Towards Harmonized Regional Style Transfer and Manipulation for Facial Images

Cong Wang
Jilin University
Changchun, Jilin, China
cwang16@mails.jlu.edu.cn

Fan Tang
Jilin University
Changchun, Jilin, China
tangfan@jlu.edu.cn

Yong Zhang
Tencent
Shenzhen, China
zhangyong201303@gmail.com

Tieru Wu
Jilin University
Changchun, Jilin, China
wutr@jlu.edu.cn

Weiming Dong
CASIA
Beijing, China
wmlake@gmail.com

Abstract

Regional facial image synthesis conditioned on semantic mask has achieved great attention in the field of computational visual media. However, the appearance of different regions may be inconsistent with each other when conducting regional editing. In this paper, we focus on harmonized regional style transfer for facial images. A multi-scale encoder is proposed for accurate style code extraction. As the key part of our work, a multi-region style attention module is introduced to adapt multiple regional style embeddings from a reference image to a target image for generating harmonious result. We also propose style mapping networks for multi-modal style synthesis. Furthermore, we employ an invertible flow model which can serve as mapping network to fine-tune the style code by reversing the code to latent space. Finally, we conduct experiments on three widely used face datasets and we evaluate our model by transferring the regional facial appearance between datasets. Experimental results show that our model can generate reliable style transfer and multi-modal manipulation results compared with SOTAs.

1. Introduction

Semantic image synthesis [18, 54, 4, 44, 31, 56, 55] that aims to generate realistic natural images from semantic labels is an active research topic in the past few years. Based on the difference in the way of involving new styles for synthesis, there are two types of mainstream methods to generate diverse images: injecting random noise [18, 47, 54] or transfer from referenced images [13, 25, 55, 43]. Researchers have made great progress in both fields. Choi *et al*. [7] employ a style extraction net for facial style transfer and a mapping network adapted from StyleGAN [21, 22] to transform Gaussian noise into style codes.

(a) Target (b) Reference (c) SEAN (d) Ours

Figure 1. An example of skin transfer. R-ST&M method can modify the transferred skin according to the global lighting condition of target image. However, without considering the relationship between different regions, the synthesis region in (c) is not harmonized with other regions.

SPADE [31] adopts the idea of VAE [24] to encode the image style and enables both tasks. However, SPADE is just able to transfer facial style globally, thus limiting practical usage. Recent works [13, 55, 56] propose to extract style codes of all semantic components separately, enabling regional style transfer and manipulation (R-ST&M) for facial images.

R-ST&M provides a flexible way for facial image editing. However, new problems arise at the same time: regional appearance editing (i.e., transfer or manipulation) will lead to the appearance of different regions inconsistent with each other. For example, when transferring skin style from a target facial image to another captured with different light conditions, the new skin style generated by other method such as SEAN [55] will be abrupt to the rest regions in target image (Fig. 1). Similar problems have been realized in the field of image composition [9, 8, 41, 53]. To the best of our knowledge, there are no prior works that focus on style consistency and harmony for R-ST&M.

In this paper, we propose a framework which takes style

Figure 2. Examples of color editing by our model.

consistency of different regions into consideration for R-ST&M. We design a multi-scale encoder which incorporates feature maps from all original layers in SEAN encoder to extract style codes with richer style information, since low-level features are important for reconstruction [18, 54, 34].

In order to make the generated image with transferred style look plausible in synthesized image, we employ a multi-region style attention (MRSA) module where the relevance between the reference and target image is computed to synthesize a calibrated reference style. Apart from regional style transfer, we employ style mapping networks to map random vectors from latent space to the style spaces for region-wise multi-modal style synthesis. The idea of style mapping networks is inherited from StarGAN-v2 [7]. Differently, instead of training the mapping networks by adversarial loss of fake/real images, we calculate the adversarial loss on the style embedding space. The multi-scale encoder outputs multi-region style spaces with relationship among different regions, building the mapping networks directly from distributions can generate reliable regional styles. Furthermore, we train a continuous normalizing flow (CNF) [5, 12] which can reverse style code generated by the multi-scale encoder to latent space. Thus, we can fine-tune style codes from real images in latent space. To further evaluate "harmony" of synthetic images, we use a binary classification network to distinguish natural photographs from composite ones as done in [53]. With the proposed approaches, we set up two facial editing applications. Fig. 2 shows the example of harmonized color editing application.

To summarize, our main contributions are as follows:

- We focus on the appearance harmony among regions for R-ST&M tasks and introduce a multi-scale encoder that incorporate low- and high-level features to extract regional styles and style mapping networks to generate random styles for different semantics.
- We introduce a multi-region style attention module which facilitates harmony and consistency in regional style transfer.
- We conduct sufficient evaluations and show two new face editing applications to proof that the proposed framework can generate high quality facial images on various R-ST&M tasks.

2. Related Work

Facial Image Manipulation with GANs. Generative adversarial nets (GANs) [11, 18, 2, 21, 22, 20] have achieved great success in image generation. A GAN consists of two competitors, *i.e.,* a generator and a discriminator. The generator is trained to synthesize images that cannot be distinguished from real ones by the discriminator. However, the original GAN [11] suffers from mode collapse. Then lots of works are proposed to improve the generation quality of GANs, such as [10, 2, 14, 29, 49].

One of the most important applications of GANs is to generate photo-realistic human face images. PGGAN [20] is proposed to grow both the generator and discriminator progressively, allowing users to produce high-resolution and high-quality face images. StyleGAN [21] and StyleGAN2 [22] introduce a novel generator architecture borrowed from style transfer literature, enabling indistinguishable face images generation. In the field of facial image editing, significant progress has been made using powerful GANs. FaceShop [32] presents a novel system for face image manipulation by providing both geometry and color constraints as user-drawn strokes. DeepFaceEditing [6] is a structured disentanglement framework designed for face images to support face manipulation with disentangled control of geometry and appearance. MichiGAN [38] explicitly disentangles hair into four orthogonal attributes and designs a corresponding condition module to process user inputs for each attribute. DualFace [17] proposes a two-stage guidance system to help users produce detailed portrait sketch with data-driven global guidance and GAN-based local guidance. InterFaceGAN [35] explores the disentanglement between various semantic attributes and edits several attributes using linear editing path. SeFa [36] proposes a general closed-form factorization method for latent semantic discovery. StyleRig [40] proposes to provide a face rig-like control over a pretrained StyleGAN. StyleFlow [1] presents to utilize normalizing flows [33] for facial attributes editing interactively with StyleGAN.

Recent works [52, 34] learn to encode facial images for StyleGAN inversion and facilitate various image editing tasks. MaskGAN [25] proposes a face dataset with fine-grained mask annotations and dense mapping network for attribute transfer and style copy. However, MaskGAN just allows global style transfer. Sun *et al*. [37] use partial dilated layers to modify a few pixels in learned feature maps and realize mask-aware continuous facial attributes manipulation. Gu *et al*. [13] proposes an end-to-end framework to learn conditional GANs guided by semantic masks, enabling facial regional style transfer. SEAN [55] proposes semantic region-adaptive normalization for GANs condi-

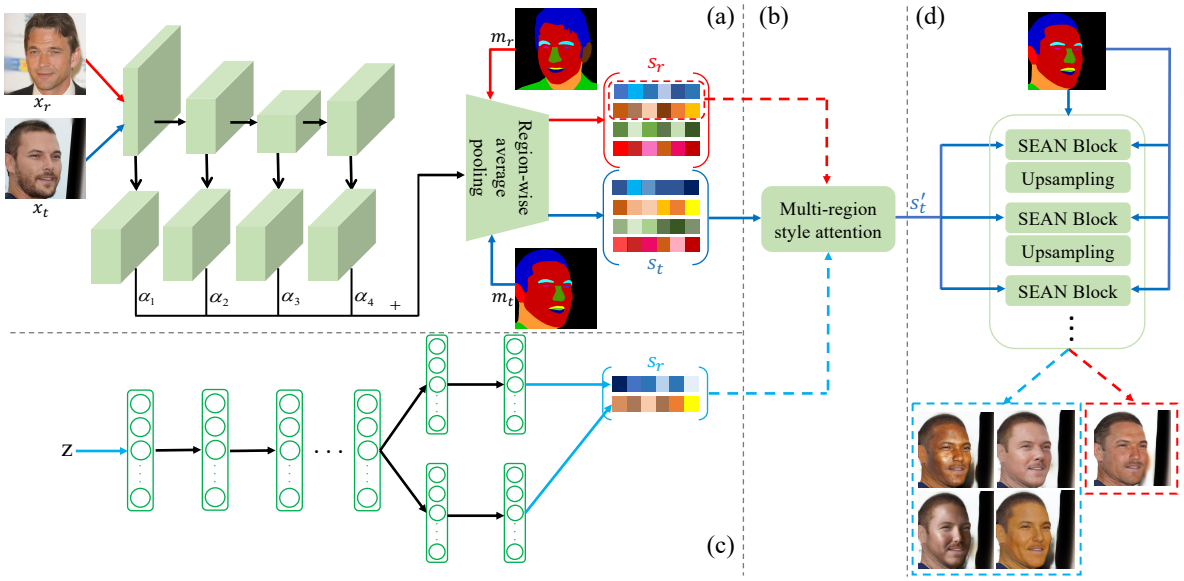

Figure 3. Whole framework of our model. (a) Multi-level feature fusion part of the encoder. (b) Multi-region style attention module. (c) An example of style mapping networks. In this example, the mapping network generates styles of skin and nose simultaneously, and multi-modal results are shown in (d). (d) SEAN generator and results of (a) and (c).

tioned on segmentation masks, and the model can control the style of each semantic region individually. Our work improves the SEAN encoder with a multi-scale structure and a multi-region style attention module for facial image harmonization. Moreover, we introduce style mapping nets to generate multi-modal styles regionally with latent codes sampled from Gaussian distribution.

Self-Attention. Self-attention is first proposed in the natural language processing literature by Transformer [42]. Then computer vision researchers extend the idea to video classification [45] and image generation [49]. Recent works generalize self-attention to extract the correspondence between source image and reference image for semantic style transfer [50, 27] and makeup transfer [19]. However, the self-attention mechanism computes the correspondence spatially, making it time-consuming and inefficient. Differently, our style attention inspired by the above works computes the correlation among semantic regional style vectors, which ensures its computation efficiency.

Multi-Modal Image Synthesis. BicycleGAN [54] models a distribution of possible outputs in a conditional generative modeling setting. To ensure that random sampling can be used during testing, the model employs KL-divergence loss to enforce the latent style distribution to be close to a standard normal distribution. [16, 26] extend the idea of multi-modal to unsupervised image-to-image translation and generate diverse images. SPADE [31] uses the same idea to encode image style for semantic image synthesis. GroupDNet [56] extends SPADE by using KL loss for all the semantic labels, thus enabling regional multi-modal synthesis. Recently, StarGAN-v2 [7] is proposed to learn a mapping network to achieve diversity, and our style mapping model is the same as StarGAN-v2 but with different training strategy which is more suitable for our framework.

Deep Image Harmonization. Deep convolutional models have achieved significant success for image harmonization in recent years. Zhu *et al*. [53] train a binary classifier to guide color adjustment for composite images. Then, an end-to-end deep CNN model is proposed to capture both the context and semantic information during harmonization. Cun *et al*. [46] spatial-separated attention module in order to learn the feature map in the foreground and background individually. DoveNet [8] translates the foreground domain to the background domain by using a domain verification discriminator and generates impressive results. Since facial regional style transfer may lead to inharmony, we propose a multi-region style attention module to adjust the transferred regional style to other regions. The proposed module is inserted in the process of style transfer. Therefore, harmonious images can be directly synthesized without a subsequent image harmonization process. Moreover, we use the method proposed by Zhu *et al*. [53] to evaluate the level of harmonization of an image.

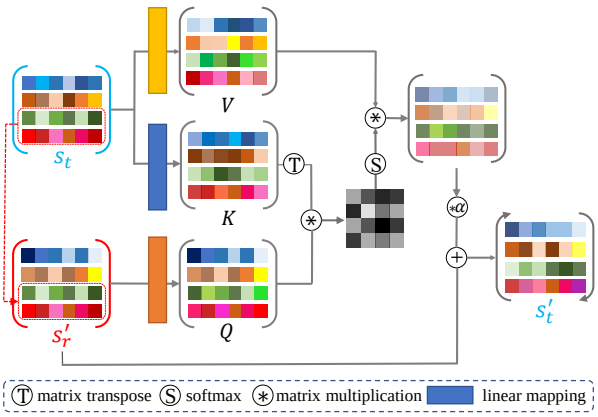

Figure 4. Multi-region style attention module. s_t denotes target style vectors of all regions. s'_r is the concatenation of styles from target regions in the reference and styles from the rest regions in the target. Linear projection metric $\mathcal{W}_v$, $\mathcal{W}_q$ and $\mathcal{W}_k$ are adopted to produce V, Q and K, respectively. Then, V and Q are used to yield an attention matrix M. Finally, the multi-region style correction is calculated by $M * V$ which is applied to s'_r.

3. Framework Architecture

Fig. 3 shows the framework of the proposed multi-region style transfer and multi-modal synthesis method. The inputs are a target image x_t that the user wants to edit with segmentation and a reference style. The reference style can either be generated from a reference *style image* x_r with segmentation for style transfer, or directly sampled from a normal Gaussian distribution for manipulation. In this section, we start from introducing the regional feature encoding, including a multi-scale encoder for input images and regional style mapping (**RSM**) subnets for multi-modal style synthesis. We then move on to the multi-region style attention (**MRSA**) module followed by a semantic region-adaptive normalization based decoder. Next, we discuss the supervised training strategy and details. Finally, we demonstrate how to fine-tune a real style code based on normalizing flow models.

3.1. Regional Feature Encoding

Multi-scale Encoder. The encoder in SEAN employs a "bottleneck" structure with plain convolutional layers to extract styles of all facial semantic regions. Since the purpose of the model is to generate images from the output of encoder, the low-level features from shallow layers are important for image reconstruction. Therefore, we compute the weighted summation of feature maps from all layers in encoder, as shown in Fig. 3(a). Concretely, we first re-scale the feature maps to a unified resolution and get new features $\{F_i\}_{i=1}^K$, where K is the number of shallow layers. Then, a set of learnable parameters $\{a_i\}_{i=1}^K$ are defined and we feed them into a softmax function for normalization as:

$$\{\alpha_i\}_{i=1}^K \leftarrow \text{softmax}(\{a_i\}_{i=1}^K). \tag{1}$$

After that, we get the final multi-scale style feature map,

$$F = \sum_{i=1}^{K} \alpha_i F_i. \tag{2}$$

The learned weights $\{a_i\}_{i=1}^K$ indicate the proportion of each scale for compositing the feature map F. Given an input target image x_t and a reference image x_r with their segmentation masks (m_t and m_r), we employ the region-wise average pooling layer [44, 55] to transform F_t and F_r to initial style vectors s_t and s_r respectively.

Regional Style Mapping. In order to synthesize multi-modal facial images with random styles, we utilize a series of regional style mapping sub-networks to learn the distributions of styles from different facial regions respectively. Facial semantic regions can be divided into several groups according to their relevance, and one network is responsible for one group. For example, some regions such as skin and nose that share same color and texture appearance are strongly correlated, so we should define one network to model them simultaneously. In practice, we only train hair and skin network as the area of these regions is large enough. As the correlations among some regions such as nose and hair are weak, we use two networks to model them separately. Fig. 3(c) shows an example of the mapping subnetwork for modeling skin and nose. Given a latent code z sampled from the Gaussian distribution, a random reference style can be generated with the mapping network $\mathcal{M}$,

$$s_r = \mathcal{M}(z). \tag{3}$$

In our method, related regions such as skin and nose or two eyes share a same mapping network. More details of the training of RSM are in Sec. 3.4. After that, we can feed s_r into the MRSA module and generator $\mathcal{G}$.

3.2. Multi-region Style Attention

If the global appearances (i.e., lighting conditions) in x_t and x_r are quite different, regional style transfer results probably become inharmonious. However, users prefer to get a harmonious image directly without a subsequent image harmonization process. To this end, we propose a multi-region style attention (MRSA) module to learn transferred styles. Fig. 4 illustrates the workflow of MRSA. Different from the attention modules in [50] and [27] that extract the spatial correspondence in pixel space, our MRSA module computes the relevance of regional semantic styles. In order to correct the styles of different regions, we first concatenate the target components in s_r with the rest components

in s_t to form a new s'_r. Then we map the style vectors using $Q = \mathcal{W}_q(s'_r)$, $K = \mathcal{W}_k(s_t)$ and $V = \mathcal{W}_v(s_t)$, where $\mathcal{W}_q$, $\mathcal{W}_k$ and $\mathcal{W}_v$ are linear mappings. After that, an attention matrix can be computed by $Q * K^\top$ followed by a softmax function within each row, *i.e.,*

$$M = \text{softmax}(Q * K^\top), \tag{4}$$

where * denotes matrix multiplication. After computing the attention matrix M, we can get the style correction $s_c = M * V$. Finally, the target style can be computed by

$$s'_t = s'_r + \alpha s_c. \tag{5}$$

3.3. Decoder

Given the style vectors generated by MSRA, the SEAN generator [55] is used as decoder by feeding them into a semantic region-adaptive normalization (SEAN) module. In the SEAN normalization, target mask along with style map generated by broadcasting style vectors to the corresponding regions are used to modulate the activation from previous layer. The decoder employs several SEAN blocks with upsampling layers and synthesizes images progressively.

3.4. Model Training

The encoder-decoder part in our model is similar with SPADE and SEAN. We use three loss functions described in SPADE and SEAN to train this part: adversarial loss, feature matching loss and perceptual loss. During training, if we use the s_r extracted from a reference different from the source image x_s, this results in an unsupervised training as there is no ground truth for the new image. To tackle this problem, we set x_r equal to x_s for training. We test the training strategy by mixing supervised with unsupervised training, but it fails to generate realistic images. The reason we suppose is that the unsupervised result would disturb supervised training pace.

As for style mapping networks $\{\mathcal{M}_j\}_{j=1}^M$, we turn to the adversarial loss imposed on s_s and s_r generated by style mapping. M is the number of mapping networks. In order to train $\{\mathcal{M}_j\}_{j=1}^M$, a set of discriminators $\{\mathcal{D}_j\}_{j=1}^M$ are employed and the adversarial objectives are as follows:

$$\mathcal{L}_j = \min_{\mathcal{M}_j} \max_{\mathcal{D}_j} \mathbb{E}[\log \mathcal{D}_j(s_s)] + \mathbb{E}[\log(1 - \mathcal{D}_j(\mathcal{M}_j(z)))]. \tag{6}$$

A similar style mapping network has been proposed in StarGAN-v2 [7] which focuses on unsupervised image-to-image translation. However, StarGAN-v2 trains it with the adversarial loss defined on image synthesis. The training strategy in StarGAN-v2 cannot effectively train our style mapping networks. The reason is that our encoder-decoder part is trained in a supervised way, and the encoder will learn expressive style information. It is more effective to learn the distributions of encoded styles directly.

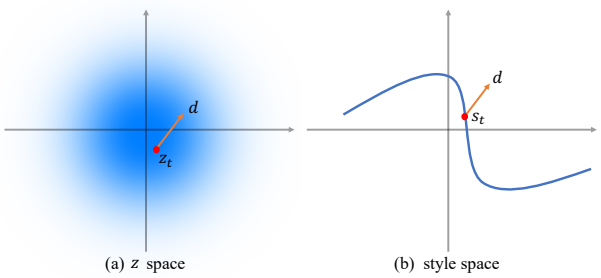

Figure 5. Style fine-tuning in z space and style space. Manipulation in style space will lead to a out-of-manifold result.

3.5. Style Random Fine-tuning

If users are not very satisfied with current style, we also provide a method for style fine-tuning that users can adjust to a new style based on the current one. A straightforward idea for doing this is to sample a random unit vector d as the tuning direction, then we can move the style code s_t forward along d. However, the style distribution in style space is supported on a low dimensional manifold since the style code s contains abundant semantic information. As shown in Fig. 5, fine-tuning in z space is more reasonable as the support of Gaussian distribution is the whole space. In order to realize style random fine-tuning in z space, we train a invertible continuous normalizing flow (CNF) [5, 12] which is utilized in [48] for point cloud generation and [1] for facial attributes manipulation.

Specifically, we first use the ODE below to get z_t corresponding to s_t,

$$z_t = s(l_0) + \int_{l_0}^{l_1} f(s(l), l) dl, \tag{7}$$

where $s(l_0) = s_t$ and l denotes time. Then we fine-tune z_t by $z'_t = z_t + \eta \cdot d$, where η is the step size. After that, a reverse-time ODE is employed to recover a modified meaningful style code,

$$s'_t = z'(l_1) + \int_{l_1}^{l_0} f(z'(l), l) dl, \tag{8}$$

where $z'(l_1) = z_t$.

4. Experiments

4.1. Experimental Settings

Datasets. We use three face datasets to evaluate our framework:

- CelebAMASK-HQ [25] consists of $30,000$ face images with segmentation masks. Each image is annotated with a semantic mask of 19 semantic categories in total. We use the first $28,000$ images for training and the remains for evaluating.

- FFHQ [21] contains $70,000$ high-quality images. We utilize a deeplab-v3 model [3] trained on CelebAMASK-HQ to parse the facial semantics. We employ the first $2,000$ images for evaluation.
- LaPa [28] is a new dataset for face parsing which consists of more than $22,000$ images with large variations in pose, facial expression and illumination. 11-category semantic label maps are provided. We discard low-resolution images in the dataset. The final training set contains $19,770$ faces and testing set contains $1,930$ faces.

Metrics. We employ several commonly used metrics to evaluate our framework and the competing state-of-the-art methods. Specifically, FID [15] computes the distance between the distributions of synthesized images and the distribution of real images, which is used to evaluate the quality of synthesized results. We also adopt PSNR, SSIM and LPIPS [51] to assess the similarity between the synthesized and the ground-truth image in face reconstruction task. In order to evaluate the performance of our model for regional multi-modal synthesis with random styles, we utilize mean Class-Specific Diversity (mCSD) and mean Other-Classes Diversity (mOCD) [56]. For a fixed semantic region, mCSD is used to assess the generation diversity of the region while mOCD is used to assess the diversity of the rest regions. Apparently, high mCSD and low mOCD indicate good performance for the fixed region.

In addition to the above metrics, we employ **harmony score** (HS) to measure the harmony degree between the transferred region and the rest for regional style transfer. The idea of harmony score is the same as realism score [53] predicted by a binary classifier. Concretely, we train a convolutional neural network to distinguish real images from synthetic ones and use the output probability as the harmony score. The real images are set as positive samples and the unrealistic composite images are set as negative samples. We use HAdobe5k [8] to train the classification network and concatenate one image and the corresponding foreground mask as an input.

Competing methods. We compare our method with five leading semantic image synthesis models: pix2pixHD [44], SPADE [31], GroupDNet [56], SEAN [55] and CLADE [39]. Specifically, pix2pixHD applies an image feature encoder network and instance-wise pooling to get image features within each object. Then, the features and the corresponding mask are feed into a coarse-to-fine generator to reconstruct the image. Therefore, pix2pixHD is suitable for regional style transfer. SPADE proposes the encoder and generator to form a VAE [24] and a new normalization for the generator, enabling global style transfer and multi-modal synthesis conditioned on semantic mask. GroupDNet extends the idea of SPADE by encoding different semantic regions separately and leveraging group decreasing generator. GroupDNet can be used for regional style transfer and multi-modal synthesis. SEAN employs similar structures of pix2pixHD encoder and SPADE generator, and it improves the generation quality significantly with the SEAN normalization. CLADE improves SPADE based on the observation that its modulation parameters benefit more from semantic-awareness rather than spatial-adaptiveness. Tan *et al*. [39] also introduce CLADE-ICPE where intra-class positional map encoding are proposed to improve spatial-adaptiveness.

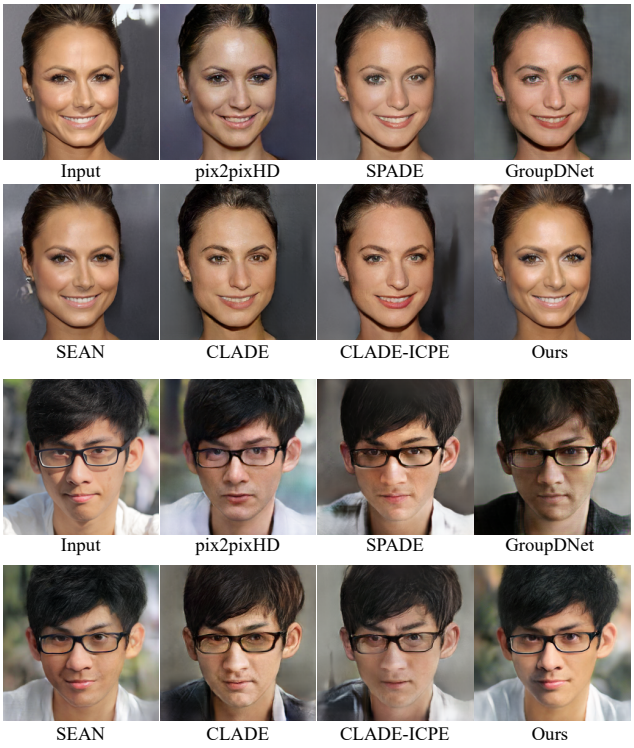

Figure 6. Results of image reconstruction.

4.2. Implementation Details

We use the TTUR [15] strategy and set the learning rate to 0.0001 and 0.0004 for the generator and discriminator, respectively. Following SPADE and SEAN, we apply Spectral Norm [30] to the encoder. Moreover, we use the ADAM solver [23] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ to optimize the model. For style mapping, we set the learning rate to 0.0002 for both mapping networks and discriminators. For both training and evaluation, the input images are resized to a fixed resolution of 256×256.

Table 1. The results of facial image reconstruction. For PSNR and SSIM, the higher the better. For LPIPS and FID, the lower the better.

	CelebAMASK-HQ				FFHQ				LaPa			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
pix2pixHD [44]	17.32	0.5387	0.2117	21.68	16.08	0.5200	0.2506	45.55	13.16	0.4387	0.3817	68.87
SPADE [31]	16.87	0.5142	0.2462	25.46	15.82	0.4894	0.2923	53.10	**14.82**	0.4607	0.3927	89.96
GroupDNet [56]	16.40	0.5184	0.2526	38.87	15.27	0.4913	0.2981	71.83	14.22	0.4454	0.3928	93.35
SEAN [55]	18.55	0.5741	0.1749	17.12	17.23	0.5368	0.2099	34.29	14.72	0.4841	**0.3281**	47.94
CLADE [39]	16.18	0.4863	0.2518	24.47	15.16	0.4653	0.2952	57.45	14.67	0.4681	0.4012	76.58
CLADE-ICPE [39]	16.57	0.4997	0.2507	23.75	15.57	0.4794	0.2967	56.46	14.07	0.4495	0.3925	85.93
Ours	**18.60**	**0.5787**	**0.1702**	**15.26**	**17.41**	**0.5510**	**0.2020**	**32.84**	14.75	**0.4891**	0.3295	**46.62**

Table 2. The FID↓ results of skin and hair transfer.

	CelebAMASK-HQ		FFHQ		LaPa	
	skin	hair	skin	hair	skin	hair
pix2pixHD	26.39	26.58	57.13	56.44	85.49	84.72
GroupDNet	45.65	44.09	77.89	78.20	104.52	104.02
SEAN	24.04	24.59	44.84	43.64	61.19	60.68
SEAN+DoveNet	29.62	24.67	52.96	44.32	65.70	63.46
Ours	**22.65**	**22.97**	**42.82**	**41.85**	**60.14**	**58.85**

4.3. Global Reconstruction

We first evaluate the effectiveness of the proposed multi-region style control and manipulation network in image reconstruction task, namely transferring the own style to itself. Only one image is employed as input. Visual comparisons are shown in Fig. 6. Overall, pix2pixHD and groupDNet cannot maintain the skin color well. Compared with SEAN, our method can reconstruct more facial details of the input, *e.g.* the wrinkles on the left side of the woman's face and the left eye under sunglasses of the man. In terms of quantitative evaluation, as shown in Table 1, our model outperforms other SOTA methods on all datasets. It is worth mentioning that although MRSA is designed for style transfer and manipulation, it exhibits the best reconstruction quality (*i.e.,* the lowest FID) on all datasets.

4.4. Regional Style Transfer

We further evaluate the effectiveness of the proposed approach in regional style transfer task. One target image and one reference image are employed as inputs. We split all testing datasets into two parts: one half as target images and the other half as reference images. SPADE is not selected for comparison since it does not support region style transfer. Fig. 7 (a) shows the quantitative results. We use FID measured on the whole image as the metric in two transfer tasks: skin (with nose) transfer and hair transfer. The quantitative results are shown in Table 2. In terms of FID, our model with style attention achieves the lowest values, indicating that it synthesizes human faces with the highest quality.

Transfer Cross dataset. Most images in CelebAMask-HQ [25] and FFHQ [21] are captured in good lighting conditions. Regional style transfer among these images can barely lead to inharmony. However, LaPa [28] consists of facial images with abundant variations in lighting conditions. Therefore, we transfer skin (with nose) and hair of facial images in the test sets of CelebAMask-HQ, FFHQ and LaPa to the test set of LaPa separately, and calculate FID and HS of synthesized faces shown in Table 3. According to the quantitative results, our method and SEAN [55] perform much better than pix2pixHD [44] and GroupDNet [56] in terms of FID, corresponding to higher image quality. We can draw the same conclusion from additional visual results in Fig. 7. HS reflects the harmony degree between the transferred region and the rest regions. Our method exhibits obviously higher harmony score than SEAN, showing the effectiveness of MRSA. However, pix2pixHD and GroupDNet reach higher harmony score than our model. As shown in Fig. 7, although the results of pix2pixHD and GroupDNet are harmonious, the two methods fail to reconstruct the transferred styles and the rest regions expected to keep their appearance change severely. In summary, our model is the best trade-off considering image quality and harmony degree.

We can see that the area outside of the region of interest is also changed a lot especially the background in Figure 7 (b). This is because the background contains rich diversity and the 512-dimensional (following SEAN) style code cannot reconstruct background accurately. It is not the transferred region that affects the background.

User Study. We conduct user studies to further compare the visual performance of ours and the SOAT methods aforementioned. Firstly, we show the participants each target-reference pair and tell them which region in the target image we want to edit. Then we show them four results, one is by our method and the others are from pix2pixHD, GroupDNet and SEAN. Each subject are assigned with 30 group results. We receive 59 responses, among which 47 responses are valid. A total of $1,410$ votes are obtained. Our model has 627 (44.45%) votes, SEAN has 410 (29.07%) votes, GroupDNet has 265 (18.78%) votes, and Pix2Pix has 108 (7.66%) votes.

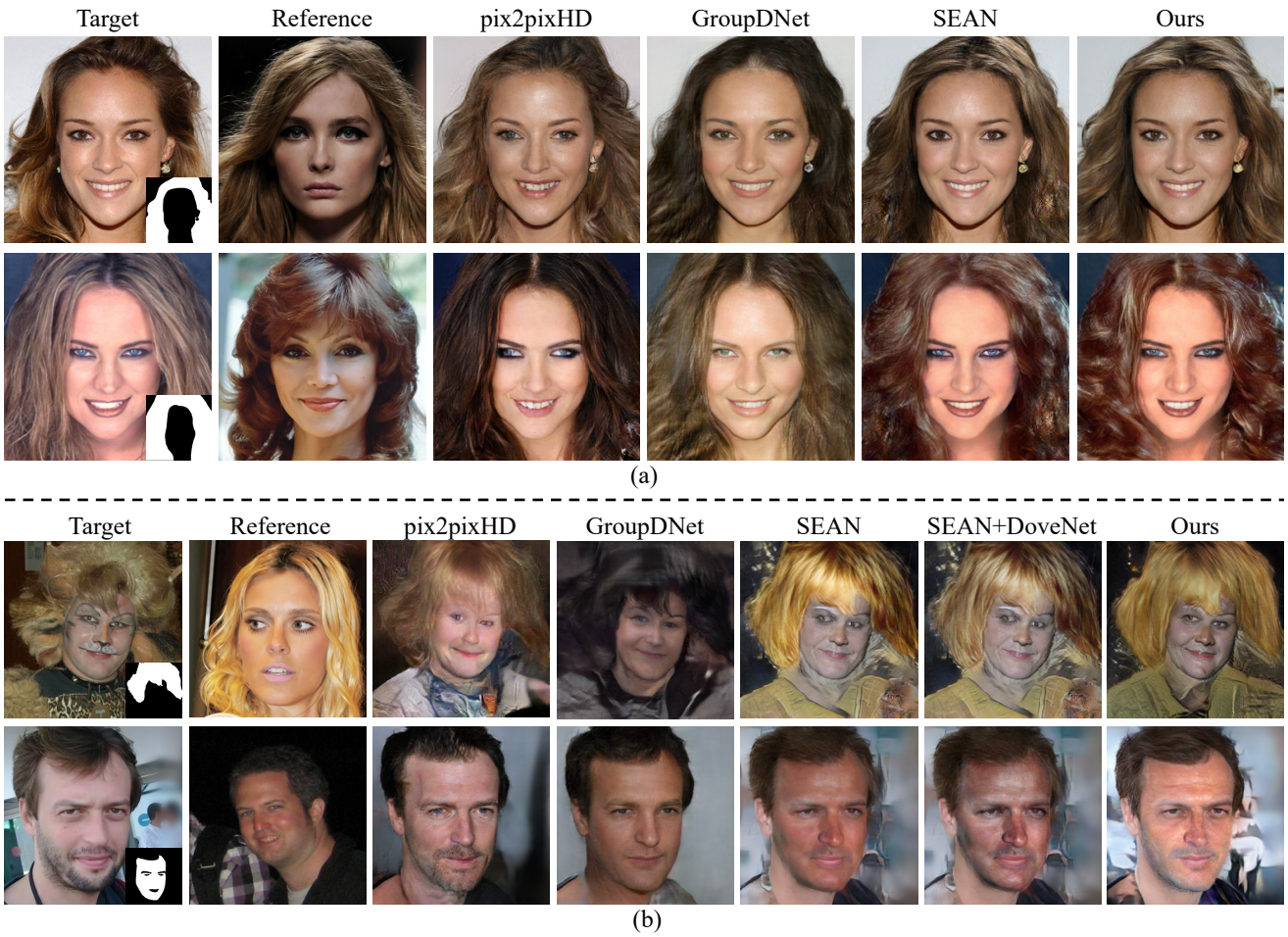

Figure 7. Results of regional style transfer (segmentation mask shown as small inset). Target and reference images in (a) are from the same dataset. (b) shows results of cross dataset transfer. We can see that our model generates more harmonious results than DoveNet [8] performed on the outputs of SEAN [55].

Table 3. Results (PSNR↓/HS↑) for regional style transfer cross dataset. Although pix2pixHD and GroupDNet perform higher HS than ours, the two methods achieve harmony by modifying other regions severely (Fig. 7). This is contradictory to regional style transfer.

	CelebAMASK-HQ→LaPa		FFHQ→LaPa		LaPa→LaPa	
	skin	hair	skin	hair	skin	hair
pix2pixHD [44]	73.62/0.7923	73.80/0.8075	77.36/0.8538	73.98/0.8137	85.49/0.8314	84.72/0.8035
GroupDNet [56]	96.20/**0.8965**	93.36/**0.8752**	94.78/**0.9131**	93.34/**0.8753**	104.52/**0.9093**	104.02/**0.8736**
SEAN [55]	48.25/0.7420	48.43/0.6940	48.17/0.7105	48.62/0.7164	61.19/0.7396	60.68/0.6996
Ours w/o softmax	52.98/0.8071	54.90/0.7353	53.10/0.8100	54.68/0.7495	66.67/0.8003	65.25/0.7348
Ours w/o SA	48.06/0.7749	47.72/0.7130	47.84/0.7598	48.39/0.7310	61.43/0.7872	59.99/0.6964
Ours	**47.46**/0.8490	**47.05**/0.7742	**46.71**/0.8341	**47.36**/0.7854	**60.14**/0.8537	**58.85**/0.7566

4.5. Comparison to DoveNet

Since our work is related to image harmonization, we utilize a recent deep harmonization model DoveNet [8] to harmonize the outputs of SEAN for comparison. As shown in Table 2, DoveNet has adverse effect on image synthesis quality as DoveNet gets higher FID scores than SEAN. For image harmonization shown in Fig. 7 (b), DoveNet indeed harmonize the output of SEAN, but it has limited effect. Since DoveNet is a subsequent and independent process and the reference is invisible to it, DoveNet changes the original tone from reference during harmonization. More importantly, a separate harmonization process will take extra time.

Figure 8. Skin multi-modal synthesis.

4.6. Regional Multi-modal Manipulation

We then evaluate the effectiveness of the proposed approach in regional multi-modal manipulation task. One target image and one vector sampled from a normal Gaussian distribution are employed as inputs. SPADE [31], GroupDNet [56], CLADE [39] and CLADE-ICPE [39] are selected for multi-modal synthesis. SPADE and CLADE are proposed for global multi-modal synthesis while GroupDNet is for regional multi-modal synthesis. Fig. 8 shows the manipulation of skin, GroupDNet affects hair more significantly than ours when doing skin multi-modal synthesis. We further conduct qualitative experiments on manipulation of skin and hair regions. Table 4 reports the FID, mCSD and mOCD calculated over the three different datasets. In terms of image quality, our method outperforms other methods by a large margin on all the datasets. In terms of diversity (mCSD), SPADE, CLADE, CLADE-ICPE and our method are in the same level. But SPADE, CLADE and CLADE-ICPE fail to preserve the appearance of other regions (higher mOCD), as they are de-

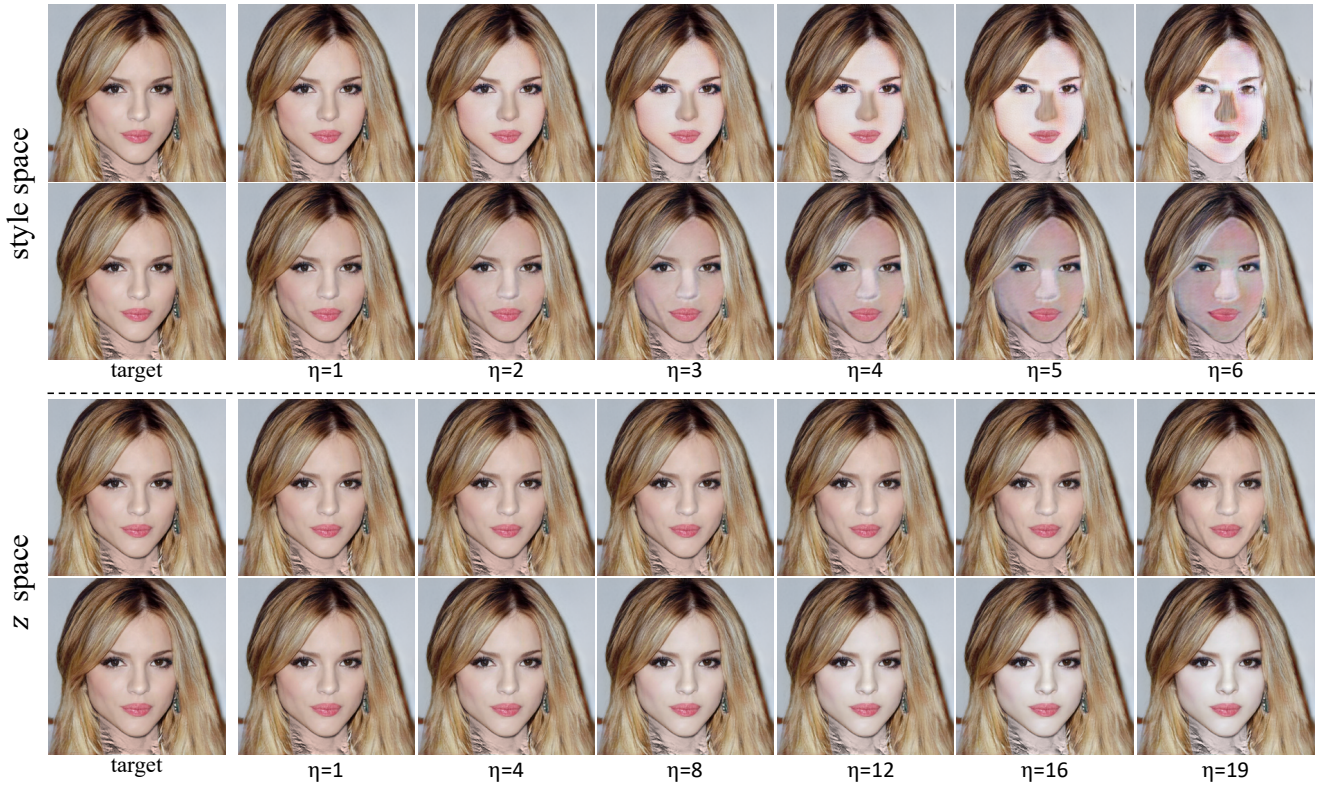

Figure 9. Two examples of style random fine-tuning for skin both in z and style spaces. Tuning in z space can achieve the goal of convincing style fine-tuning, such as gradually changing skin color and adding wrinkles.

Table 4. Regional multi-modal synthesis.

			SPADE	GroupDNet	CLADE	CLADE-ICPE	Ours w/ CNF	Ours
CelebA MASK-HQ	Skin	FID↓	21.09	39.72	21.39	19.40	**12.21**	12.67
		mCSD↑	0.0354	0.0321	**0.0437**	0.0416	0.0408	0.0395
		mOCD↓	0.2126	0.1280	0.2561	0.2382	**0.0721**	0.0752
	Hair	FID↓	21.12	50.43	21.42	19.39	12.84	**12.55**
		mCSD↑	0.1848	0.0001	0.1855	0.1954	**0.2323**	0.2078
		mOCD↓	0.1230	**0.0000**	0.1203	0.1417	0.0585	0.0505
FFHQ	Skin	FID↓	51.38	72.34	55.38	52.24	**30.57**	31.43
		mCSD↑	0.0392	0.0360	0.0395	0.0393	**0.0458**	0.0413
		mOCD↓	0.2020	0.0820	0.2097	0.2278	0.0285	**0.0279**
	Hair	FID↓	51.36	81.71	55.32	52.28	**28.45**	**28.45**
		mCSD↑	0.0723	0.0000	0.1167	**0.1533**	0.0826	0.0875
		mOCD↓	0.1920	**0.0000**	0.1757	0.1797	0.0157	0.0150
LaPa	Skin	FID↓	74.61	96.75	53.83	60.29	40.54	**40.47**
		mCSD↑	0.0455	0.0446	0.0462	0.0466	**0.0685**	0.0600
		mOCD↓	0.3005	0.1657	0.3375	0.3201	0.1185	**0.1071**
	Hair	FID↓	74.68	150.46	53.76	60.21	**41.11**	41.26
		mCSD↑	0.0512	0.0047	0.0884	0.0957	**0.1076**	0.0958
		mOCD↓	0.3080	**0.0000**	0.3299	0.3405	0.1238	0.1004

signed for global synthesis. For skin multi-modal synthesis, our method presents higher mCSD and lower mOCD than GroupDNet, even though it extends SPADE to regional style synthesis. That is to say, our method is better at maintaining the appearance of the rest regions while achieving high color and texture diversity of skin synthesis. For hair multi-modal synthesis, GroupDNet generates facial images with low diversity since mCSD and mOCD of GroupDNet

Table 5. Results (FID↓) for ablation study of RSM.

		SEAN+ GroupDNet	SEAN+ StarGAN-v2	SEAN+ Our RSM	Ours
CelebA MASK-HQ	Skin	25.32	27.19	14.53	**12.67**
	Hair	28.82	20.05	14.56	**12.55**
FFHQ	Skin	55.02	40.88	32.06	**31.43**
	Hair	58.07	33.57	30.27	**28.45**
LaPa	Skin	84.72	105.30	41.93	**40.47**
	Hair	87.29	90.75	43.30	**41.27**

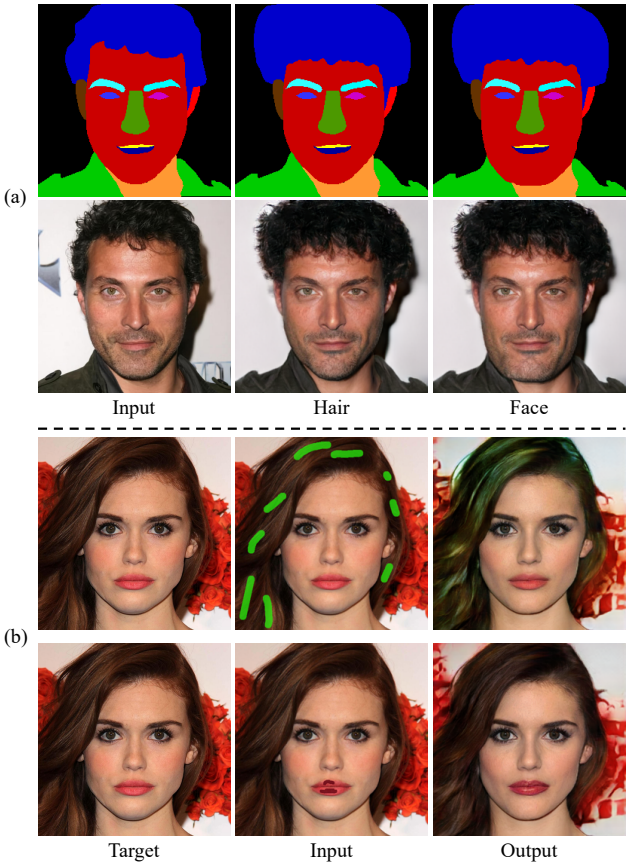

Figure 10. Applications (zoom in for details). (a) Shape editing. (b) Color editing (hair/lips).

are both close to zero. For all the multi-modal synthesis experiments, we manipulate each image using 10 random styles.

4.7. Style Fine-tuning

First, we evaluate the style synthesis quality of CNF model. As shown in Table 4, CNF performs closely to GAN on style multi-modal synthesis. Although our CNF model can be used for regional multi-modal synthesis, we only recommend it for style fine-tuning since the CNF runs much slower than a GAN model. In practical, our model with CNF takes 0.2s to generate a style code while our GAN mapping network takes 6e-4s on a single RTX3090.

Second, we validate the analysis in Section 3.5 by showing examples of style random fine-tuning both on z space and style space. Concretely, Fig. 9 demonstrates that the skin tuning results in z space are much better than in style space. Small step tuning in style space hardly affects the style of target image while larger step tuning fails to generate clear results. Recall that skin style code consists of codes of two parts, tuning skin style code will lead to deviation from the manifold. Thus, the nose presents different style from face skin even if the step size η is small. We can get the same conclusion on hair style fine-tuning. More results are shown in supplementary file.

4.8. Ablation Study

Ablation of MRSA. To validate the effects of the softmax function and MRSA module in the encoder, we conduct ablation experiments by not using them in the framework. "Ours w/o softmax" means we do not use softmax normalization and MRSA and "Ours w/o SA" means we do not use MRSA in our framework. Results of cross-dataset regional style transfer in Table 3 indicate that softmax normalization and MRSA improves image quality and harmony degree, respectively.

Ablation of RSM. To validate the effects of the RSM module, we conduct ablation experiments by combining the encoders and training strategies used in GroupDNet [56], StarGAN-v2 [7] and ours with the SEAN generator, respectively. We can get three variations for comparisions, *i.e.*, "SEAN+GroupDNet", "SEAN+StarGAN-v2" and "SEAN+Our RSM". Table 5 illustrates FID of skin and hair multi-modal synthesis on all the datasets. "SEAN+Our RSM" method performs much better than the two variations in terms of image quality. Our RSM uses similar mapping network as StarGAN-v2 but different training strategy. If we use the training strategy in StarGAN-v2, the generator would be trained in an unsupervised way. However, our encoder-decoder part is trained in a supervised manner and the mapping network shares the same generator with the encoder. Thus, different objectives would misguide the generator. Visual comparisons can be found in Fig. 8.

5. Applications

Our framework can enable various applications in facial image synthesis. Sections 4.4 and 4.6 demonstrate the effectiveness of regional style transfer cross facial images and multi-modal synthesis with random styles, respectively. We now introduce other two applications of interactive face editing.

Shape editing. Our framework allows users to edit the shape of facial components directly on segmentation mask to manipulate face interactively. Fig. 10 (a) shows an example of hair and face shape editing.

Color editing. By drawing simple color strokes on facial components, our method enables color editing on facial semantic regions. The two rows in Fig. 10 (b) demonstrate hair and lips color editing, respectively.

6. Conclusion

In this paper, we focus on the harmonized region style editing for facial images. The proposed framework follows the encoding-fusion-decoding fashion. For the encoder, we employ a multi-scale structure in order to extract regional styles more effectively. Then a multi-region style attention (MRSA) module is proposed for harmonious regional style transfer, especially when the target and reference face images are with different lighting conditions. For the sake of regional multi-modal synthesis, we introduce the regional style mapping (RSM) net to map random noise to styles.

Although our model can generate high quality regional multi-modal results with random styles, the styles of specific region are still in weak control condition. The regional style transfer is the only way to provide strong control information. If we want to randomly synthesize regions with specified appearance, our model, SPADE and GroupDNet will be helpless. This problem remains to be resolved and it will be our future work.

References

[1] R. Abdal, P. Zhu, N. J. Mitra, and P. Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *arXiv preprint arXiv:2008.02401*, 2020. 2, 5

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017. 2

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 6

[4] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[5] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. In *NIPS'18 Proceedings of the 32nd International Conference on Neural Information Processing Systems*, volume 31, pages 6572–6583, 2018. 2, 5

[6] S.-Y. Chen, F.-L. Liu, Y.-K. Lai, P. L. Rosin, C. Li, H. Fu, and L. Gao. DeepFaceEditing: Deep face generation and editing with disentangled geometry and appearance control. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2021)*, 40(4):90:1–90:15, 2021. 2

[7] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8188–8197, 2020. 1, 2, 3, 5, 11

[8] W. Cong, J. Zhang, L. Niu, L. Liu, Z. Ling, W. Li, and L. Zhang. Dovenet: Deep image harmonization via domain verification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8394–8403, 2020. 1, 3, 6, 8

[9] X. Cun and C.-M. Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing*, 29:4759–4771, 2020. 1

[10] E. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, volume 28, pages 1486–1494, 2015. 2

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3:2672–2680, 2014. 2

[12] W. Grathwohl, R. T. Q. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2018. 2, 5

[13] S. Gu, J. Bao, H. Yang, D. Chen, F. Wen, and L. Yuan. Mask-guided portrait editing with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3436–3445, 2019. 1, 2

[14] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of wasserstein gans. In *NIPS'17 Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, pages 5769–5779, 2017. 2

[15] 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 *Advances in Neural Information Processing Systems*, volume 30, pages 6626–6637, 2017. 6

[16] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 3

[17] Z. Huang, Y. Peng, T. Hibino, C. Zhao, H. Xie, T. Fukusato, and K. Miyata. dualface: Two-stage drawing guidance for freehand portrait sketching. *Computational Visual Media*, 8(1):63–77, 2022. 2

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017. 1, 2

[19] W. Jiang, S. Liu, C. Gao, J. Cao, R. He, J. Feng, and S. Yan. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5194–5202, 2020. 3

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *international conference on learning representations*, 2018. 2

[21] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 6, 7

[22] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2

[23] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *ICLR 2015 : International Conference on Learning Representations 2015*, 2015. 6

[24] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR 2014 : International Conference on Learning Representations (ICLR) 2014*, 2014. 1, 6

[25] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5549–5558, 2020. 1, 2, 5, 7

[26] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. K. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018. 3

[27] J. Lee, E. Kim, Y. Lee, D. Kim, J. Chang, and J. Choo. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5801–5810, 2020. 3, 4

[28] Y. Liu, H. Shi, H. Shen, Y. Si, X. Wang, and T. Mei. A new dataset and boundary-attention semantic segmentation for face parsing. In *AAAI*, pages 11637–11644, 2020. 6, 7

[29] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821, 2017. 2

[30] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. 6

[31] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2337–2346, 2019. 1, 3, 6, 7, 9

[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)*, 2018. 2

[33] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1530–1538, 2015. 2

[34] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-or. Encoding in style: a stylegan encoder for image-to-image translation. In *arxiv:cs.CV*, 2021. 2

[35] Y. Shen, J. Gu, X. Tang, and B. Zhou. Interpreting the latent space of gans for semantic face editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9243–9252, 2020. 2

[36] Y. Shen and B. Zhou. Closed-form factorization of latent semantics in gans. *arXiv preprint arXiv:2007.06600*, 2020. 2

[37] R. Sun, C. Huang, H. Zhu, and L. Ma. Mask-aware photorealistic facial attribute manipulation. *Computational Visual Media*, pages 1–12, 2021. 2

[38] Z. Tan, M. Chai, D. Chen, J. Liao, Q. Chu, L. Yuan, S. Tulyakov, and N. Yu. Michigan: multi-input-conditioned hair image generation for portrait editing. *ACM Transactions on Graphics*, 39(4):95, 2020. 2

[39] Z. Tan, D. Chen, Q. Chu, M. Chai, J. Liao, M. He, L. Yuan, G. Hua, and N. Yu. Efficient semantic image synthesis via class-adaptive normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):1–1, 2021. 6, 7, 9

[40] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Perez, M. Zollhofer, and C. Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6142–6151, 2020. 2

[41] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, and M.-H. Yang. Deep image harmonization. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2799–2807, 2017. 1

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017. 3

[43] M. Wang, G.-Y. Yang, R. Li, R.-Z. Liang, S.-H. Zhang, P. M. Hall, and S.-M. Hu. Example-guided style-consistent image synthesis from semantic labeling. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1495–1504, 2019. 1

[44] 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 1, 4, 6, 7, 8

[45] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 3

[46] C. Xiaodong and P. Chi-Man. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing*, 2020. 3

[47] D. Yang, S. Hong, Y. Jang, T. Zhao, and H. Lee. Diversity-sensitive conditional generative adversarial networks. In *International Conference on Learning Representations*, 2019. 1

[48] G. Yang, X. Huang, Z. Hao, M.-Y. Liu, S. Belongie, and B. Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4541–4550, 2019. 5

[49] H. Zhang, I. J. Goodfellow, D. N. Metaxas, and A. Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363, 2018. 2, 3

[50] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen. Cross-domain correspondence learning for exemplar-based image translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5143–5153, 2020. 3, 4

[51] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[52] J. Zhu, Y. Shen, D. Zhao, and B. Zhou. In-domain gan inversion for real image editing. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 2

[53] J.-Y. Zhu, P. Krahenbuhl, E. Shechtman, and A. A. Efros. Learning a discriminative model for the perception of realism in composite images. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3943–3951, 2015. 1, 2, 3, 6

[54] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 2017. 1, 2, 3

[55] P. Zhu, R. Abdal, Y. Qin, and P. Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5104–5113, 2020. 1, 2, 4, 5, 6, 7, 8

[56] Z. Zhu, Z. Xu, A. You, and X. Bai. Semantically multi-modal image synthesis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5467–5476, 2020. 1, 3, 6, 7, 8, 9, 11