FaceShapeGene: A Disentangled Shape Representation for Flexible Face Image Editing

Sen-Zhe Xu†‡*, Hao-Zhi Huang†, Shi-Min Hu†, Wei Liu‡
†Tsinghua University ‡Tencent AI Lab
xsz15@mails.tsinghua.edu.cn, huanghz08@gmail.com, shimin@tsinghua.edu.cn, wl2223@columbia.edu

Abstract

Existing methods for face image manipulation generally focus on editing the expression, changing some predefined attributes, or applying different filters. However, users lack the flexibility of controlling the shapes of different semantic facial parts in the generated face. In this paper, we propose an approach to compute a disentangled shape representation for a face image, namely the FaceShapeGene. The proposed FaceShapeGene encodes the shape information of each semantic facial part separately into a 1D latent vector. On the basis of the FaceShapeGene, a novel part-wise face image editing system is developed, which contains a shape-remix network and a conditional label-to-face transformer. The shape-remix network can freely recombine the part-wise latent vectors from different individuals, producing a remixed face shape in the form of a label map, which contains the facial characteristics of multiple subjects. The conditional label-to-face transformer, which is trained in an unsupervised cyclic manner, performs part-wise face editing while preserving the original identity of the subject. Experimental results on several tasks demonstrate that the proposed FaceShapeGene representation correctly disentangles the shape features of different semantic parts. Comparisons to existing methods demonstrate the superiority of the proposed method on accomplishing novel face editing tasks.

1. Introduction

With the rapid development of deep generative models, image generation results evolve to be more and more realistic. Face is one of the most significant image categories that people care about, due to its wide applications in various aspects. The leading technologies for image generation include Generative Adversarial Networks (GANs) [11, 4, 2, 12, 20, 27], Variational Auto-Encoders (VAEs) [23, 34, 24, 15, 32, 3], and Flow-based Generative Models (FGMs) [9, 10, 22]. Among them, GANs receive most of the attentions and implicate a lot of variants that enable a variety of face generation applications.

Besides randomly synthesizing a realistic face image based on a latent vector [11, 4, 2, 12, 20], people are more interested in developing approaches to support more creative and flexible face editing tools. Transforming a label map or a sketch into a corresponding face image can be formulated as an image-to-image translation problem [18, 41, 39]. By editing the label map or the sketch manually, we can generate face images with different overall layouts. However, the identity of the generated person is out of control in this scenario. By training with an auxiliary classification loss, the image-to-image translation technique can also be applied to manipulating one or multiple prede-
In summary, the main contributions of this paper are four-fold. Firstly, we present a method for extracting a disentangled shape representation of a face image. Secondly, we propose a shape-remix network to recombine the shape features of different individuals to easily generate a desired label map. Thirdly, we propose a conditional label-to-face transformer to generate a face image according to the given label map while preserving the original identity. Fourthly, some novel face editing manipulations are developed by coupling the shape-remix network and the conditional label-to-face transformer together.

2. Related Work

General Image-to-Image Translation. Image-to-image translation is a problem of translating one possible representation of an image into another representation. Isola et al. [18] proposed Pix2Pix to give a supervised solution to general image-to-image translation based on conditional adversarial networks. Afterwards, some unsupervised methods [41, 29] were proposed by introducing the cycle-consistent constraints. The above methods make a simplifying assumption that image-to-image translation is a problem of learning a deterministic one-to-one mapping. However, one-to-many or many-to-many mappings exist in most of the image-to-image translation tasks. Recently, methods like MUNIT [17] and DRIT [26] tried to tackle the multimodal image-to-image translation problem by decomposing the latent representation of an image into a domain-invariant content code and a domain-specific style code, greatly reducing mode collapse and producing diverse multimodal translation results. There are also some approaches that intend to perform image-to-image translation at higher resolution, under either a supervised setting [39] or an unsupervised setting [27].

Face Editing with GANs. Besides studying the general image-to-image translation problem, a large number of GANs focus on face editing. GANimation, proposed by Pumarola et al. [33], delves deeper into the expression editing task and utilizes the Facial Action Coding System (FACS) for describing a continuous embedding of muscle movements. By training a GAN model conditioned on the FACS codes, GANimation can manipulate the expression of the face images. Choi et al. proposed StarGAN [8] to integrate multiple two-domain facial attribute transformations in a single model. Zhao et al. proposed ModularGAN [40] to accomplish editing multiple facial attributes at the same time by stacking multiple two-domain models in series. Besides transforming some predefined facial attributes, there are also some methods working on generating faces from arbitrary viewpoints, while preserving the identity. Shen et al. [36] proposed a three-player GAN, which considers pose, identity, and realism simultaneously. In this paper, we propose a novel disentangled shape representation for face images, which is the basis of a novel part-aware face edit-
Figure 2. The overall pipeline of our system. A disentangling encoder is trained to extract the proposed FaceShapeGene representation, which divides the shape information of the whole face into seven semantic parts. Based on this representation, a shape-remix network is proposed to produce a remixed face label map by recombining the FaceShapeGenes of multiple faces. In addition, a conditional label-to-face transformer is proposed to transform the remixed label map into a corresponding photo-realistic face image while preserving the identity of the conditional input.

Our approach can also preserve the identity of the input face by utilizing cycle-consistent losses. While previous approaches [33, 8, 40] merely employ a cycle-consistent loss in the image domain, we also introduce a cycle-consistent loss in the label domain to ensure that the shape-remix network works as expected.

3. Approach

We propose FaceShapeGene, a disentangled shape representation which benefits the development of flexible face editing tools. The overall pipeline is shown in Fig. 2. The proposed FaceShapeGene encodes the shape information of different facial parts into 1D latent vectors, respectively. On the basis of FaceShapeGene, a part-wise face editing system is developed. Our proposed system consists of two components. The first component is a shape-remix network, which recombines the FaceShapeGenes from multiple faces and produces a remixed face shape in the form of a label map. Exploiting the proposed shape-remix network, we can conveniently transfer the shape of a facial part from a reference image to a target image. The second component is a conditional label-to-face transformer, which takes the remixed label map and a conditional face image as input to generate a remixed face image. While the remixed label map provides the desired shape information, the conditional face image offers the identity information we want to keep. The shape-remix network and the conditional label-to-face transformer are coupled together and trained in an unsupervised cyclic manner to learn to generate a realistic face image in the desired shape with the target identity.

3.1. Learning The FaceShapeGene

We propose a disentangled shape representation FaceShapeGene for face images, with which we can modify the shape of each facial part individually by editing the corresponding part-wise feature. In this paper, we focus on seven facial parts, including hair, eyebrows, eyes, nose, mouth, face shape, and body. As shown in Fig. 3, the shape information of different parts are disentangled by training an individual local face parser \((G_E^i, G_D^i)\) for each facial part. For the \(i\)-th facial part, an encoder \(G_E^i\) encodes \(x\) into a 1D latent vector, while a decoder \(G_D^i\) decodes the 1D latent vector into a partial label map \(y^i\). Our label map is a three-channel image in the RGB color space, in which each facial part is represented by a unique color, except for the mouth and body. For the mouth, in order to increase the representative capacity, we use three different colors to denote the upper lip, lower lip, and teeth. For the body, we use two different colors for the body skin and clothes. Since the supervision signal is a label map, which contains no texture and color of the original face, the local face parser will...
automatically discard the detailed appearance information during training.

We formulate the local face parsing task as a regression problem rather than a dense classification problem. This is because when formulating the label map prediction task as a dense classification problem using the cross-entropy loss as in the scene parsing task [5, 6, 7], the minority category with fewer pixels, like the eyes, tends to be neglected. On the contrary, formulating the task as a regression problem can alleviate the above class imbalance issue. To train the local face parsers, a combination of an $L_1$ loss, a VGG loss and a GAN loss is adopted. The $L_1$ loss is a pixel-wise reconstruction loss defined as:

$$
\mathcal{L}_{L_1}(G_E^i, G_D^i) = \mathbb{E}_{x,y} \left[ \|G_D^i(G_E^i(x)) - y \|_1 \right], \tag{1}
$$

where $x$ denotes the input image, $y^i$ denotes the ground-truth partial label map corresponding to the $i$-th part. However, in our experiments, using only an $L_1$ loss leads to blurry generated label maps. To alleviate this issue, a VGG loss and a GAN loss are introduced. The VGG loss is a perceptual reconstruction loss [19, 25] defined as:

$$
\mathcal{L}_{VGG}(G_E^i, G_D^i) = \mathbb{E}_{x,y} \left[ \|\phi(G_D^i(G_E^i(x))) - \phi(y)\|_1 \right], \tag{2}
$$

where $\phi$ denotes the features extracted by a pretrained VGG19 network [37]. In our experiments, we use the conv1-1, conv2-2, conv3-2, conv4-4, and conv5-4 layers to extract features. The VGG loss encourages the generated result to be similar to the ground-truth in the semantic feature domain. We also adopt a discriminator $D_i$ for the $i$-th facial part to calculate the GAN loss:

$$
\mathcal{L}_{GAN}(G_E^i, G_D^i, D^i) = \mathbb{E}_{x,y} \left[ \|D_i(x, y^i)\|_2^2 \right] + \mathbb{E}_x \left[ 1 - D_i(x, G_D^i(G_E^i(x)))\right]. \tag{3}
$$

Here, we use LSGAN [31] and PatchGAN [18] for stable training. $\mathcal{L}_{GAN}$ ensures that the generated partial label map $\hat{y}^i$ stays in the label domain, which brings more perceptual details. The full objective function for training the local face parser $(G_E^i, G_D^i)$ is:

$$
\mathcal{L}_{total}(G_E^i, G_D^i) = \mathcal{L}_{L_1} + \lambda_{VGG}\mathcal{L}_{VGG} + \lambda_{GAN}\mathcal{L}_{GAN}.
$$

We optimize the objective function by alternately updating the local face parser $(G_E^i, G_D^i)$ and the discriminator $D^i$: $\arg\min_{G_E^i, G_D^i} \max_{D^i} \mathcal{L}_{total}(G_E^i, G_D^i, D^i)$.

Once the training of all the local face parsers is done, we train an overall decoder $G_D$ to gather the shape information of the whole face, as shown at the bottom of Fig. 3. We name the collection of all the part-wise encoders as a disentangling encoder $G_E$, which concatenates all the part-wise 1D latent vectors to formulate an overall 1D latent vector $f$. Then, $f$ is fed to the overall decoder $G_D$ to produce the final whole-face label map $y$. Similarly, we use a combination of $L_1$, VGG, and GAN losses to train $G_D$. $G_E$ is fixed during training $G_D$.

![Figure 4. The shape-remix network.](image)

3.2. The Shape-Remix Network

Based on the disentangled shape representation, we propose a shape-remix network $S$ to recombine the part-wise shape features from different people, generating a new whole-face label map containing the shape characteristics of multiple faces. This operation resembles the genetic editing process. As shown in Fig. 4, an input image $x_A$ is treated as a receptor and a reference image $x_B$ acts as a donor. For both the receptor $x_A$ and the donor $x_B$, we employ the disentangling encoder $G_E$ to extract two FaceShapeGenes $f_A$ and $f_B$. By replacing the hair gene of $f_A$ with that of $f_B$, a remixed FaceShapeGene $f_r$ is obtained. Then, $f_r$ is decoded into a remixed label map $y_r$ by the overall decoder $G_D$. The same operation can be applied to other semantic parts.

3.3. The Conditional Label-to-Face Network

On the basis of the shape-remix network $S$, we propose a conditional label-to-face transformer $F$ to complete the task of the part-wise shape editing of a face image. The transformer $F$ learns to transform a whole-face label map to a photo-realistic face image, while preserving the identity of a conditional input image. Since there is no ground-truth data for the supervised training of $F$, we adopt both adversarial losses and cycle-consistent losses to train $F$ in an unsuper-
Figure 5. The cyclic training framework. The shape-remix network \( S \) takes a pair of images \((x_A, x_B)\), producing a remixed whole-face label map \( y_r \), by transferring the \( i \)-th part-wise shape feature of \( x_B \) to \( x_A \). The conditional label-to-face transformer \( F \) takes \( y_r \) and \( x_A \) to generate a remixed face \( x_r \). Symmetrically, the remixed face image \( x_r \) and the receptor image \( x_A \) are fed back to \( S \) to reconstruct a remixed label map \( y'_r \). \( F \) takes \( y'_r \) and the remixed face \( x_r \) as input, generating a reconstructed image \( x'_A \).

A more detailed description follows:

Receptive manner. The complete training process is illustrated in Fig. 5. At the beginning, the shape-remix network \( S \) takes a pair of images \((x_A, x_B)\) and produces a remixed whole-face label map \( y_r = S(x_A, x_B, i) \). Here, \( i \) indicates that the index of the targeted facial part. \( i \) is randomly chosen for each iteration. Then, the transformer \( F \) takes \( y_r \) and \( x_A \) as input, generating a remixed face \( x_r = F(y_r, x_A) \). Here, \( x_A \) is a conditional input to provide the identity information. In our experiments, we remove the background of the conditional input for more stable training. Note that from the inconsistency between \( y_r \) and \( x_A \), the transformer \( F \) can implicitly identify the targeted part. The remixed face \( x_r \) is supposed to be corresponding to the remixed label map \( y_r \), while preserving the identity of the conditional input \( x_A \). Currently, the color or texture of the targeted facial part is generated randomly, because we only constrain the shape. The strategy for assigning the desired texture will be left for the future work. Symmetrically, the remixed face image \( x_r \) and the receptor image \( x_A \) are fed back to the shape-remix network to reconstruct a remixed label map \( y'_r = S(x_r, x_A, i) \). Note that \( y'_r \) is supposed to be the same as \( y_A \), which is the ground-truth label map for \( x_A \). Next, \( F \) takes the reconstructed label map \( y'_r \) and the remixed face \( x_r \) as input, generating a reconstructed image \( x'_A \). Here, \( x'_A \) is supposed to be the same as the input \( x_A \).

**Cycle-Consistent Losses.** According to the above process, two cycle-consistent losses can be defined. The first one is in the label domain:

\[
\mathcal{L}_{\text{cyc.L}} = \mathbb{E}_{y'_r, y_A} \left[ \| y'_r - y_A \|_1 + \lambda_{\text{VGG}} \| \phi(y'_r) - \phi(y_A) \|_1 \right].
\]  

(4)

On one hand, \( \mathcal{L}_{\text{cyc.L}} \) provides a supervision signal to fine-tune the overall decoder \( G_D \) in the shape-remix network \( S \), when regarding the remixed face \( x_r \) as a receptor and the original receptor face \( x_A \) as a donor, respectively. On the other hand, since the shape features except for the \( i \)-th part are provided by \( x_r \), \( \mathcal{L}_{\text{cyc.L}} \) encourages the remixed face \( x_r \) to preserve the correct part-wise shapes to reconstruct a label map \( y'_r \) which resembles \( y_A \). The second cycle-consistent loss is in the image domain:

\[
\mathcal{L}_{\text{cyc.I}} = \mathbb{E}_{x'_A, y_A} \left[ \| x'_A - x_A \|_1 + \lambda_{\text{VGG}} \| \phi(x'_A) - \phi(x_A) \|_1 \right].
\]  

(5)

Here, \( x'_A = F(y'_r, x_r) \) is a reconstructed face, which should be similar to the receptor face \( x_A \). Note that the background of \( x_A \) is also removed when computing \( \mathcal{L}_{\text{cyc.I}} \). It is reasonable to assume that \( y'_r \) resembles \( y_A \), given the cycle-consistent loss in the label domain. Thus, \( \mathcal{L}_{\text{cyc.I}} \) provides a supervised ground-truth to train \( F \) to generate a photo-realistic image \( x'_A \) which is corresponding to the label map \( y_A \). In addition, \( \mathcal{L}_{\text{cyc.I}} \) also encourages the remixed face \( x_r \) to provide correct identity information when acting as a conditional input. If \( x_r \) does not preserve the identity of \( x_A \), it cannot provide the correct identity information for reconstructing face \( x'_A \).

**Adversarial Losses.** To ensure that \( F \) is learning to generate a photo-realistic face image, an adversarial loss in the image domain is added:

\[
\mathcal{L}_{\text{GAN.I}} = \mathbb{E}_{x_A, x_B} \left[ \| D_I(x_A) \|_2^2 + \| D_I(x_B) \|_2^2 \right] + \mathbb{E}_{x_r} \left[ \| 1 - D_I(x_r) \|_2^2 \right] + \mathbb{E}_{x'_A} \left[ \| 1 - D_I(x'_A) \|_2^2 \right],
\]  

(6)

where \( x_r = F(y_r, x_A) \) and \( x'_A = F(y'_r, x_r) \). Note that the overall decoder \( G_D \) also undergoes finetuning during the training process. Thus, we need to make sure that \( G_D \) still generates a meaningful label map by introducing a label-domain adversarial loss \( \mathcal{L}_{\text{GAN.L}} \) similarly.
Identity Constraint. Since the traditional label-to-face transformation is a one-to-many mapping, we may have multiple face images corresponding to the same label map. The role of the conditional input image is to offer the identity information. The cycle-consistent losses mentioned above have already implicitly encouraged the transformer $F$ to learn the identity preservation. In order to explicitly constrain the identity, we propose a masked identity loss:

$$
L_{ID} = \mathbb{E}_{x, x_A} \left[ \left\| (x_r - x_A) \cap (1 - M_1) \right\|_1 \right] + \mathbb{E}_{x, x'_A} \left[ \left\| (x'_r - x'_A) \cap (1 - M_2) \right\|_1 \right].
$$

(7)

Here, $M_1$ and $M_2$ are editing masks computed by combining the $i$-th part-wise masks of both the receptor and donor, indicating the potential editing area. This loss term encourages that the content outside the editing area remains untouched.

The Final Objective. The final objective function is:

$$
L_{total} = L_{cyc} + \lambda_{CL} L_{cyc} + \lambda_{GL} L_{GAN} + \lambda_{ID} L_{ID},
$$

which can be optimized in the form of a min-max game:

$$
\arg \min_{F, G_D, \lambda} \max_{D_1, D_L} L_{total}(F, G_D, D_1, D_L).
$$

4. Experiments

Dataset. An internal face parsing dataset, containing 17,975 face images with ground-truth label maps, is utilized to train and test our network. The face images in this dataset are from CelebA [30]. For each image, we hire people to manually draw a pixel-wise label map, while each part is labeled with a unique color. The specific parts include hair, eyebrows, eyes, nose, upper lip, lower lip, teeth, face skin, body skin, clothes, and background. We split the dataset into three parts: 14,403 pairs for training, 1,781 pairs for validation, and 1,791 pairs for testing.

Implementation Details. During training, all the images and label maps are resized to $256 \times 256$. A random affine transformation is adopted for data augmentation. We adapt the architecture proposed by Johnson et al. [19] to build our system. Our part-wise encoder consists of 3 convolutional layers, 4 residual blocks [13], and 2 additional convolutional layers to generate a feature map of size $128 \times 1 \times 1$. Our part-wise decoder or the overall decoder consists of 2 transposed convolutional layers, 5 residual blocks, 2 more transposed convolutional layers, and one additional convolutional layer in the end to transform the feature map to a three-channel RGB label map. The architecture of $F$ consists of 5 convolutional layers, 9 residual blocks, and 4 transposed convolutional layers. Please refer to the appendix for architecture details. The Adam solver [21] with $lr = 0.0002$ and $(\beta_1, \beta_2) = (0.5, 0.999)$ is adopted. To learn the disentangled shape representation, we train the part-wise encoder-decoders for 100 epochs, and the overall decoder for 50 epochs. During the cyclic training of the whole system, we fix the disentangling encoder and train all the other networks together for 15 epochs. The best snapshot for each component is recorded according to the validation performance. When conducting the adversarial training, we adopt the gradient penalty term proposed by Gulrajani et al. [12] for a more stable convergence. We set the batch size to 1, and $\{\lambda_{VGG}, \lambda_{GAN}, \lambda_{ID}, \lambda_{GL}, \lambda_{GI}\} = \{1.0, 0.1, 1.0, 0.1, 0.1\}$ in our experiments. Since our conditional label-to-face transformer only focuses on generating the foreground, we adopt an off-the-shelf image inpainting method [28] to complete the background, on which the generated foreground is pasted.

4.1. Losses for Training The Overall Decoder

We test three different loss settings for training the overall decoder, including $L_1$, $L_1$+VGG, and $L_1$+VGG+GAN. From Fig. 6 we can see that using merely an $L_1$ loss leads to blurry results. Introducing a VGG loss can bring more details. By combining the $L_1$, VGG and GAN losses together,
we can enforce the overall decoder to produce high-quality label maps with finer details.

### 4.2. Part-Wise Shape Editing for Faces

In this session, we will show some part-wise shape editing manipulations supported by our proposed system. Firstly, as shown in Fig. 7, we can easily transfer the hair style of a reference face $x_B$ to the input face $x_A$, while preserving the identity of $x_A$. Secondly, by interpolating the part-wise shape feature for the hair, we can gradually change the hair style of the target person from one to another, as shown in Fig. 8. This indicates that our Face-ShapeGene lies on a continuous manifold, where similar shapes have similar features. Note that the shapes of other parts keep unchanged throughout the whole interpolation process, which indicates that we successfully disentangle the shape features of different facial parts. Thirdly, besides using the shape-remix network to provide a remixed label map, we can also manually edit a label map to flexibly control the shape of each facial part. Fig. 9 demonstrates that we can generate a desired photo-realistic face image by conveniently editing the label map.

Other than the hair, we can also manipulate the features of other facial parts to accomplish partial face editing. Examples of exchanging the eyebrows, mouths, and noses between two faces are shown in Fig. 10.

### 4.3. Comparison to Existing Methods

To our knowledge, our method is the first end-to-end system designed for automatically transferring the partial shape of one face to another. Thus, it is impossible to compare our full system to other methods. To demonstrate the superiority of the proposed training strategy for the conditional label-to-face transformer, we compare our method against two classic image-to-image translation methods Pix2Pix [18] and CycleGAN [41]. We also evaluate several variants of our proposed methods.

#### Qualitative Comparisons

Fig. 11 shows some visual results. Pix2Pix and CycleGAN cannot preserve the identity of the receptor $x_A$, while our method produces a photo-realistic remixed face keeping the identity of $x_A$. Removing $\mathcal{L}_{ID}$ impedes the identity preservation ability a little bit. The quality gets worse when $\mathcal{L}_{cyc,L}$ and $\mathcal{L}_{cyc,I}$ are also removed. Besides the desired part, other facial parts also change accordingly. Hence, the transformer fails to preserve the identity of the conditional input.

#### Quantitative Comparisons

To quantitatively compare our method with the CycleGAN and Pix2Pix baselines, and the variants of our method, we use Fréchet Inception Distance (FID) score [14], Inception Score (IS) [35], and Open-Face Score [1] as the evaluation metrics. FID uses the Inception network [38] to extract features from an intermediate layer, and then evaluate the distance between the feature distributions of the ground-truth images and the generated images. A smaller FID score indicates more photo-realistic results. Inception Score is computed as the KL-divergence between the conditional class distribution and the marginal class distribution, where the output class label of the generated image is predicted by the Inception network. To evaluate the identity preservation ability, we exploit the Open-Face model [1] to compute the face-id features of face images. The OpenFace score is computed as the dot product between the face-id features of the output face and the receptor face. Please refer to the supplementary material for the details of all the evaluation metrics.

We use all the images in the test set as receptors, while for each receptor we randomly choose a donor for it. The choice of the donor is fixed when testing different methods. The average scores over the test set are reported in Table 1. According to Table 1, our method produces the most photo-realistic results and has the best identity preservation ability. When $\mathcal{L}_{ID}$ is removed, the scores decline because the preservation of the identity is weakened. The scores get even worse when $\mathcal{L}_{cyc,I}$ and $\mathcal{L}_{cyc,L}$ are also removed. We can also find that $\mathcal{L}_{cyc,I}$ has a greater impact on the scores than $\mathcal{L}_{cyc,L}$, because $\mathcal{L}_{cyc,I}$ directly constrains the reconstructed image while $\mathcal{L}_{cyc,L}$ only constrains the intermediate result.

#### User Study

We also conducted a user study to subjectively compare different methods. Our method was compared to CycleGAN, Pix2Pix, and “w/o $\mathcal{L}_{ID}$”, respectively. We randomly selected 10 groups of hair editing examples for testing. In each example, we showed 4 images, namely the receptor $x_A$, the donor $x_B$, our result, and the result of an alternative method. On one hand, each participant was asked to choose a face image with a higher visual quality, considering both the overall fidelity and the success degree.

![Figure 9. The conditional label-to-face transformer can also be applied to a manually edited label map.](image)

| Method            | FID Score | Inception Score | OpenFace |
|-------------------|-----------|-----------------|----------|
| CycleGAN          | 45.07     | 2.296           | 1.947    |
| Pix2Pix           | 25.50     | 1.946           | 1.700    |
| w/o $\mathcal{L}_{ID}$ | 18.55     | 1.959           | 1.322    |
| w/o $\mathcal{L}_{ID} + \mathcal{L}_{cyc,I}$ | 18.15     | 2.214           | 1.252    |
| w/o $\mathcal{L}_{ID} + \mathcal{L}_{cyc,L}$ | 21.97     | 2.726           | 1.538    |
| Ours              | 14.22     | 1.912           | 1.116    |

Table 1. Quantitative Comparisons.
of changing the hair style. On the other hand, each participant was required to choose the result which preserves the identity better. 31 people participated in this study. When compared to CycleGAN, 84% of the participants thinks our synthesis quality is better, while 99% thinks our method preserves the identity better. When compared to Pix2Pix, the numbers are 56% and 93%. When compared to “w/o \( L_{\text{ID}} \)”, the numbers are 76% and 97%. The user study results show that our method performs better than other alternatives subjectively.

5. Conclusion

In this paper, we proposed FaceShapeGene, a disentangled shape representation for face, which encodes the shape information of each facial part separately. Exploiting the FaceShapeGene, we developed a novel face editing system, which includes a shape-remix network and a conditional label-to-face transformer. A cyclic training strategy was further proposed to train the system in an unsupervised manner. The extensive experiments demonstrate that our system can achieve state-of-the-art partial editing results.
References

[1] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016. 7

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In ICML, 2017. 1

[3] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. Cvae-gan: fine-grained image generation through asymmetric training. In ICCV, 2017. 1

[4] D. Berthelot, T. Schumm, and L. Metz. Began: boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717, 2017. 1

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI, 2018. 4

[6] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05887, 2017. 4

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, 2018. 4

[8] 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 CVPR, 2018. 2, 3

[9] L. Dinh, D. Krueger, and Y. Bengio. Nice: Non-linear independent components estimation. arXiv preprint arXiv:1410.8516, 2014. 1

[10] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. arXiv preprint arXiv:1605.08803, 2016. 1

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In NeurIPS, 2014. 1

[12] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. NeurIPS, 2017. 1, 6

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016. 6

[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In NeurIPS, 2017. 7

[15] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In ICLR, 2017. 1

[16] R. Huang, S. Zhang, T. Li, R. He, et al. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In ICCV, 2017. 2

[17] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In ECCV, 2018. 2

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In CVPR, 2017. 1, 2, 4, 7

[19] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In ECCV, 2016. 4, 6

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In ICLR, 2018. 1

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In ICLR, 2015. 6

[22] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In NeurIPS, 2018. 1

[23] D. P. Kingma and M. Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013. 1

[24] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In ICML, 2016. 1

[25] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In CVPR, 2017. 4

[26] H.-Y. Lee, J.-H. Tseng, J.-B. Huang, M. K. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations, 2018. 2

[27] M. Li, H. Huang, L. Ma, W. Liu, T. Zhang, and Y. Jiang. Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks. In ECCV, 2018. 1, 2

[28] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In ECCV, 2018. 6

[29] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In NeurIPS, 2017. 2

[30] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In ICCV, 2015. 2, 6

[31] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In ICCV, 2017. 4

[32] L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In ICML, 2017. 1

[33] A. Pumarola, A. Agudo, A. Martinez, A. Sanfeliu, and O. Winther. Least squares generative adversarial networks. In NeurIPS, 2016. 1

[34] A. Pumarola, A. Agudo, A. Martinez, A. Sanfeliu, and O. Winther. Least squares generative adversarial networks. In NeurIPS, 2016. 1

[35] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In NeurIPS, 2016. 7

[36] Y. Shen, P. Luo, J. Yan, X. Wang, and X. Tang. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In CVPR, 2018. 2

[37] K. Simonanyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015. 4

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In CVPR, 2015. 7
[39] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In CVPR, 2018. 1, 2

[40] B. Zhao, B. Chang, Z. Jie, and L. Sigal. Modular generative adversarial networks. In ECCV, 2018. 2, 3

[41] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In ICCV, 2017. 1, 2, 7