Pixel Sampling for Style Preserving Face Pose Editing

Xiangnan Yin, Liming Chen, Di Huang, Hongyu Yang, Zehua Fu, Yunhong Wang

Pixel Sampling for Style Preserving Face Pose Editing

Xiangnan Yin[1], Di Huang[2], Hongyu Yang[2], Zehua Fu[1], Yunhong Wang[2], Liming Chen[1]
[1]Department of Mathematics and Informatics, Ecole Centrale de Lyon, Lyon, 69134, France
[2]School of Computer Science and Engineering, Beihang University, Beijing, 100191,China
{yin.xiangnan,Liming.chen, zehua.fu}@ec-lyon.fr, {dhuang,hongyuyang,yhwang}@buaa.edu.cn

Abstract

The existing auto-encoder based face pose editing methods primarily focus on modeling the identity preserving ability during pose synthesis, but are less able to preserve the image style properly, which refers to the color, brightness, saturation, etc. In this paper, we take advantage of the well-known frontal/profile optical illusion and present a novel two-stage approach to solve the aforementioned dilemma, where the task of face pose manipulation is cast into face inpainting. By selectively sampling pixels from the input face and slightly adjust their relative locations with the proposed "Pixel Attention Sampling" module, the face editing result faithfully keeps the identity information as well as the image style unchanged. By leveraging high-dimensional embedding at the inpainting stage, finer details are generated. Further, with the 3D facial landmarks as guidance, our method is able to manipulate face pose in three degrees of freedom, i.e., yaw, pitch, and roll, resulting in more flexible face pose editing than merely controlling the yaw angle as usually achieved by the current state-of-the-art. Both the qualitative and quantitative evaluations validate the superiority of the proposed approach.

1. Introduction

Face pose editing aims to change the pose of an input face image while keeping its original identity unchanged. It has many potential applications, *e.g.*, face recognition, movie industry and entertainment. The current state-of-the-art has featured two main research lines in this field, *i.e.*, 3D reconstruction-based, and simple 2D based.

For 3D reconstruction-based approaches, face pose editing is achieved by either mapping the 2D face images to 3D face models with fixed or regressed parameters [25, 50, 44] or directly regressing the UV map [8, 11] of the input face. The advantage of such models is that pose control is not demanding. With the reconstructed 3D face, face images at any target pose can be obtained by 3D geometrical transformation and 2D projection. However, regressing either the parameters of predefined 3D models or the UV map requires large amounts of high-quality training data. Moreover, due to the restriction of the predefined model and the missing texture of extreme poses, fine details of the images are ignored. As a result, the faces generated by these approaches are generally not photo-realistic enough and require further refinements[13].

Thanks to the development of Generative Adversarial Networks (GAN) [14], a number of GAN based 2D approaches to face pose editing have been proposed in recent years. GAN has achieved great success in face image inpainting and facial attribute editing [5, 19, 33, 32, 6]. However, the existing methods are generally only capable of editing the subtle attributes or local regions of the image, whereas the global structure remains almost unchanged. Regarding face pose manipulation, when changing the view angle from side to front, not only the local texture but also the global shape of the face image dramatically changes. Despite these difficulties, there still exist significant efforts tackling this problem [39, 18, 17, 37, 34]. Most of the methods are implemented by an encoder-decoder structured network, with a bottleneck layer in the middle, where the faces are first encoded into a low dimensional feature vector, and then decoded into the image space conditioned by the pose information, *e.g.*, CR-GAN [37], DR-GAN [39]. However, there exists an intrinsic trade-off between the image style conserving capability and the identity preserving ability in the compact deep feature space, *i.e.*, it is hard to model the expertise of both the face identity and other image properties, such as lightning condition, saturation, background color, *etc.*

To highlight the aforementioned dilemma that commonly incurs in current 2D based methods, we remove the face classification branch of DR-GAN [39] (with the latent feature dimensionality of 320) and train the model only with the adversarial loss and the reconstruction loss. In this case, an adversarial auto-encoder (AE) is achieved, where the reconstruction loss aims to efficiently preserve the style of the

978-1-7281-9186-7/20/$31.00 ©2020 IEEE

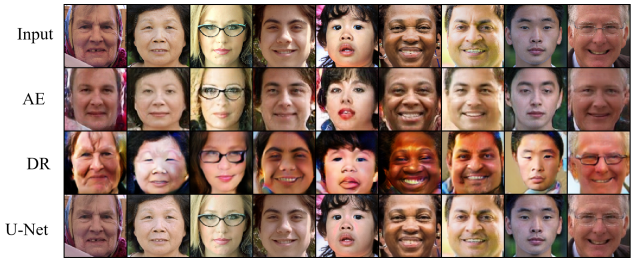


Figure 1. Illustration of the trade-off between identity preserving and style preserving.

input image, and the adversarial loss enforces the generated images photo-realistic. Figure 1 illustrates the input images (the first row) and the results obtained by the adversarial auto-encoder (second row) and DR-GAN (third row), respectively. As can be seen, the auto-encoder properly preserves the style of the input image, but it fails maintaining the identities. The reconstructed faces by DR-GAN successfully catch the identity characteristics of the input images, whereas the output ones are distorted and present obvious artifacts. If it is even painful for the model to faithfully rebuild the given input face in terms of both style and the identity without any pose manipulation, how can we further expect it to preserve them after changing the pose?

To fight the trade-off incurred by the low-dimensional restriction in the feature space, we seek solutions from the high-dimensional embeddings. But to make the condition label not ignored by the decoder, the encoding dimension should not be simply increased. The classical structure of U-Net [35], which adds skip-connections between symmetric layers of the encoder and the decoder, is able to prevent the problem of over-compression by concatenating features from the shallow layers in the reconstruction path. The last row of Figure 1 shows the corresponding reconstruction results, where both the identity and the style of the input face is well preserved. High-dimensional embedding is indeed promising in image synthesis, however, structures like U-Net convey too much low-level details, making it much more challenging to edit the face pose than on the low-dimension features, especially for the extreme shape changes. Therefore, how to enable face pose editing in the high dimensional feature space is the main problem to be solved.

To tackle the challenge above, we present a novel two-stage method and a module named "**P**ixel **A**ttention **S**ampling" (PAS) in this paper. Inspired by the fact that face images of different view angles also share a large number of similar pixels as highlighted by the optical illusion of face images [40, 41] in Figure 2(a), we believe that these pixels are significant to construct the texture of a face image in the target view through sampling. Specifically, given a target pose, this PAS module selects pixels from the input image and slightly change their relative locations in a learning manner to match the target pose (similarly to a non-linear image warping). Thus the recovered face editing result possesses the target pose and shares the original texture simultaneously, faithfully keeping the identity information and image style unchanged. Due to the lack of texture in invisible regions, the results of PAS would possibly contain noises and holes, then the main task can be cast as image inpainting, which has been extensively studied. We feed the intermediate pose-edited face image into the aforementioned U-Net, so that the noises can be filtered out and holes filled. By incorporating the module of PAS, the low-level details preserved by the U-Net are no longer burdensome for the task of pose editing, instead, they become useful information for generating the visually compelling face images.

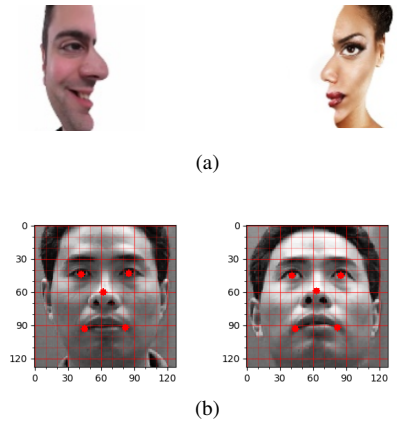


Figure 2. (a) Example of front/profile optical illusion. Indicating that face images in different view angles still share pixel-level similarities. (b) The ambiguity of representing 3D face pose by 2D landmarks. The two faces above have almost the same landmark distribution, but are in different poses.

Further, by introducing the 3D landmarks rather than 2D ones to represent the head pose more precisely, we achieve a better flexibility of pose manipulation. On the contrary, the traditional methods like DR-GAN and CR-GAN merely manipulate face images in several discrete yaw angles, and TP-GAN [18] can only frontalize face images. Although CAPG-GAN [17] uses 2D landmarks to guide the generation, it cannot generate faces in arbitrary poses as it claims, since using 2D landmarks to represent 3D angles can bring ambiguity as Figure 2(b) shows. Besides, the 3D landmarks tends to provide richer shape-related information, further facilitating the synthesis of face images.

In summary, our main contributions are as follows:

- A novel two-stage face pose editing method is proposed, which casts the task of face pose manipulation as face inpainting, thereby enabling it fully utilize the fine details of the given input image by exploiting high-dimensional embedding.

- A new "**P**ixel **A**ttention **S**ampling" module is designed, which effectively resolves the conflict between the identity and style preserving.

- The 3D facial landmarks is introduced to represent face poses for the first time, resulting in more flexible pose editing than using the discrete one-hot pose label or ambiguous 2D facial landmarks.

- The proposed method demonstrates competitive performance in comparison with the current state-of-the-art, both qualitatively and quantitatively.

2. Related Work

2.1. Generative Adversarial Network (GAN)

In recent years, Generative Adversarial Networks (GAN) has been one of the most popular research directions for image generation. Traditional GAN is composed of a generator and a discriminator. The training follows an adversarial paradigm. To overcome the problems of unstable gradient and mode collapse, Wasserstein GAN (WGAN) [1] proposes the earth move distance as metric in the discriminator's loss function. To enforce the Lipschitz constraint of the discriminator, SN-GAN [29] applies spectral normalization to the weight parameters. Due to its simplicity and promising effect, most of the recent GAN based algorithms make use of this technique, including SN-GAN [46], BigGANs [3], StyleGAN [22], *etc*. In our method, SN-GAN is also adopted in the structure.

2.2. Image-to-Image Translation

The combination of auto-encoder with discriminator has achieved impressive results in image-to-image translation [6, 51, 48, 33]. In multi-domain image translation tasks, the domain information is provided either to the bottleneck layer of the auto-encoder [48, 16, 39], or to the entry of the encoder/generator [6, 33, 17], by simply concatenating the domain label with the features or input images. Conditional batch normalization [7] and conditional instance normalization (CIN) [10] provide another way of introducing the conditional label in addition to concatenation, via predicting the affine parameters of the normalized feature map (either by batch normalization or by instance normalization) from the input label. Here, the CIN technique is exploited in our decoder to avoid the operation of duplicating the label.

In multi-domain image translation, the discriminator is used to not only estimate the image quality, but also control the target domain of the generated image. Our approach borrows the idea of projection discriminator [30], which introduces the reality score and the inner product of the embedded label with the features of the input data.

2.3. Face Pose Manipulation

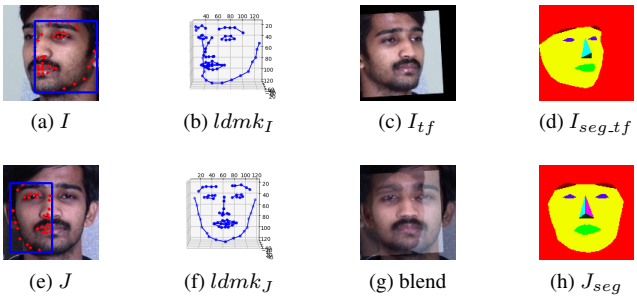
The existing methods can be roughly divided into two categories: 3D reconstruction based, and simple 2D based.

For the 3D based models, DA-GAN [50] uses a predefined 3D face model to produce the synthesized faces with arbitrary poses, and the dual agents serve to keep the identity information stable and improve the realism, Feng *et al.* [11] train a model to regress the UV map from a single 2D image directly, which records the 3D shape information. Tran *et al.* [38] proposes a framework to learn a nonlinear 3DMM model from a large set of unconstrained face images. FF-GAN [44] incorporates 3DMM [2] into the GAN based structure, where the 3DMM coefficients provide the low-frequency information, while the input image injects high-frequency local information.

For 2D based models, DR-GAN [39] learns a disentangled representation of face identity with the supervision of an auxiliary face classifier of the discriminator. TP-GAN [18] employs a two-pathway architecture to preserve both global and local texture information separately, and generates the frontalized face images. With the guidance of 2D facial landmarks, CAPG-GAN [17] is able to generate faces of arbitrary poses, where the couple-agent discriminator distinguishes the generated face/landmark pairs and profile/front pairs from ground-truth pairs, such design enables the algorithm generate face images of target poses while keeping the identity unchanged. CR-GAN [37] trains the generator to produce face images directly from the noises, together with the training of pose manipulation, maintaining the completeness of the learned embedding space. FNM [34] employs unsupervised training and synthesizes normalized face images of Multi-PIE [15] style. Most of the above methods only focus on modeling the identity preserving ability, whereas they generally ignore the image style preserving ability, such as color, facial expression, lightning, *etc.* Although it is claimed that the synthesized frontal face images improve the face verification accuracy, the generated face images are visually far from the input images, thus greatly limits their further usage scenarios other than face recognition.[47] frontalizes the face image by predicting the pixel displacement. However, it's hard to extend to the arbitrary face pose editing problem due to the time consuming SIFT feature extraction.

3. Method

The goal of our method is to keep not only the identity but also the image style during face pose manipulation. We first define several notations: (I, J) denotes paired face images in the training set, where I is the source image, and J is the target one. The 3D facial landmarks are denoted as $ldmk_I$ and $ldmk_J$, which could be detected by an off-the-shelf 2D\3D facial landmark detector [4]. I_{tf} represents



(a) I (b) $ldmk_I$ (c) I_{tf} (d) I_{seg_tf}

(e) J (f) $ldmk_J$ (g) blend (h) J_{seg}

Figure 3. (a) and (e) are the source image and the target one, respectively, where the corresponding landmarks and the bounding boxes of their bigger side are shown to illustrate our aligning strategy. (b) and (f) are the 3D facial landmarks detected by [4]. (c) shows the aligned image I_{tf}. (g) is the alpha blend of I_{tf} and J, illustrating that the target image shares pixel-level similarities with the source image. (d) and (h) are the segmentation maps of I_{tf} and J transformed from their 2D facial landmarks.

the input image after similarity transformation. To guide the training, the landmark based segmentation maps of I_{tf} and J are also required, which we denote as I_{seg_tf} and J_{seg}. These notations are visualized in Figure 3.

Our approach is composed of three major steps: preprocessing, pixel attention sampling, and image inpainting. They are described in detail subsequently.

3.1. Preprocessing

Given the fact that human faces are roughly left-right symmetrical, thus a face at an arbitrary pose always has at least one side fully exposed to the camera. This preprocessing step aims to align the fully exposed side of face I to that of a target face J.

The inputs of this step are the input face image I, its 3D facial landmarks $ldmk_I$, and the landmarks $ldmk_J$ of the image J at a target pose. We first find the fully exposed side by calculating the bounding box region of the projected facial landmarks, as illustrated in Figure 3(a) and Figure 3(e). Then, the least square regression on the corresponding landmarks is applied to calculate the transformation matrix, based on which the aligned image I_{tf} could be obtained, as shown in Figure 3(c). From Figure 3(g), we observe that I_{tf} and J indeed share pixel-level similarities. Finally, with the 2D facial landmarks of I and the transformation matrix obtained above, we obtain the 2D facial landmarks of I_{tf}, as well as the landmark based segmentation map I_{seg_tf}. Besides, to guide the training process of the PAS module, the segmentation map of the target image J_{seg} is also prepared at this stage.

3.2. Pixel Attention Sampling

The previous preprocessing step delivers the input face image with the larger side aligned to the target pose. Despite the fact that the transformed input face image I_{tf} and the face at the target pose J share many similarities in terms of texture, there still exist great gaps between them, from the global shape to the finer details of textures. Therefore, our goal at this stage is to preserve and fine-tune their similar face regions while eliminating the major differences. This is achieved by a novel pixel sampling based module, which we call **P**ixel **A**ttention **S**ampling module (PAS), since the process of sampling mainly "focuses" on bridging the gaps. Figure 4 depicts the corresponding diagram.

Specifically, given the transformed image I_{tf} and the target pose $ldmk_J$, PAS generates a two-channel coordinate sampling map of the same size as I_{tf}. The first channel holds the abscissa while the second one for the ordinate. Each pixel location of the map is registered a coordinate, indicating which input pixel of I_{tf} that location will sample from. Note, the original pixel indices are converted into decimal coordinates ranging from -1 to 1, for the purpose of gradient backpropagation, and the final sampling is achieved by interpolating the adjacent pixels. Our sampling map is similar to the one used in the spatial transformer network[20]. The difference lies in that the one in [20] is determined by a 2D affine transform matrix, with only six parameters, whereas our sampling map is directly predicted by the neural network, resulting in $height \times width \times 2$ parameters in total. The PAS module is composed of two parts *i.e.*, the image embedder and the sampler. The embedder consists of stacked convolution layers, conditional instance normalization [10] layers (CIN), and self-attention [46] layers (SA). The CIN layers incorporates the 3D facial landmarks of the target pose to guide the embedding, and the SA layers enable the embedder focus more on the global structure of the face.The embedder finally outputs a 512-dimensional feature vector, which is further fed into the sampler to generate the sampling maps. The sampler is composed of fully connected layers and ReLU layers. After applying the obtained sampling map to the transformed input face image I_{tf} and its corresponding segmentation map I_{seg_tf}, we could obtain the intermediate face image at the target pose, denoted as $\hat{J}_{fake}$, and its corresponding fake segmentation map, denoted as J_{seg_fake}. In order to maintain the reconstruction ability of the module, the original image I and its corresponding 3D landmarks $ldmk_I$ are also fed into the PAS module and the reconstructed output $\hat{I}_{recon}$ is achieved.

The training process of the PAS is guided by the following losses:

Pixel-wise loss between $\hat{J}_{fake}$ and J, $\hat{I}_{recon}$ and I, which is commonly used in the image-to-image translation algorithms. It can be formulated as:

$$L_{pix} = L_1(\hat{J}_{fake}, J) + 0.1 \cdot L_1(\hat{I}_{recon}, I) \qquad (1)$$

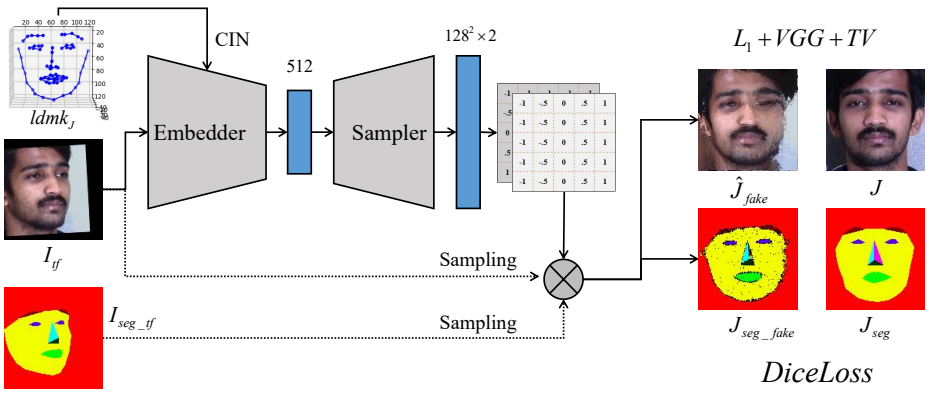


Figure 4. Structure of the proposed **P**ixel **A**ttention **S**ampling (PAS) module.

where

$$L_1(I,J) = \frac{1}{WHC}\sum_{x,y,c=1}^{W,H,C} |I(x,y,c) - J(x,y,c)| \quad (2)$$

Since it does not take much effort to learn an identity mapping, we set the weight of the reconstruction loss to 0.1, which makes the PAS module concentrate much more on the pose manipulation task.

Segmentation loss between J_{seg_fake} and J_{seg}. Based on the assumption that if the sampled image $\hat{J}_{fake}$ is close to the target image J, the segmentation map J_{seg_fake} should be close to the target segmentation map J_{seg} as well. We therefore introduce a segmentation-related loss so as to push $\hat{J}_{fake}$ close to J. To facilitate the training converge, the segmentation loss is used as a complement to the aforementioned pixel-wise loss L_{pix}. Here, we make use of the Dice loss[28], which has been widely exploited in image segmentation tasks, and it can be formulated as:

$$L_{seg} = \sum_{c=1}^{N} 1 - \frac{2\sum_{x,y} J^c_{seg}(x,y) \cdot J^c_{seg\_fake}(x,y)}{\sum_{x,y} J^c_{seg}(x,y) + \sum_{x,y} J^c_{seg\_fake}(x,y)} \quad (3)$$

where c represents the different classes of facial attributes. Since each pixel location (x,y) of the segmentation map is represented by a c-dimensional one-hot vector, the fraction in Equation 3 is thus a simple intersection over union. The benefit of the adopted loss function is that it is independent to the amount of pixels of different classes.

Perceptual loss[21] between $\hat{J}_{fake}$ and J. Perceptual loss is significant to preserve the identity information and high-level semantic features of the face images. We follow the work of [45] and employ the pre-trained VGG-Face [31] network to extract the features:

$$L_{per} = VGG_{loss}(\hat{J}_{fake}, J) \quad (4)$$

with

$$VGG_{loss}(I,J) = \sum_i |VGG_{Face}(I)_i - VGG_{Face}(J)_i| \quad (5)$$

where i is the layer index of the pre-trained model and $i \in \{3,8,15,22,29\}$, which are the last convolutional layer of each feature map scale.

Total variation loss. Total variation [27] loss has been widely used in GAN based algorithms for its powerful ability of reducing the noises and smoothing the generated results. In our PAS module, there inevitably exist obvious noises, since the resultant face image is pixel-wise sampled from the transformed input face image. Therefore, the TV loss is incorporated:

$$L_{tv} = TV(\hat{J}_{fake}) + TV(\hat{I}_{recon}) \quad (6)$$

where

$$\begin{aligned} TV(I) = &\sum_{x,y,c=1}^{W-1,H,C} |I(x+1,y,c) - I(x,y,c)|^2 \\ &+ \sum_{x,y,c=1}^{W,H-1,C} |I(x,y+1,c) - I(x,y,c)|^2 \end{aligned} \quad (7)$$

The overall training loss of the PAS module is a sum of the above losses:

$$L_{sampler} = L_{pix} + L_{seg} + L_{per} + L_{tv} \quad (8)$$

Thanks to the PAS module, we achieve a face image whose facial attributes have been aligned to the target pose location, with the original identity and style characteristics well preserved. It should be noted that, as the sampling is accomplished by interpolating adjacent pixels, it only modifies the location of the pixels within a small area around them, the PAS module is thus not able to sample for instance the left eye from the right one or the opposite. As

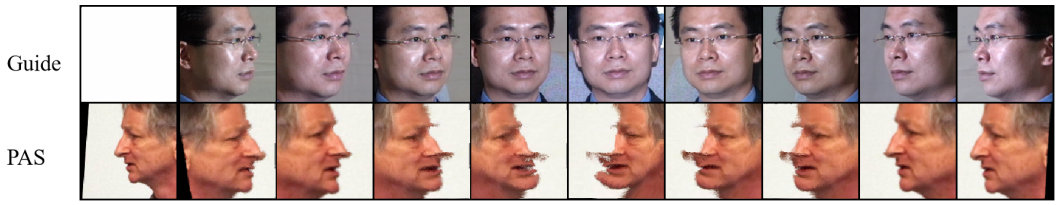


Figure 5. The result of PAS. The first row shows the guiding face images at target poses, the first image in the second row is the input face image, and the remaining images are synthesized faces based on the landmarks of the guiding face images. We can see that the pixels are sampled and adjusted to the target pose. The noises and holes will be removed or filled at the next image inpainting stage.

a result, the sampled face images possibly contain artifacts, holes and noises, as illustrated in Figure 5. In order to further improve the generated image quality, image inpainting is introduced subsequently.

3.3. Image Inpainting

The image inpainting stage is to restore the holes and remove the noises and artifacts on the intermediate faces generated by PAS, and finally generate photo-realistic face images. To accomplish this goal, we introduce a Conditional Adversarial Auto-Encoder, where the discriminator is implemented by a projection discriminator [30], and the auto-encoder is based on the U-Net structure [35]. We also make use of CIN layer to merge the information provided by 3D facial landmarks, the identity features, and the image features, thereby making the generated face image in desired pose and shape. More precisely, the inputs of the encoder are the images generated by PAS together with their target poses, *i.e.*, $\hat{J}_{fake}$ with $ldmk_J$ for the task of pose manipulation, and $\hat{I}_{recon}$ with $ldmk_I$ for the task of reconstruction. To well preserve the face identity, the decoder is conditioned by the high level feature extracted by the pre-trained LightCNN [42] model, where the parameters of the fully connected layer is fine-tuned during training. The outputs of the auto-encoder are denoted as J_{fake} and I_{recon}, whose ground truths are J and I, respectively. To further improve the model's generalization ability, unpaired face images could also be exploited to supplement the training set. This is achieved by feeding the network with the partially occluded face images S_{occ} and their 3D landmarks, and expect the network to output S_{recon} restoring the original S. For the discriminator, we feed all the generated images, including J_{fake}, I_{recon} and S_{recon}, as fake samples, while their corresponding ground truth as the genuine ones, with dis_{real} and dis_{fake} as output, respectively.

The loss function of the inpainting network is composed of four parts:

Pixel-wise loss, formulated as:

$$L_{pix} = L_1(J_{fake}, J) + \lambda \cdot L_1(I_{recon}, I) + L_1(S_{recon}, S) \tag{9}$$

where L_1 is defined in equation 2.

Perceptual loss to capture the semantic similarity:

$$L_{per} = VGG_{loss}(J_{fake}, J) + VGG_{loss}(S_{recon}, S) \tag{10}$$

where VGG_{loss} is defined in equation 5. We do not include (I_{recon}, I) here, because compared to image reconstruction task, image inpainting and pose manipulation are more likely to lose the identity consistency.

Identity loss to maintain the identity-related characteristics stable. We use the pre-trained LightCNN [42] to extract the identity feature of the synthesized image and the target image, and minimize the L_1 loss of them:

$$L_{id} = \frac{1}{N}\sum_{i=1}^{N} |F(J_{fake})_i - F(J)_i| \tag{11}$$

Adversarial loss to guarantee the generated image quality:

$$L_{adv} = -dis_{fake} \tag{12}$$

Besides, we also incorporate the total variation loss to reduce the spike artifacts. The overall training loss of the generator of the image inpainting network is a sum of the aforementioned losses:

$$L_{gen} = L_{pix} + L_{per} + L_{id} + L_{tv} + L_{adv} \tag{13}$$

Following the work of [24], the **discriminator loss** is defined as:

$$L_{dis} = \mathbf{max}(1 - dis_{real}, 0) + \mathbf{max}(1 + dis_{fake}, 0) \tag{14}$$

4. Experiments

Given an input face image, the proposed method aims to manipulate its pose while keeping the identity unchanged along with its style. Correspondingly, we evaluate it in two aspects: the style-conserving skill and the identity-preserving ability during face pose editing. In this section, we present the training details first, then the qualitative analysis for face style conserving, followed by the quantitative results for identity preserving. Ablation studies are also carried out to highlight the effectiveness of the proposed PAS module.

4.1. Training details

The training is based on four databases: Multi-PIE [15], 300W-LP [52], CAS-PEAL-R1 [12], and CelebA [26]. **Multi-PIE** has four sessions with face images under 13 poses and 20 illuminations. We follow Setting 1 of TP-GAN [18] and train the proposed algorithm on the first 150 subjects of session 1, then test on the remained 99 subjects. **300W-LP** contains large-pose face images synthesized from 300W [36]. After manually filtering out the low-quality images, we have 40,159 images from 2,815 subjects in total. **CAS-PEAL-R1** contains 1,040 subjects. For each subject, gray-scale images across 21 different poses are included. **CelebA** is a large-scale face attributes dataset with more than 200K celebrity images in it.

During the training process, we use the occluded face images as input, and train the U-Net based generator to restore the original face images. This operation improves the generalization ability of the network, and make the generated images photo-realistic. All of the training images are cropped to 128×128 pixels. The learning rate is set to $1\mathbf{e}^{-4}$, and the Adam [23] optimizer is utilized with betas of [0.9, 0.999]. We first pre-train the generator and the discriminator on CelebA for 20000 iterations, making it a fundamental image inpainting model, which facilitates the subsequent training procedure. Then, we train the proposed PAS model and the image inpainting model jointly for 110000 iterations in total. Observing that the CAS-PEAL-R1 dataset consists of gray-scale images, which degenerates the color saturation of the generated images, we thus exclud the data of CAS-PEAL-R1 for the last 10000 iterations.

4.2. Style-conserving validation

Multi-PIE images under different poses are used as the guiding images, the pose of faces from CelebA are edited accordingly. As shown in Figure 6, the synthesized face images comply with the guiding faces in term of pose. They are visually photo-realistic and both the identities and the styles are well preserved, clearly validating the effectiveness of the proposed method. A further qualitative comparison of our method and CR-GAN [37], DR-GAN [39] and FNM [34] are demonstrated in Figure 7(a). As can be seen, our results are more visually convincing and the styles are closer to the input images compared to DR-GAN, and the identities are better preserved than CR-GAN. As for FNM, the generated image style is more similar to the training set, where the lighting and expressions are normalized, and the color has been changed, by contrast, our results better preserve those characteristics of the input face images. Moreover, the proposed approach is able to manipulate face poses in three degrees of freedom, resulting in more flexible pose editing results than merely controlling the yaw angles as usually achieved by previous methods. Figure 7(b) shows the results of editing both pitch and yaw angles of input face images (leftmost).

Table 1. FID score of frontalized face images (lower is better)

CR-GAN	DR-GAN	FNM	ours
204	122	150	105

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

Method	$\pm 90^\circ$	$\pm 75^\circ$	$\pm 60^\circ$	$\pm 45^\circ$	$\pm 30^\circ$	$\pm 15^\circ$
HPN [9]	29.82	47.57	61.24	72.77	78.26	84.23
c-CNN [43]	47.26	60.7	74.4	89	94.1	97.0
TP-GAN [18]	64.0	84.1	92.9	98.6	99.9	99.8
PIM [49]	75.0	91.2	97.7	98.3	99.4	99.9
CAPG-GAN [17]	77.1	87.4	93.7	98.3	99.4	99.9
FNM [34]	55.8	81.3	93.7	98.2	99.5	99.9
Light CNN [42]	2.6	10.5	32.7	71.2	95.1	99.8
Ours	45.5	78.7	90.0	99.6	99.9	100

Quantitative evaluations are further performed. We calculate the FID score of the above models, on the frontalized large pose face images from CelebA, the results are shown in Table 1, indicating that the proposed method generates face images with styles closer to the input face images, which can be applied to more perceptual applications.

4.3. Identity-preserving ability evaluation

There are 249 subjects in Session 1 of Multi-PIE. Following the Setting 1 of TP-GAN, we use the first 150 subjects for training, and the remaining 99 subjects for testing. The identity preserving ability is evaluated by Rank-1 recognition rate. The face with frontal view and normal illumination in the testing set compose the gallery, and the rest non-frontal images are used as probe.

The evaluation is conducted based on the features extracted by the pre-trained Light-CNN model. We directly extract the features of the probe images as baseline. For the proposed method, we first frontalize the probe face images, based on which their face representations are extracted. As can be seen from Table 2, the proposed method achieves similar or even better Rank-1 recognition rate in comparison with the baseline and state-of-the-art algorithms when the rotation angle is smaller than 60°. For larger rotation angles ($\geq 60^\circ$), the proposed algorithm drastically outperforms the baseline, whereas it does not perform as well as the SOTA algorithms. There exist two possible reasons: 1) The face images of extreme poses share relatively less pixels with the face images of front view, thus the pixels sampled by the PAS module are not sufficient enough for the following inpainting stage, and 2) most of the SOTA algorithms normalize the face images into a consistent style, where the information irrelevant to identity is filtered out, in contrast, our method preserves relatively more style information.

4.4. Ablation Study

To highlight the effectiveness of the PAS module, the ablation study is conducted by removing it and training the

Figure 6. The final result of our approach. The first row shows the guidance images. The input images are in the first column, and the simple reconstructed images are in the second column. The rest images are the pose editing results based on the landmarks of the exemplars.

(a) (b) (c)

Figure 7. (a) From top to bottom shows the input images, results of CR-GAN, DR-GAN, FNM and our method. (b) From left to right are the input images and the generated images with both yaw angles and pitch angles changed. (c) From left to right are the input images, half-face aligned images, frontalized images w/o PAS module, results of the proposed algorithm, and the ground-truth images.

U-Net based conditional adversarial auto-encoder directly. For the sake of fair comparison, we apply the same preprocessing pipeline (*i.e.*, align the larger side of the input image to match the target pose) and train the model with the same number of iterations. Figure 7(c) shows the results. As can be seen, the synthesized images without PAS are blurred. More specifically, in the second row and the fourth row, the mouths are not well aligned, and the unexpected edges of the aligned input images are not well removed. The results indicate that it is indeed difficult for the single U-Net based model to change the original patterns of the input image thus results in undesired artifacts.

5. Conclusion

In this work, we first carefully analyze the trade-off between the style-preserving ability and the identity-preserving ability of the existing 2D based pose manipulation methods. Based on the observation that face images in different poses share a large number of pixels, we propose a novel pose editing method and a sophisticatedly designed PAS module. The method selectively samples pixels from the input face and adjust their relative locations with the PAS module, so that the recovered face editing result match the target pose and faithfully keeps the original identity and style information unchanged. In this way, we convert the pose manipulation problem to a image inpainting problem, and further make the best of the finer details in the original face images to obtain convincing pose editing results. We also utilize 3D facial landmarks to represent the face pose, which is more precise and flexible comparing to the one-hot labels and the 2D facial landmarks adopted in previous studies. Extensive experiments validate that the proposed pose editing approach preserves the style information of the input images better than the existing methods.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] V. Blanz, T. Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194, 1999.

[3] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[5] J. Cai, H. Hu, S. Shan, and X. Chen. Fcsr-gan: End-to-end learning for joint face completion and super-resolution. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[6] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[7] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville. Modulating early visual processing by language. In *Advances in Neural Information Processing Systems*, pages 6594–6604, 2017.

[8] J. Deng, S. Cheng, N. Xue, Y. Zhou, and S. Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7093–7102, 2018.

[9] C. Ding and D. Tao. Pose-invariant face recognition with homography-based normalization. *Pattern Recognition*, 66:144–152, 2017.

[10] V. Dumoulin, J. Shlens, and M. Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016.

[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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 534–551, 2018.

[12] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao. The cas-peal large-scale chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 38(1):149–161, 2007.

[13] B. Gecer, B. Bhattarai, J. Kittler, and T.-K. Kim. Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 217–234, 2018.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[15] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[16] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 2019.

[17] Y. Hu, X. Wu, B. Yu, R. He, and Z. Sun. Pose-guided photorealistic face rotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8398–8406, 2018.

[18] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 2439–2448, 2017.

[19] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[20] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[21] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[22] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] J. H. Lim and J. C. Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017.

[25] F. Liu, Q. Zhao, D. Zeng, et al. Joint face alignment and 3d face reconstruction with application to face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[26] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[27] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015.

[28] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[29] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[30] T. Miyato and M. Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018.

[31] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[32] T. Portenier, Q. Hu, A. Szabo, S. A. Bigdeli, P. Favaro, and M. Zwicker. Faceshop: Deep sketch-based face image editing. *ACM Transactions on Graphics (TOG)*, 37(4):99, 2018.

[33] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833, 2018.

[34] Y. Qian, W. Deng, and J. Hu. Unsupervised face normalization with extreme pose and expression in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9851–9858, 2019.

[35] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[36] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[37] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas. Cr-gan: learning complete representations for multi-view generation. *arXiv preprint arXiv:1806.11191*, 2018.

[38] L. Tran and X. Liu. On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[39] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1415–1424, 2017.

[40] Combine side and front portrait shots to create optical illusion — photoshop cc. `https://www.youtube.com/watch?v=KxdXaAxdKBQ`. Accessed: 2020-2-25.

[41] Photoshop tutorial: How to make a bi-directional, optical illusion, photo portrait. `https://www.youtube.com/watch?v=H8PpBInBEDM&t=29s`. Accessed: 2020-2-25.

[42] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[43] C. Xiong, X. Zhao, D. Tang, K. Jayashree, S. Yan, and T.-K. Kim. Conditional convolutional neural network for modality-aware face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3667–3675, 2015.

[44] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker. Towards large-pose face frontalization in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3990–3999, 2017.

[45] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. *arXiv preprint arXiv:1905.08233*, 2019.

[46] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[47] 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 Transactions on Image Processing*, 28(5):2187–2199, 2018.

[48] Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5810–5818, 2017.

[49] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, et al. Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2207–2216, 2018.

[50] J. Zhao, L. Xiong, P. K. Jayashree, J. Li, F. Zhao, Z. Wang, P. S. Pranata, P. S. Shen, S. Yan, and J. Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *Advances in Neural Information Processing Systems*, pages 66–76, 2017.

[51] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[52] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.