Ferrari *et al.* [12]	-	94.29
LFW-3D [19]	93.62	88.36
LFW-HPEN [57]	96.25	99.39
FF-GAN [52]	96.42	99.45
CAPG-GAN [21]	99.37	99.90
Ours	**99.48**	**99.93**

TABLE V

RANK-1 RECOGNITION RATES (%) OF OUR FFLOWGAN AND ITS VARIANTS UNDER SETTING-1 OF MULTI-PIE. "F" REPRESENTS THE FACE FLOW MODULE AND "S" DENOTES THE SELF-SELECTIVE FACE FUSION

Methods	$\pm 90°$	$\pm 75°$	$\pm 60°$	$\pm 45°$	$\pm 30°$	$\pm 15°$
w/o $\mathcal{L}_{idt}$	13.71	17.88	27.70	54.49	68.18	65.15
w/o $\mathcal{L}_G$	89.07	96.26	98.94	99.55	99.92	99.97
w/o $\mathcal{L}_{pix}$	90.88	98.13	99.42	99.87	100	99.85
w/o $\mathcal{L}_{tv}$	90.70	96.35	97.41	**99.77**	99.85	99.80
w/o F,w/o S	89.90	89.22	87.15	90.63	95.13	98.23
w/o F,w/i S	88.86	90.40	93.76	92.70	97.90	99.55
w/i F,w/o S	89.73	94.60	97.30	99.32	99.77	99.87
FFlowGAN	**91.62**	**96.62**	**98.23**	99.57	**99.87**	**100**

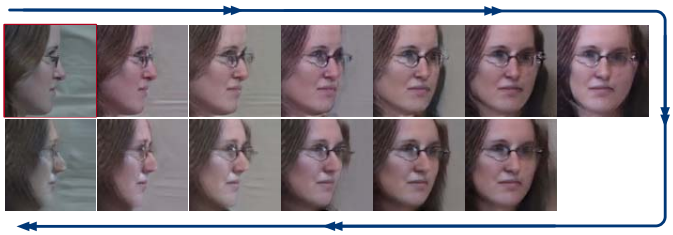


Fig. 11. Qualitative results of FFlowGAN with 12 successive rotation steps on clockwise and counterclockwise models. The input is marked in red and the arrows show the rotation direction.

self-selective face fusion; 3) with the face flow module but without the self-selective face fusion; 4) without the identity information (w/o $\mathcal{L}_{idt}$); 5) without adversarial learning (w/o $\mathcal{L}_G$); 6) without the pixel-wise supervision (w/o $\mathcal{L}_{pix}$) and 7) without addressing the spike artifacts (w/o $\mathcal{L}_{tv}$). We show both qualitative and quantitative results on setting-1 of Multi-PIE.

The face recognition performance is shown in Table V. We can see that the $\mathcal{L}_{idt}$ contributes the most, while $\mathcal{L}_{tv}$ contributes little to the result. Furthermore, for the plain variant that without the face flow module and without self-selective face fusion (w/o F,w/o S), compared with the FFlowGAN, this variant can still keep a good performance on the angle of $\pm 90°$ (89.90% vs. 91.62%). However, it drops significantly on the angles of $\pm 75°$, $\pm 60°$, $\pm 45°$ and $\pm 30°$. Compared with the face flow module without the self-selective face fusion (w/o F,w/i S) and the plain variant (w/o F,w/o S), we can see that (w/o F,w/i S) performs only slightly better than the plain network (w/o F,w/o S) (even worse on the large angle of $\pm 90°$). On the contrary, the variant (w/i F,w/o S) outperforms the plain variant (w/o F,w/o S) by a large margin, which shows that the face flow module contributes the most to

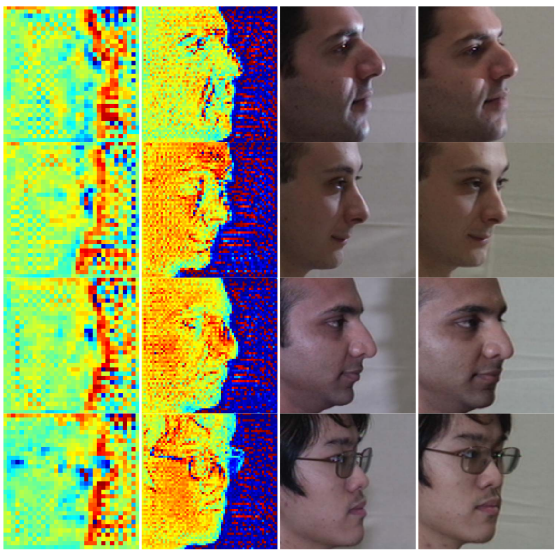


Fig. 12. Visualization results on the spatial weighted mask M_s. The 1_{st}, 2_{nd} columns are the spatial weighted masks with resolutions of 32×32 and 64×64 respectively. And the 3_{rd}, 4_{th} columns indicate the source and the next target faces respectively. The brighter color indicates a higher value and vice versa.

the recognition result. Meanwhile, the final model FFlowGAN (w/i F,w/i S) outperforms all the other variants.

Fig. 10 illustrates the qualitative results of our FFlowGAN and its variants. As expected, each loss term boosts the qualitative performance. If without the face flow module and without the self-selective face fusion, the qualitative results tend to be blurry and lack fine details. On the other hand, if the self-selective face fusion is removed from our FFlowGAN, although some fine details can be preserved, many other artifacts are introduced. The last column shows the results of our final model, which provides more details and is more visual-appealing.

To examine how the performance degrades with respect to rotation steps, we also show the results of chaining 12 rotation steps with our FFlowGAN. As shown in Fig. 11, given a profile face image with angle of 90° as input, our FFlowGAN can rotate it to different views progressively. The upper row presents the rotation results of the clockwise model and the bottom shows the results of counterclockwise model by taking the synthesized frontal face as input. After 12 rotations, we can see that our model can still preserve important facial features and identity, with acceptable accumulative errors.

E. Analysis on Self-Selected Face Fusion

In this section, we give the analysis on our self-selected face fusion by visualizing its output spatial weighted mask M_s. As shown in Fig. 12, images in the first column are M_s^1 produced by the first self-selected face fusion, and those in the second column are M_s^2 by the second SSF block (see "**SSF**" block in Fig. 3). The third columns are the input source faces and fourth columns are their corresponding target faces of the next view. We can see the brighter area (which means a higher value) in M_s^1 mainly lie on the "edges" between the foreground faces and the background. Compared them with source and target faces, we can see that the M_s^1 mainly eliminating the artifacts introduced by face flows globally. Meanwhile, the second spatial weighted mask M_s^2 mainly focuses on the differences between two faces, especially on the local facial textures, such as mouths, noses or eyebrows. Our two self-selected face fusion could eliminate the artifacts in a coarse-to-fine manner. Cooperated with the face flow module, they could preserve the vivid facial details meanwhile eliminating the unsatisfied artifacts.

V. Conclusion

In this paper, we propose a new face flow-guided GAN (FFlowGAN) method for multi-view face synthesis in a progressive manner. Our method simplifies the challenging large-angle face synthesis problem by proposing a series of easy small-angle synthesis steps. The proposed FFlowGAN takes advantages of both adversarial learning and flow-based methods, resulting in identity- and detail-preserved multi-view faces. Extensive experiments demonstrate the effectiveness of the proposed method and it outperforms state-of-the-art methods on four benchmark datasets.

A limitation of our FFlowGAN is that the degree of views must be a multiple of the "small-angle", which is determined by the angle stride of the dataset (*i.e.*, Multi-PIE [17] in our experiment). In the future, we aim to loosen this limitation to a free-form rotation.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
[2] J. Cao, Y. Hu, B. Yu, R. He, and Z. Sun, "3D aided duet GANs for multi-view face image synthesis," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 8, pp. 2028–2042, Aug. 2019.
[3] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Learning a high fidelity pose invariant model for high-resolution face frontalization," in *Proc. NeurIPS*, 2018, pp. 2867–2877.
[4] X. Chai, S. Shan, X. Chen, and W. Gao, "Locally linear regression for pose-invariant face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 7, pp. 1716–1725, Jul. 2007.
[5] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3025–3032.
[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Feb. 2005, pp. 886–893.
[7] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou, "UV-GAN: Adversarial facial UV map completion for pose-invariant face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7093–7102.
[8] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 3, p. 37, 2016.
[9] C. Ding and D. Tao, "Pose-invariant face recognition with homography-based normalization," *Pattern Recognit.*, vol. 66, pp. 144–152, Jun. 2017.
[10] A. Dosovitskiy *et al.*, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.
[11] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3D face reconstruction and dense alignment with position map regression network," in *Proc. ECCV*, 2018, pp. 534–551.
[12] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo, "Effective 3D based frontalization for unconstrained face recognition," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 1047–1052.

[13] S. Ge, C. Li, S. Zhao, and D. Zeng, "Occluded face recognition in the wild by identity-diversity inpainting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3387–3397, Oct. 2020.

[14] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 2051–2062, Apr. 2019.

[15] S. Ge, S. Zhao, C. Li, Y. Zhang, and J. Li, "Efficient low-resolution face recognition via bridge distillation," *IEEE Trans. Image Process.*, vol. 29, pp. 6898–6908, 2020.

[16] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NeurIPS*, 2014, pp. 2672–2680.

[17] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, May 2010.

[18] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. ECCV*, 2016, pp. 87–102.

[19] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4295–4304.

[20] T. Hassner *et al.*, "Pooling faces: Template based face recognition with pooled face images," in *Proc. CVPR Workshops*, 2016, pp. 59–67.

[21] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun, "Pose-guided photorealistic face rotation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8398–8406.

[22] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment, Recognit.*, 2008.

[23] R. Huang, S. Zhang, T. Li, and R. He, "Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2439–2448.

[24] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2462–2470.

[25] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. NeurIPS*, 2015, pp. 2017–2025.

[26] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016, pp. 694–711.

[27] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive auto-encoders (SPAE) for face recognition across poses," in *Proc. CVPR*, Jun. 2014, pp. 1883–1890.

[28] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.

[30] B. F. Klare *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA janus benchmark a," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1931–1939.

[31] C. Li, S. Ge, D. Zhang, and J. Li, "Look through masks: Towards masked face recognition with de-occlusion distillation," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3016–3024.

[32] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[33] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, May 1999, pp. 1150–1157.

[34] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. ICML*, 2010, pp. 807–814.

[35] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y. Yang, "HoloGAN: Unsupervised learning of 3D representations from natural images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7588–7597.

[36] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, vol. 1, 2015, p. 6.

[37] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[38] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9243–9252.

[39] Y. Shih, S. Paris, C. Barnes, W. T. Freeman, and F. Durand, "Style transfer for headshot portraits," *ACM TOG*, vol. 33, no. 4, p. 148, 2014.

[40] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas, "CR-GAN: Learning complete representations for multi-view generation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 942–948.

[41] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1415–1424.

[42] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proc. CVPR*, 1991, pp. 586–591.

[43] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022

[44] D. Wang, C. Otto, and A. K. Jain, "Face search at scale," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1122–1136, Jun. 2017.

[45] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. ECCV*, 2016, pp. 499–515.

[46] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2884–2896, Nov. 2018.

[47] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Proc. NeurIPS*, 2015, pp. 802–810.

[48] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, and T.-K. Kim, "Conditional convolutional neural network for modality-aware face recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3667–3675.

[49] X. Xu, K. Li, C. Xu, and S. He, "GDFace: Gated deformation for multi-view face image synthesis," in *Proc. AAAI*, vol. 34, 2020, pp. 12532–12540.

[50] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in *Proc. CVPR*, 2013, pp. 3539–3545.

[51] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim, "Rotating your face using multi-task deep neural network," in *Proc. CVPR*, 2015, pp. 676–684, 2015.

[52] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Towards large-pose face frontalization in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3990–3999.

[53] Z. Zhang, X. Chen, B. Wang, G. Hu, W. Zuo, and E. R. Hancock, "Face frontalization using an appearance-flow-based convolutional neural network," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2187–2199, May 2019.

[54] J. Zhao *et al.*, "Towards pose invariant face recognition in the wild," in *Proc. CVPR*, 2018, pp. 2207–2216.

[55] J. Zhao *et al.*, "3D-Aided deep pose-invariant face recognition," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, vol. 2, no. 3, p. 11.

[56] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," in *Proc. ECCV*, 2016, pp. 286–301.

[57] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 787–796.

[58] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3D total solution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 78–92, Jan. 2019.

[59] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 113–120.

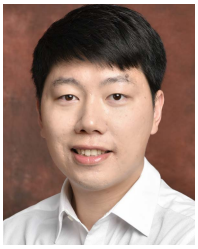
Yangyang Xu is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His current research interests include computer vision, image processing, computer graphics, and deep learning.

Xuemiao Xu received the B.S. and M.S. degrees in computer science and engineering from the South China University of Technology in 2002 and 2005, respectively, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong in 2009. She is currently a Professor with the School of Computer Science and Engineering, South China University of Technology. Her research interests include object detection, tracking, recognition, and image, video understanding and synthesis, particularly their applications in the intelligent transportation.

Jianbo Jiao (Member, IEEE) received the Ph.D. degree in computer science from the City University of Hong Kong in 2018. From 2017 to 2018, he was a Visiting Scholar with the Beckman Institute, University of Illinois at Urbana–Champaign. He is currently a Postdoctoral Researcher with the Department of Engineering Science, University of Oxford. His research interests include computer vision and machine learning. He was a recipient of the Hong Kong Ph.D. Fellowship.

Keke Li received the B.S. and M.S. degrees in computer science and engineering from the South China University of Technology in 2017 and 2020, respectively. Her research interests include face editing and synthesizing.

Cheng Xu received the B.Eng. degree in software engineering from Xiangtan University, China, in 2016. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, South China University of Technology. His research interests include computer vision, image processing, and deep learning.



Shengfeng He (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Macau University of Science and Technology in 2009 and 2011, respectively, and the Ph.D. degree from the City University of Hong Kong in 2015. He is currently an Associate Professor with the School of Computer Science and Engineering, South China University of Technology. His research interests include computer vision, image processing, computer graphics, and deep learning. He serves on the Editorial Board of *Neurocomputing*.