Computational Visual Media
https://doi.org/10.1007/s41095-022-0284-6 Vol. 9, No. 2, June 2023, 351–366

Research Article

# Towards harmonized regional style transfer and manipulation for facial images

**Cong Wang**[1], **Fan Tang**[2] (✉), **Yong Zhang**[3], **Tieru Wu**[1,2] (✉), **and Weiming Dong**[4]

**Abstract** Regional facial image synthesis conditioned on a semantic mask has achieved great attention in the field of computational visual media. However, the appearances of different regions may be inconsistent with each other after performing 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. The key part of our work is a multi-region style attention module. It adapts multiple regional style embeddings from a reference image to a target image, to generate a harmonious result. We also propose style mapping networks for multi-modal style synthesis. We further employ an invertible flow model which can serve as mapping network to fine-tune the style code by inverting the code to latent space. Experiments on three widely used face datasets were used to evaluate our model by transferring regional facial appearance between datasets. The results show that our model can reliably perform style transfer and multi-modal manipulation, generating output comparable to the state of the art.

**Keywords** face manipulation; style transfer; generative models; facial harmonization

## 1 Introduction

Semantic image synthesis [1–7], that aims to generate realistic natural images from semantic labels, has been an active research topic in the past few years. Based on the different ways of involving new styles in synthesis, there are two types of mainstream methods to generate diverse images: injecting random noise [1, 2, 8], or transfer from reference images [7, 9–11]. Researchers have made great progress in both fields. Choi et al. [12] employ a style extraction net for facial style transfer and a mapping network adapted from StyleGAN [13, 14] to transform Gaussian noise into style codes.

SPADE [5] adopts the idea of VAE [15] to encode the image style and enables both tasks. However, SPADE is just able to transfer facial style globally, thus limiting its practical usage. Recent works [6, 7, 9] extract style codes for all semantic components separately, enabling regional style transfer and manipulation (R-ST&M) for facial images.

R-ST&M provides flexible facial image editing. However, new problems arise at the same time: regional appearance editing (i.e., transfer or manipulation) can lead to different regions having mutually inconsistent appearance. For example, when transferring skin style from a target facial image to another face captured under different lighting conditions, the new skin style generated by methods such as SEAN [7] may mismatch other regions in the target image (see Fig. 1). Similar problems occur in the field of image composition [16–19]. To the best of our knowledge, no prior work focuses on style consistency and harmony for R-ST&M.

In this paper, we propose a framework which takes style 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 the SEAN encoder to extract style codes with rich style information, since low-level features are important for reconstruction [1, 2, 20].

[1] School of Mathematics, Jilin University, Changchun 130012, China. E-mail: C. Wang, cwang16@mails.jlu.edu.cn; T. Wu, wutr@jlu.edu.cn (✉).

[2] School of Artificial Intelligence, Jilin University, Changchun 130012, China. E-mail: tanhgfan@jlu.edu.cn (✉).

[3] AI Lab, Tencent, Shenzhen 518054, China. E-mail: zhangyong201303@gmail.com.

[4] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. E-mail: wmlake@gmail.com.

Manuscript received: 2022-02-18; accepted: 2022-03-16

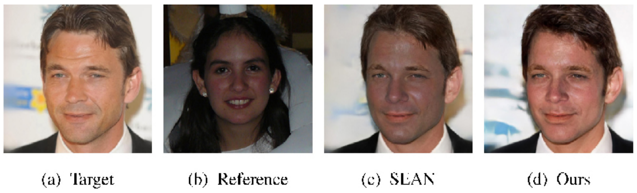


**Fig. 1** An example of skin transfer. The R-ST&M method can modify the transferred skin to match the global lighting of the target image. However, without considering the relationships between different regions, the synthesised region in (c) is not in harmony with other regions.

In order to make the synthesized image with transferred style look plausible, we employ a multi-region style attention (MRSA) module where the relationship 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 the style mapping networks is inherited from StarGAN-v2 [12]. However, instead of training the mapping networks by adversarial loss of fake and real images, we calculate the adversarial loss on the style embedding space. The multi-scale encoder outputs multi-region style spaces with relationships between different regions; building the mapping networks directly from distributions can generate reliable regional styles. Furthermore, we train a continuous normalizing flow (CNF) [21, 22] which can invert the 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 the "harmony" of synthetic images, we use a binary classification network to distinguish natural photographs from composite ones, following Ref. [19]. The proposed approaches are used as a basis for two facial editing applications. Figure 2 shows an example of our harmonized color editing application.

To summarize, our main contributions are as follows:

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

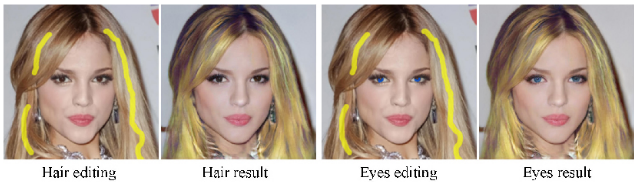


**Fig. 2** Examples of color editing using our model.

## 2 Related work

### 2.1 Facial image manipulation with GANs

Generative adversarial nets (GANs) [1, 13, 14, 23–25] 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 [23] suffers from mode collapse. Many works, e.g., Refs. [24, 26–29], have proposed improvements to the generation quality of GANs.

One of the most important applications of GANs is to generate photo-realistic human face images. PGGAN [25] grows both the generator and discriminator progressively, allowing users to produce high-resolution and high-quality face images. StyleGAN [13] and StyleGAN2 [14] introduce a novel generator architecture borrowed from the style transfer literature, enabling indistinguishable face image generation. In the field of facial image editing, significant progress has been made using powerful GANs. FaceShop [30] presents a novel system for face image manipulation, providing both geometry and color constraints as user-drawn strokes. DeepFaceEditing [31] is a structured disentanglement framework designed for face images to support face manipulation with disentangled control of geometry and appearance. MichiGAN [32] explicitly disentangles hair into four orthogonal attributes and designs a corresponding condition module to process user inputs for each attribute. DualFace [33] proposes a two-stage guidance system to help users produce detailed portrait sketches with data-driven global guidance and GAN-based local guidance. InterFaceGAN [34] explores disentanglement of various semantic attributes and edits several attributes using linear editing paths. SeFa [35] proposes a general closed-form factorization method for latent semantic discovery. StyleRig [36] provides face rig-like control over a pretrained StyleGAN. StyleFlow [37] utilizes normalizing flows [38] for editing facial attributes interactively with StyleGAN.

Recent works [20, 39] learn to encode facial images for StyleGAN inversion and facilitate various image editing tasks. MaskGAN [10] 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. [40] use partial dilated layers to modify a few pixels in learned feature maps and realize mask-aware continuous facial attribute manipulation. Gu et al. [9] propose an end-to-end framework to learn conditional GANs guided by semantic masks, enabling regional facial style transfer. SEAN [7] proposes semantic region-adaptive normalization for GANs conditioned on segmentation masks; it 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 distributions.

### 2.2 Self-attention

Self-attention was first proposed in the natural language processing literature in the form of Transformer [41]. Computer vision researchers then extended the idea to video classification [42] and image generation [29]. Recent works generalize self-attention to extract correspondences between source and reference image for semantic style transfer [43, 44] and makeup transfer [45]. However, the self-attention mechanism computes the correspondence spatially, making it time-consuming and inefficient. Instead, our style attention, inspired by the above works, computes the correlation between semantic regional style vectors, which ensures its computational efficiency.

### 2.3 Multi-modal image synthesis

BicycleGAN [2] 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. Refs. [46, 47] extend the multi-modal idea to unsupervised image-to-image translation, to generate diverse images. SPADE [5] uses the same idea to encode image style for semantic image synthesis. GroupDNet [6] extends SPADE by using KL loss for all semantic labels, thus enabling regional multi-modal synthesis. Recently, StarGAN-v2 [12] was proposed; it learns a mapping network to achieve diversity. Our style mapping model follows StarGAN-v2, but has a different training strategy more suited to our framework.

### 2.4 Deep image harmonization

Deep convolutional models have achieved significant success for image harmonization in recent years. Zhu et al. [19] train a binary classifier to guide color adjustment for composite images. Then, an end-to-end deep CNN model captures both the context and semantic information during harmonization. Cun and Pun [48] use a spatially-separated attention module in order to learn separate foreground and background feature maps. DoveNet [17] translates the foreground domain to the background domain by using a domain verification discriminator, with impressive results. Since facial regional style transfer may lead to disharmony, we propose a multi-region style attention module to adjust the transferred regional style to other regions. The proposed module is incorporated into the style transfer process, allowing harmonious images to be directly synthesized without a subsequent image harmonization process. We use the method of Zhu et al. [19] to evaluate the degree of harmony of an image.

## 3 Architecture

Figure 3 shows the framework of our proposed multi-region style transfer and multi-modal synthesis method. The inputs are a segmented target image $x_{\mathrm{t}}$ that the user wishes to edit and a reference style. The reference style can either be generated from a segmented reference *style image* $x_{\mathrm{r}}$ for style transfer, or directly sampled from a normal Gaussian distribution for manipulation. In this section, we start by introducing the regional feature encoding, including a multi-scale encoder for input images and regional style mapping (RSM) subnets for multi-

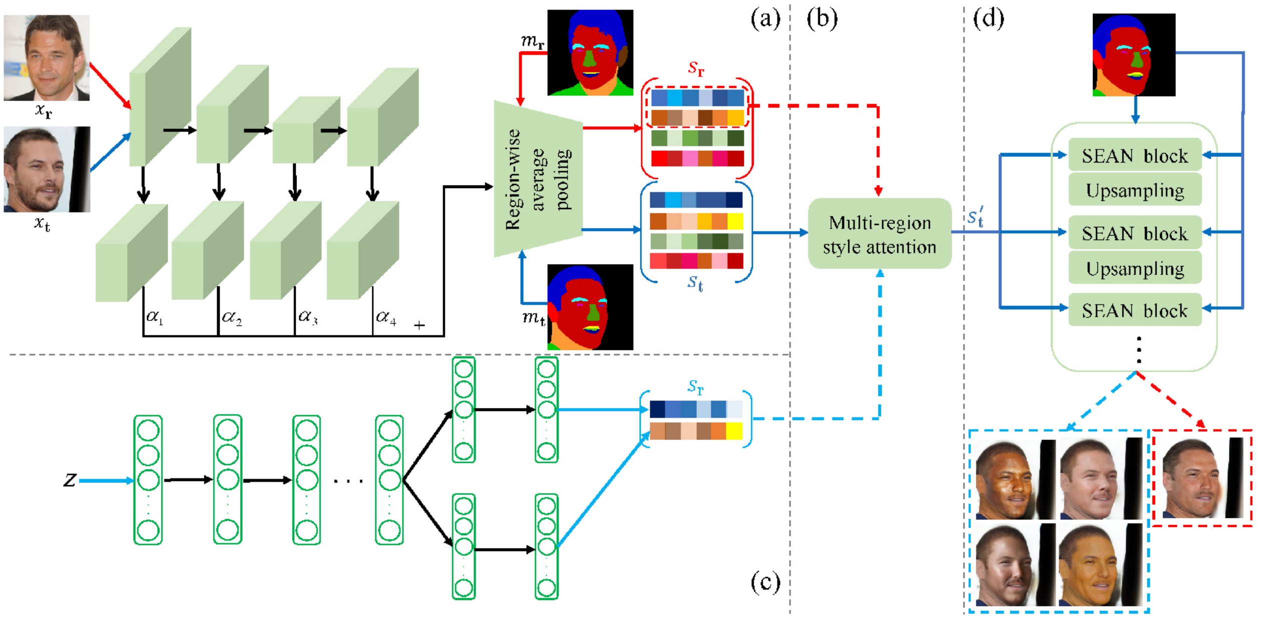


**Fig. 3** Framework. (a) Multi-level feature fusion part of the encoder. (b) Multi-region style attention module. (c) An example style mapping network, in which the mapping network generates styles for skin and nose simultaneously; multi-modal results are shown in (d). (d) SEAN generator with results of (a) and (c).

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

### *3.1.1 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 encoder, low-level features from the shallow layers are important for image reconstruction. Therefore, we compute a weighted sum of feature maps from all layers in the encoder, as shown in Fig. 3(a). Concretely, we first re-scale the feature maps to a uniform resolution to get new features $\{F_i\}_{i=1}^K$, where $K$ is the number of shallow layers. Then, a set of learnable parameters $\{a_i\}_{i=1}^K$ is fed into a softmax function for normalization:

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

We then get the final multi-scale style feature map using

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

The learned weights $\{a_i\}_{i=1}^K$ indicate the proportion of each scale to use when compositing the feature map $F$. Given an input target image $x_t$ and a reference image $x_r$ with segmentation masks $m_t$ and $m_r$ respectively, we employ a region-wise average pooling layer [4, 7] to transform $F_t$ and $F_r$ to initial style vectors $s_t$ and $s_r$ respectively.

### *3.1.2 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 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 networks as the area of these regions is large enough. As the correlations between some regions, such as nose and hair, are weak, we use two networks to model them separately. Figure 3(c) shows an example of the mapping sub-network for modeling skin and nose. Given a latent code $z$ sampled from the Gaussian distribution, a random reference style can be generated using the mapping network $\mathcal{M}$:

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

In our method, related regions such as skin and nose or the two eyes share a common mapping network. More details of training the RSM are given in Section 3.4. Next, we feed $s_{\mathrm{r}}$ into the MRSA module and generator $\mathcal{G}$.

### 3.2 Multi-region style attention

If the global appearances (i.e., lighting conditions) of $x_{\mathrm{t}}$ and $x_{\mathrm{r}}$ are quite different, regional style transfer results may be disharmonious. However, users would prefer to get a harmonious image directly without a need for a subsequent image harmonization process. To this end, we propose a multi-region style attention (MRSA) module to learn transferred styles. Figure 4 illustrates the workflow of the MRSA module. Unlike the attention modules in Refs. [43, 44] that extract spatial correspondences in pixel space, our MRSA module computes relationships between regional semantic styles. In order to correct the styles of different regions, we first concatenate the target components in $s_{\mathrm{r}}$ with the remaining components in $s_{\mathrm{t}}$ to form a new $s'_{\mathrm{r}}$. Then we map the style vectors using $Q = \mathcal{W}_q(s'_{\mathrm{r}})$, $K = \mathcal{W}_k(s_{\mathrm{t}})$, and $V = \mathcal{W}_v(s_{\mathrm{t}})$, where $\mathcal{W}_q$, $\mathcal{W}_k$, and $\mathcal{W}_v$ are linear mappings. Next, an attention matrix can be computed by $QK^{\mathrm{T}}$ followed by a softmax function within each row:

$$M = \mathrm{softmax}(QK^{\mathrm{T}}) \tag{4}$$

After computing the attention matrix $M$, we can get the style correction: $s_{\mathrm{c}} = MV$. Finally, the target style can be computed using

$$s'_{\mathrm{t}} = s'_{\mathrm{r}} + \alpha s_{\mathrm{c}} \tag{5}$$

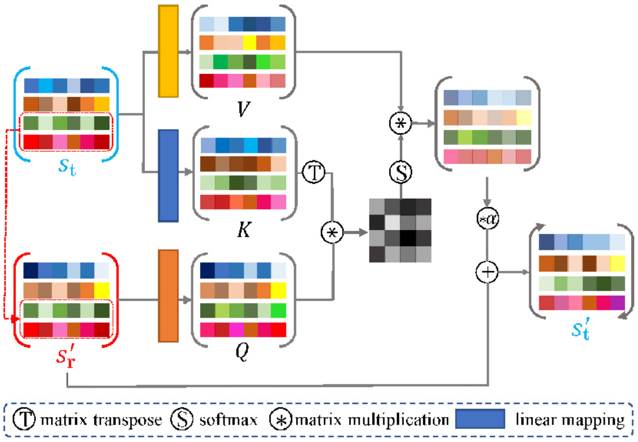


**Fig. 4** Multi-region style attention module. $s_{\mathrm{t}}$: target style vectors of all regions. $s'_{\mathrm{r}}$: concatenation of styles from target regions in the reference and styles from the remaining regions in the target. Linear projection metrics $\mathcal{W}_v$, $\mathcal{W}_q$, and $\mathcal{W}_k$ are used 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 applying $MV$ to $s'_{\mathrm{r}}$.

### 3.3 Decoder

Given the style vectors generated by MSRA, the SEAN generator [7] is used as a decoder by feeding them into a semantic region-adaptive normalization (SEAN) module. In SEAN, a target mask and a style map generated by broadcasting style vectors to the corresponding regions are used to modulate the activation from the previous layer. The decoder employs several SEAN blocks with upsampling layers and synthesizes images progressively.

### 3.4 Model training

The encoder–decoder part of our model is similar to that in SPADE and SEAN. We use three loss functions described in SPADE and SEAN to train it: adversarial loss, feature matching loss, and perceptual loss. During training, if we use $s_{\mathrm{r}}$ extracted from a reference different from the source image $x_{\mathrm{s}}$, this results in unsupervised training as there is no ground truth for the new image. To tackle this problem, we set $x_{\mathrm{r}}$ equal to $x_{\mathrm{s}}$ for training. We tried a training strategy mixing supervised and unsupervised training, but it failed to generate realistic images. The reason we suppose is that the unsupervised result disturbs the supervised training pace.

As for style mapping networks $\{\mathcal{M}_j\}_{j=1}^{M}$, we turn to the adversarial loss imposed on $s_{\mathrm{s}}$ and $s_{\mathrm{r}}$ 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 $\{\mathcal{D}_j\}_{j=1}^{M}$ is employed, with adversarial objectives as Eq. (6):

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

A similar style mapping network was proposed in StarGAN-v2 [12], 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, as our encoder–decoder is trained in a supervised way, and the encoder learns expressive style information. It is more effective to learn the distributions of encoded styles directly.

### 3.5 Style random fine-tuning

If users are not satisfied with the current style, we also provide a method for style fine-tuning that users can use to 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 move the style code $s_{\rm t}$ 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 an invertible continuous normalizing flow (CNF) [21, 22], as utilized in Ref. [49] for point cloud generation and Ref. [37] for facial attribute manipulation.

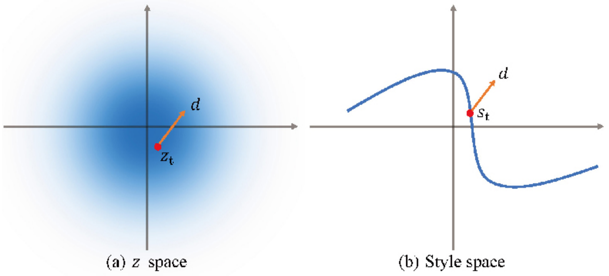
Specifically, we first use the ODE below to get $z_{\rm t}$ corresponding to $s_{\rm t}$:

$$z_{\rm t} = s(l_0) + \int_{l_0}^{l_1} f(s(l), l)\mathrm{d}l \tag{7}$$

where $s(l_0) = s_{\rm t}$ and $l$ denotes time. Then we fine-tune $z_{\rm t}$ by $z'_{\rm t} = z_{\rm t} + \eta d$, where $\eta$ is the step size. Then a reverse-time ODE is employed to recover a modified meaningful style code:

$$s'_{\rm t} = z'(l_1) + \int_{l_1}^{l_0} f(z'(l), l)\mathrm{d}l \tag{8}$$

where $z'(l_1) = z_t$.



(a) $z$ space (b) Style space

**Fig. 5** Style fine-tuning in $z$ space and style space. Manipulation in style space leads to an out-of-manifold result.

# 4 Experiments

## 4.1 Experimental setting

### *4.1.1 Datasets*

We used three face datasets to evaluate our framework.

CelebAMASK-HQ [10] consists of 30,000 face images with segmentation masks. Each image is annotated with a semantic mask of 19 semantic categories in total. We used the first 28,000 images for training and the remainder for evaluation.

FFHQ [13] contains 70,000 high-quality images. We utilized a deeplab-v3 model [50] trained on CelebAMASK-HQ to parse the facial semantics. We employed the first 2000 images for evaluation.

LaPa [51] 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 discarded low-resolution images in the dataset. The final training and test sets contained 19,770 and 1930 faces respectively.

### *4.1.2 Metrics*

We used several common metrics to evaluate our framework and competing methods. Specifically, FID [53] computes the distance between the distributions of synthesized images and of real images, and is used to evaluate the quality of synthesized results. We also adopt PSNR, SSIM, and LPIPS [54] to assess the similarity between synthesized and ground-truth images in the face reconstruction task. In order to evaluate the performance of our model for regional multi-modal synthesis with random styles, we utilized mean class-specific diversity (mCSD) and mean other-class diversity (mOCD) [6]. For a fixed semantic region, mCSD assesses the generated diversity of the region while mOCD assesses the diversity of the remaining regions. High mCSD and low mOCD indicate good performance for a fixed region.

In addition to the above metrics, we employ *harmony score* (HS) to measure the degree of harmony between the transferred region and the remainder for regional style transfer. The idea of harmony score follows that of realism score [19] 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. Real images are set as positive samples and unrealistic composite images are set as negative samples. We used HAdobe5k [17] to train the classification network and concatenated an image and its corresponding foreground mask to form each individual input.

### *4.1.3 Competing methods*

We compared our method to five leading semantic image synthesis models: pix2pixHD [4], SPADE [5], GroupDNet [6], SEAN [7], and CLADE [52]. 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. Thus, pix2pixHD is suitable for regional style transfer. SPADE uses an encoder and generator to form a VAE [15] and a new normalization for the generator, enabling global style transfer and multi-modal synthesis conditioned on the semantic mask. GroupDNet extends the idea of SPADE by encoding different semantic regions separately and leveraging a group decreasing generator. GroupDNet can be used for regional style transfer and multi-modal synthesis. SEAN employs similar structures to the pix2pixHD encoder and SPADE generator; SEAN normalization improves the generation quality significantly. CLADE improves SPADE based on the observation that its modulation parameters benefit more from semantic-awareness rather than spatial-adaptivity. Tan et al. [52] also introduce CLADE-ICPE, where intra-class positional map encoding improves spatial-adaptivity.

### 4.2 Implementation details

We use the TTUR [53] 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 [55] to the encoder. Moreover, we use the ADAM solver [56] 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.

### 4.3 Global reconstruction

We first evaluated the effectiveness of the proposed multi-region style control and manipulation network in the image reconstruction task, namely transferring the image's own style to itself. Only one image was employed as input. Visual comparisons are shown in Fig. 6. Overall, pix2pixHD and groupDNet fail to maintain skin color well. Compared to SEAN, our method can reconstruct more facial details of the input, e.g., the wrinkles on the left of the woman's face and the left eye of the man under the sunglasses. A quantitative evaluation is provided in Table 1. Our model outperforms the other state-of-the-art methods on all datasets. It is worth noting that although MRSA is designed for style transfer and manipulation, it exhibits the best reconstruction quality (lowest FID) on all datasets.

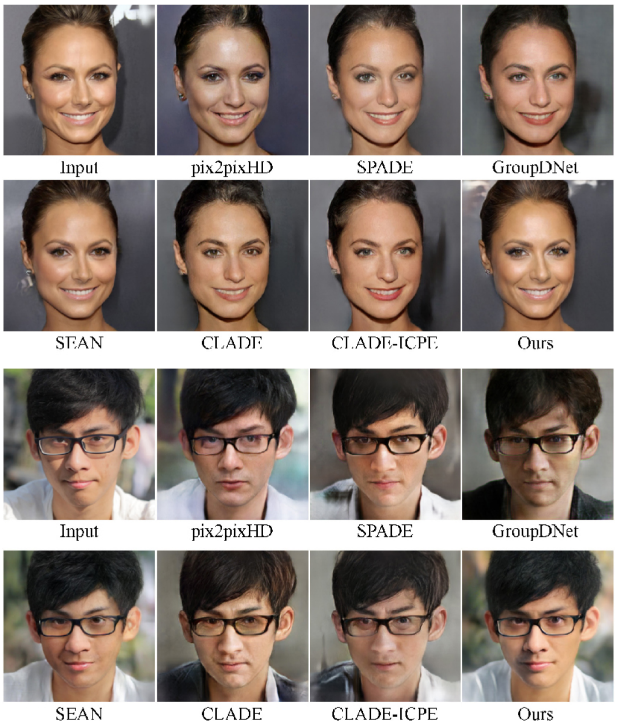


**Fig. 6** Image reconstruction results.

**Table 1** Facial image reconstruction results. Higher PSNR and SSIM are better, but lower LPIPS and FID

| | CelebAMASK-HQ | | | | FFHQ | | | | LaPa | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | LPIPS↓ | FID↓ | PSNR↑ | SSIM↑ | LPIPS↓ | FID↓ | PSNR↑ | SSIM↑ | LPIPS↓ | FID↓ |
| pix2pixHD [4] | 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 [5] | 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 [6] | 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 [7] | 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 [52] | 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 [52] | 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** |

## 4.4 Regional style transfer

### *4.4.1 Initial evaluation*

We further evaluated the effectiveness of the proposed approach in the regional style transfer task. One target image and one reference image were employed as inputs. We split all test datasets into two parts: one half as target images and the other half as reference images. SPADE was not considered for comparison since it does not support region style transfer. Figure 7(a) shows some results. We used FID measured on the whole image as the metric in two transfer tasks: skin (with nose) transfer, and hair transfer. Quantitative results are provided 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.

### *4.4.2 Cross-dataset transfer*

Most images in CelebAMask-HQ [10] and FFHQ [13] were captured under good lighting, so regional style transfer between these images can hardly lead to disharmonious results. However, LaPa [51] consists of facial images with abundant variations in lighting conditions. Therefore, we transferred skin (with nose) and hair of facial images from the test sets of CelebAMask-HQ, FFHQ, and LaPa to the test set of LaPa separately, and calculated FID and HS of synthesized faces, with results shown in Table 3. Our method and SEAN [7] perform much better than pix2pixHD [4] and GroupDNet [6] 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 remaining regions. Our method exhibits obviously higher harmony scores than SEAN, showing the effectiveness of MRSA. However, pix2pixHD and GroupDNet reach higher harmony scores 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 severe changes occur to the remaining regions which should keep their appearance. In summary, our model provides the best trade-off between image quality and degree of harmony.

**Table 2** FID↓ results for 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** |

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

### *4.4.3 User study*

We conducted user studies to further compare the visual performance of our method to the other selected methods. Firstly, we showed the participants each target-reference pair and told them which region in the target image we wanted to edit. Then we showed them four results, one from our method and the others from pix2pixHD, GroupDNet, and SEAN. Each subject was assigned with 30 group results. We received 59 responses, among which 47 responses were valid. A total of 1410 votes were obtained. Our model had 627 (44.45%) votes, SEAN had 410

**Table 3** PSNR↓ and HS↑ results for cross-dataset regional style transfer. Although pix2pixHD and GroupDNet have better HS than our method, these two methods achieve harmony at the expense of severely modifying other regions (see Fig. 7), contradicting the goals of regional style transfer. "Ours w/o softmax" means our method without softmax normalization and MRSA; "Ours w/o SA" means our method without MRSA

| | CelebAMASK-HQ→LaPa | | FFHQ→LaPa | | LaPa→LaPa | |
|---|---|---|---|---|---|---|
| | Skin | Hair | Skin | Hair | Skin | Hair |
| pix2pixHD [4] | 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 [6] | 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 [7] | 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 |

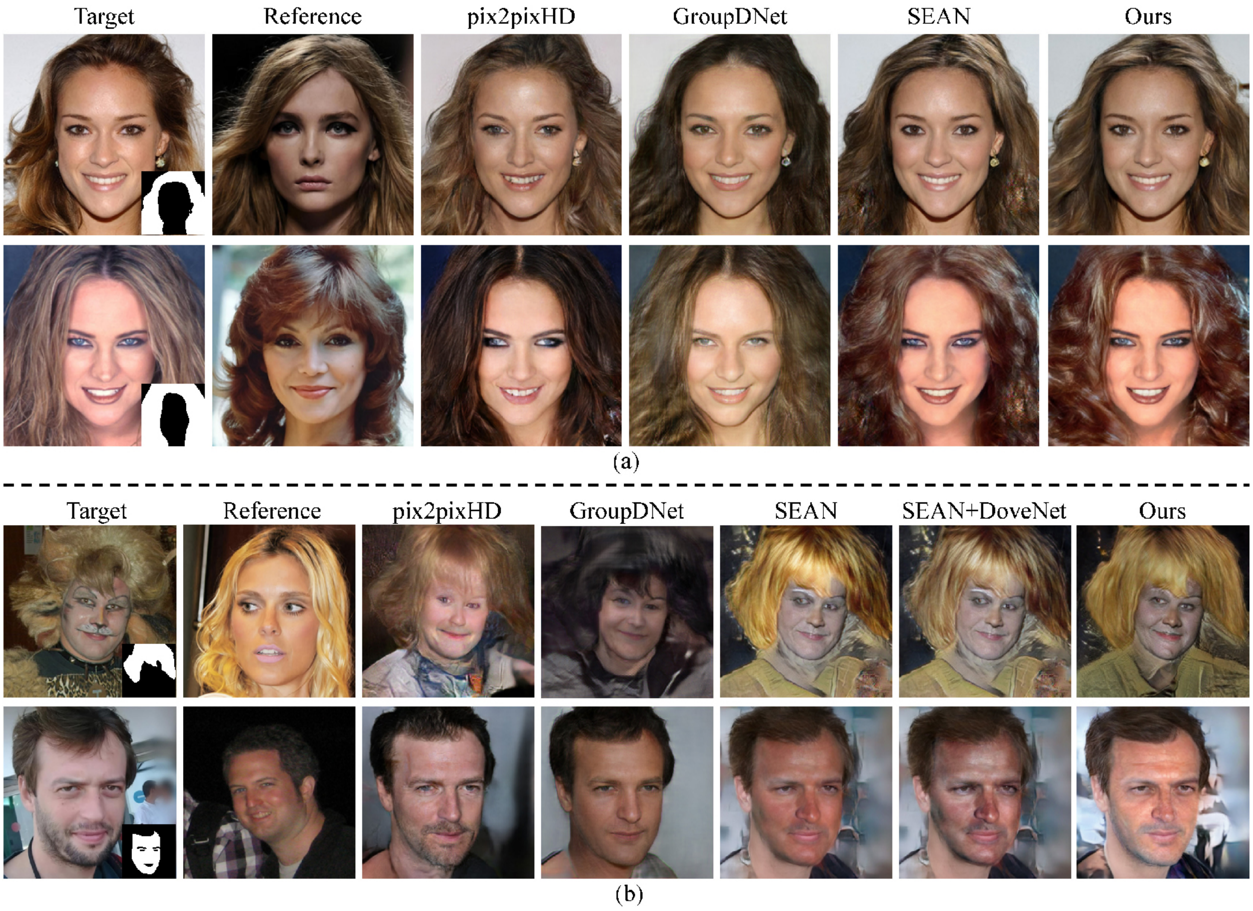


**Fig. 7** Regional style transfer results (segmentation mask shown as inset). (a) Target and reference images from the same dataset. (b) Cross dataset transfer. Our model generates more harmonious results than DoveNet [17] performed on the outputs of SEAN [7].

(29.07%) votes, GroupDNet had 265 (18.78%) votes, and Pix2Pix had 108 (7.66%) votes.

### 4.5 Comparison to DoveNet

Since our work concerns image harmonization, we utilized a recent deep harmonization model, DoveNet [17], to harmonize the outputs of SEAN for comparison. As shown in Table 2, DoveNet has an adverse effect on image synthesis quality: DoveNet gets higher FID scores than SEAN. Consider the image harmonization shown in Fig. 7(b). DoveNet indeed harmonizes 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 that of the reference during harmonization. More importantly, separate harmonization will take extra time.

### 4.6 Regional multi-modal manipulation

We next evaluated the effectiveness of the proposed approach in the regional multi-modal manipulation task. One target image and one vector sampled from a normal Gaussian distribution were employed as inputs. SPADE [5], GroupDNet [6], CLADE [52], and CLADE-ICPE [52] were selected for comparison. SPADE and CLADE are intended for global multi-modal synthesis while GroupDNet is intended for regional multi-modal synthesis. Figure 8 shows the manipulation of skin. GroupDNet affects hair more significantly than our method in skin multi-modal synthesis. We further conducted qualitative experiments on manipulation of skin and hair regions. Table 4 reports FID, mCSD, and mOCD calculated for the three different datasets. In terms of image quality, our method outperforms other methods by a large margin on all datasets. In terms of diversity (mCSD), SPADE, CLADE, CLADE-ICPE, and our method perform at the same level, but SPADE, CLADE, and CLADE-ICPE fail to preserve the appearance of the other regions (higher mOCD), as they are designed for global synthesis. For skin multi-

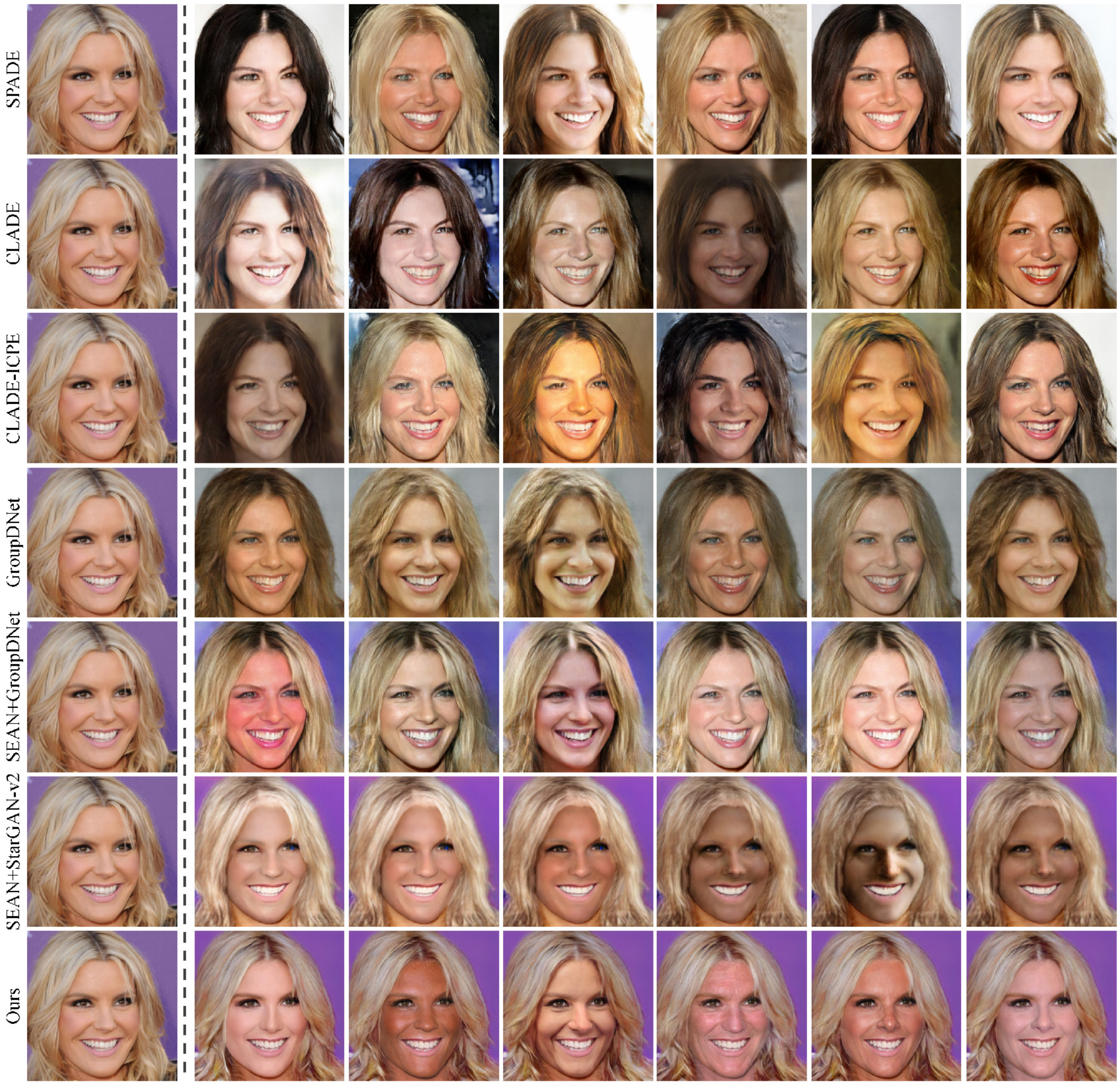


**Fig. 8** Skin multi-modal synthesis.

modal synthesis, our method exhibits higher mCSD and lower mOCD than GroupDNet, even though it extends SPADE to regional style synthesis: our method is better at maintaining the appearance of the remaining regions while achieving high color and texture diversity in skin synthesis. For hair multi-modal synthesis, GroupDNet generates facial images with low diversity: mCSD and mOCD of GroupDNet are both close to zero. In all multi-modal synthesis experiments, we manipulated each image using 10 random styles.

### 4.7 Fine-tuning style

First, we evaluated 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 more slowly than a GAN model: our model with CNF takes 0.2 s to generate a style code while our GAN mapping network takes $6 \times 10^{-4}$ s on a single RTX3090.

Secondly, we validate the analysis in Section 3.5 by

**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 |

showing examples of style random fine-tuning both in $z$ space and style space. Figure 9 demonstrates that skin tuning results using $z$ space are much better than those in style space. Small tuning steps in style space hardly affect the style of the target image while larger steps fail to generate clear results. As the skin style code consists of two parts, tuning the skin style code will lead to deviation from the manifold. Thus,

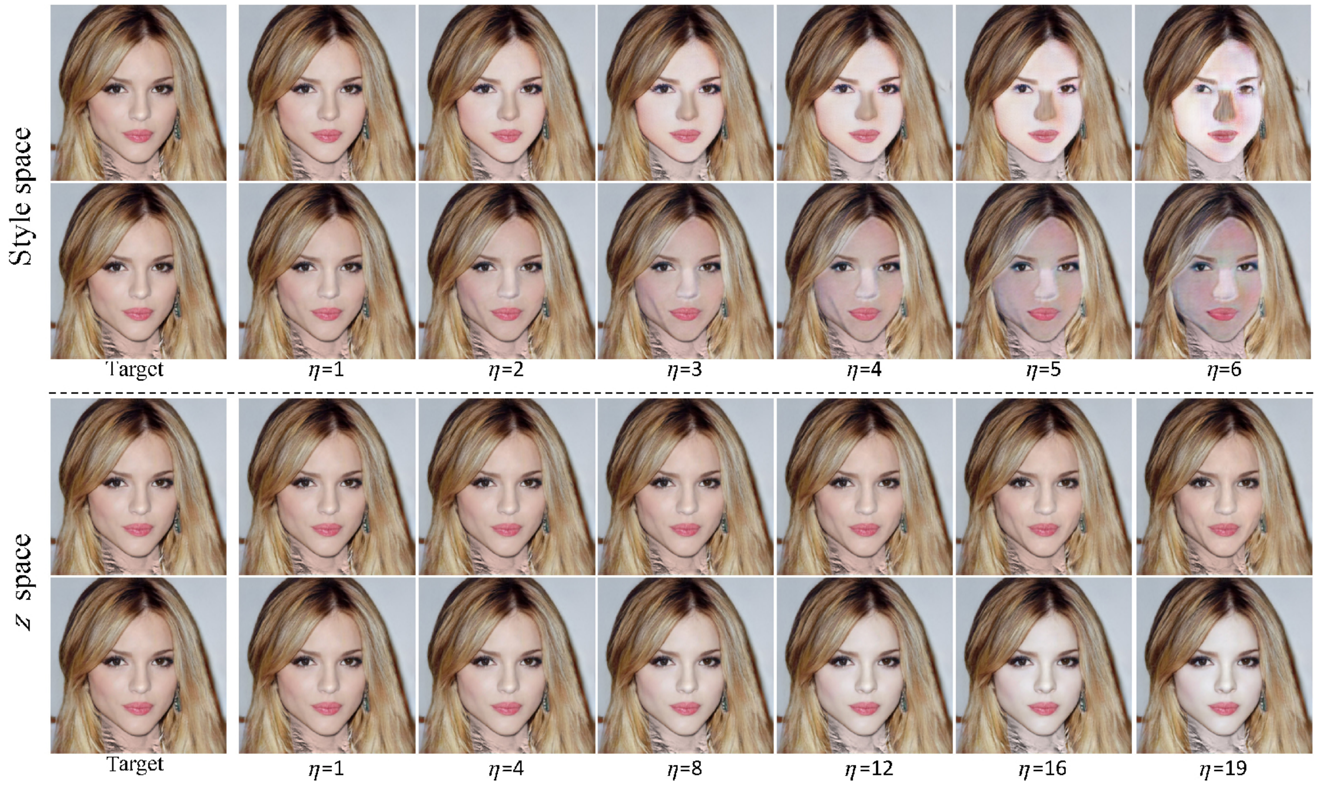


**Fig. 9** Two examples of style random fine-tuning for skin in both $z$ and style spaces. Tuning in $z$ space can achieve the goal of convincingly fine-tuning style, such as gradually changing skin color and adding wrinkles.

the nose presents a different style from the facial skin even if the step size is small. A similar conclusion holds for hair style fine-tuning.

### 4.8 Ablation study

#### *4.8.1 MRSA*

To validate the effectiveness of the softmax function and MRSA module in the encoder, we conducted ablation experiments omitting them from the framework. Results of cross-dataset regional style transfer in Table 3 indicate that softmax normalization and MRSA improve the image quality and degree harmony degree, respectively.

#### *4.8.2 RSM*

To validate the effectiveness of the RSM module, we conduct ablation experiments by combining the encoders and training strategies used in GroupDNet [6], StarGAN-v2 [12], and ours with the SEAN generator. This gave three variants for comparison: “SEAN+GroupDNet”, “SEAN+StarGAN-v2”, and “SEAN+Our RSM”. Table 5 provides FID for skin and hair multi-modal synthesis for all datasets. The “SEAN+Our RSM” method provides much better terms of image quality than the two variants. Our RSM uses a similar mapping network to StarGAN-v2 but a different training strategy. If we used the training strategy from StarGAN-v2, the generator would be trained in an unsupervised way. However, our encoder–decoder is trained in a supervised manner and the mapping network shares the same generator with the encoder; different objectives would misguide the generator. A visual comparison of the results can be found in Fig. 8.

**Table 5** FID↓ for RSM ablation study

| | | SEAN+<br>GroupDNet | SEAN+<br>StarGAN-v2 | SEAN+<br>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** |

## 5 Applications

Our framework can underpin various applications of facial image synthesis. Sections 4.4 and 4.6 demonstrate the effectiveness of regional style transfer across facial images and multi-modal synthesis with random styles, respectively. Two other applications are interactive face shape editing and face color editing.

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

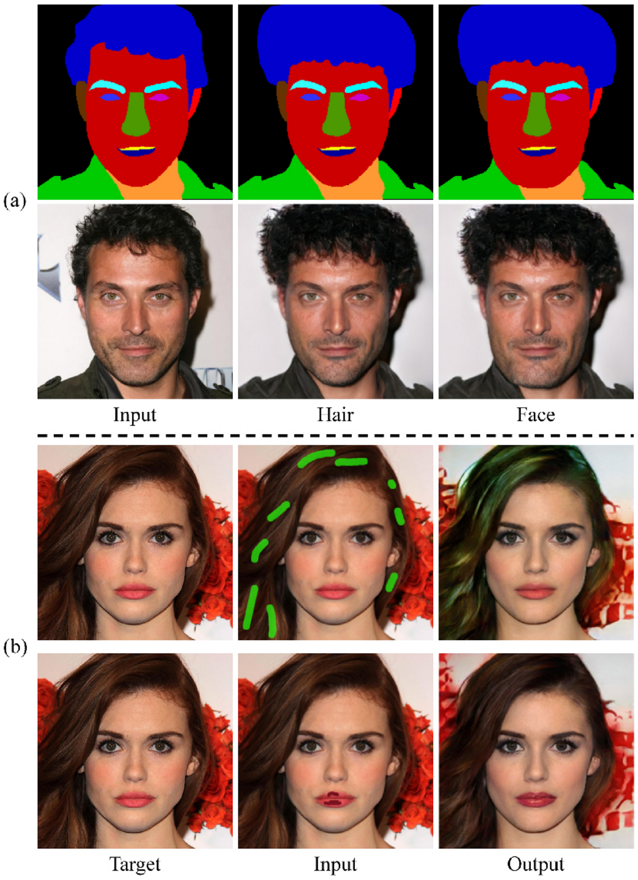


**Fig. 10** Applications: (a) shape editing, and (b) color editing (hair/lips).

## 6 Conclusions

In this paper, we focus on the harmonized region style editing for facial images. The proposed framework follows the encoding–fusion–decoding pattern. 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 used for harmonious regional style transfer; it is especially effective when the target and reference face images have different lighting conditions. For 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 regions are still only weakly controllable. Regional style transfer is the only way to provide strong control information. Our model, SPADE, and GroupDNet are all powerless to randomly synthesize regions with specific appearance. Resolving this problem remains as our future work.

## Acknowledgements

This work was partly supported by the National Key R&D Program of China (No. 2020YFA0714100) and the National Natural Science Foundation of China (Nos. 61872162, 62102162, 61832016, U20B2070).

## Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

## Electronic Supplementary Material

Supplementary material is available in the online version of this article at `https://doi.org/10.1007/s41095-022-0284-6`.

## References

[1] Isola, P.; Zhu, J. -Y.; Zhou, T.; Efros, A. A. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5967–5976, 2017.

[2] Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; Shechtman, E. Toward multimodal image-to-image translation. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 465–476, 2017.

[3] Chen, Q.; Koltun, V. Photographic image synthesis with cascaded refinement networks. In: Proceedings of the IEEE International Conference on Computer Vision, 1520–1529, 2017.

[4] Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8798–8807, 2018.

[5] Park, T.; Liu, M.-Y.; Wang, T.-C.; Zhu, J.-Y. Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2332–2341, 2019.

[6] Zhu, Z.; Xu, Z.; You, A.; Bai, X. Semantically multi-modal image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5466–5475, 2020.

[7] Zhu, P.; Abdal, R.; Qin, Y.; Wonka, P. SEAN: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5103–5112, 2020.

[8] Yang, D.; Hong, S.; Jang, Y.; Zhao, T.; Lee, H. Diversity sensitive conditional generative adversarial networks. In: Proceedings of the International Conference on Learning Representations, 2019.

[9] Gu, S.; Bao, J.; Yang, H.; Chen, D.; Wen, F.; Yuan, L. Mask-guided portrait editing with conditional GANs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3431–3440, 2019.

[10] Lee, C.-H.; Liu, Z.; Wu, L.; Luo, P. MaskGAN: Towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5549–5558, 2020.

[11] Wang, M.; Yang, G.-Y.; Li, R.; Liang, R.-Z.; Zhang, S.-H.; Hall, P. M.; Hu, S.-M. Example-guided style-consistent image synthesis from semantic labeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1495–1504, 2019.

[12] Choi, Y.; Uh, Y.; Yoo, J.; Ha, J. W. StarGAN v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8185–8194, 2020.

[13] Karras, T.; Laine, S.; Aila, T. M. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4396–4405, 2019.

[14] Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of StyleGAN. In: Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8107–8116, 2020.

[15] Kingma, D. P.; Welling, M. Auto-encoding variational bayes. In: Proceedings of the International Conference on Learning Representations, 2014.

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

[17] Cong, W. Y.; Zhang, J. F.; Niu, L.; Liu, L.; Ling, Z. X.; Li, W. Y.; Zhang, L. DoveNet: Deep image harmonization via domain verification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8391–8400, 2020.

[18] Tsai, Y.-H.; Shen, X.; Lin, Z.; Sunkavalli, K.; Lu, X.; Yang, M.-H. Deep image harmonization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2799–2807, 2017.

[19] Zhu, J.-Y.; Krahenbuhl, P.; Shechtman, E.; Efros, A. A. Learning a discriminative model for the perception of realism in composite images. In: Proceedings of the IEEE International Conference on Computer Vision, 3943–3951, 2015.

[20] Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; Cohen-Or, D. Encoding in style: A StyleGAN encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2287–2296, 2021.

[21] Chen, R. T. Q.; Rubanova, Y.; Bettencourt, J.; Duvenaud, D. Neural ordinary differential equations. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 6572–6583, 2018.

[22] Grathwohl, W.; Chen, R. T. Q.; Bettencourt, J.; Sutskever, I.; Duvenaud, D. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In: Proceedings of the International Conference on Learning Representations, 2018.

[23] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, Vol. 3, 2672–2680, 2014.

[24] Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, Vol. 70, 214–223, 2017.

[25] Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. In: Proceedings of the International Conference on Learning Representations, 2018.

[26] Denton, E.; Chintala, S.; Szlam, A.; Fergus, R. Deep generative image models using a Laplacian pyramid of adversarial networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, Vol. 1, 1486–1494, 2015.

[27] Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A. Improved training of wasserstein GANs. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 5769–5779, 2017.

[28] Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; Smolley, S. P. Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, 2813–2821, 2017.

[29] Zhang, H.; Goodfellow, I. J.; Metaxas, D. N.; Odena, A. Self-attention generative adversarial networks. In: Proceedings of the 36th International Conference on Machine Learning, 7354–7363, 2019.

[30] Portenier, T.; Hu, Q.; Szabo, A.; Bigdeli, S. A.; Favaro, P.; Zwicker, M. Faceshop: Deep sketch-based face image editing. *ACM Transactions on Graphics* Vol. 37, No. 4, Article No. 99, 2018.

[31] Chen, S.-Y.; Liu, F.-L.; Lai, Y.-K.; Rosin, P. L.; Li, C.; Fu, H.; Gao, L. DeepFaceEditing: Deep face generation and editing with disentangled geometry and appearance control. *ACM Transactions on Graphics* Vol. 40, No. 4, Article No. 90, 2021.

[32] Tan, Z.; Chai, M.; Chen, D.; Liao, J.; Chu, Q.; Yuan, L.; Tulyakov, S.; Yu, N. MichiGAN: Multi-input-conditioned hair image generation for portrait editing. *ACM Transactions on Graphics* Vol. 39, No. 4, Article No. 95, 2020.

[33] Huang, Z.; Peng, Y.; Hibino, T.; Zhao, C.; Xie, H.; Fukusato, T.; Miyata, K. DualFace: Two-stage drawing guidance for freehand portrait sketching. *Computational Visual Media* Vol. 8, No. 1, 63–77, 2022.

[34] Shen, Y. J.; Gu, J. J.; Tang, X. O.; Zhou, B. L. Interpreting the latent space of GANs for semantic face editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9240–9249, 2020.

[35] Shen, Y. J.; Zhou, B. L. Closed-form factorization of latent semantics in GANs. *arXiv preprint* arXiv:2007.06600, 2020.

[36] Tewari, A.; Elgharib, M.; Bharaj, G.; Bernard, F.; Seidel, H.-P.; Perez, P.; Zollhofer, M.; Theobalt, C. StyleRig: Rigging StyleGAN for 3D control over portrait images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6141–6150, 2020.

[37] Abdal, R.; Zhu, P. H.; Mitra, N.; Wonka, P. StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. *arXiv preprint* arXiv:2008.02401, 2020.

[38] Rezende, D.; Mohamed, S. Variational inference with normalizing flows. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning, Vol. 37, 1530–1538, 2015.

[39] Zhu, J.; Shen, Y.; Zhao, D.; Zhou, B. In-domain GAN inversion for real image editing. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12362*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 592–608, 2020.

[40] Sun, R. Q.; Huang, C.; Zhu, H. L.; Ma, L. Z. Mask-aware photorealistic facial attribute manipulation. *Computational Visual Media* Vol. 7, No. 3, 363–374, 2021.

[41] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 6000–6010, 2017.

[42] Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7794–7803, 2018.

[43] Zhang, P.; Zhang, B.; Chen, D.; Yuan, L.; Wen, F. Cross-domain correspondence learning for exemplar-based image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5142–5152, 2020.

[44] Lee, J.; Kim, E.; Lee, Y.; Kim, D.; Chang, J.; Choo, J. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5800–5809, 2020.

[45] Jiang, W.; Liu, S.; Gao, C.; Cao, J.; He, R.; Feng, J.; Yan, S. PSGAN: Pose and expression robust spatial-aware GAN for customizable makeup transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5193–5201, 2020.

[46] Huang, X.; Liu, M. Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11207*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 179–196, 2018.

[47] Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M. K.; Yang, M.-H. Diverse image-to-image translation via disentangled representations. In: Proceedings of the European Conference on Computer Vision, 2018.

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

[49] Yang, G. D.; Huang, X.; Hao, Z. K.; Liu, M. Y.; Belongie, S.; Hariharan, B. PointFlow: 3D point cloud generation with continuous normalizing flows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 4540–4549, 2019.

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

[51] Liu, Y.; Shi, H.; Shen, H.; Si, Y.; Wang, X.; Mei, T. A new dataset and boundary-attention semantic segmentation for face parsing. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 07, 11637–11644, 2020.

[52] Tan, Z.; Chen, D.; Chu, Q.; Chai, M.; Liao, J.; He, M.; Yuan, L.; Hua, G.; Yu, N. Efficient semantic image synthesis via class-adaptive normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 44, No. 9, 4852–4866, 2022.

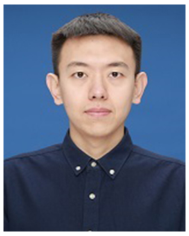
[53] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 6629–6640, 2017.

[54] Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 586–595, 2018.

[55] Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral normalization for generative adversarial networks. In: Proceedings of the International Conference on Learning Representations, 2018.

[56] Kingma, D. P.; Ba, J. L. Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference for Learning Representations, 2015.

**Cong Wang** is a Ph.D. candidate in the School of Mathematics at Jilin University. His research interests include image processing, computer vision, and deep learning.



**Fan Tang** is an assistant professor in the School of Artificial Intelligence, Jilin University. He received his Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2019. His research interests include computer graphics, computer vision, and machine learning.

**Yong Zhang** received his Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences in 2018. From 2015 to 2017, he was a visiting scholar with the Rensselaer Polytechnic Institute. He is currently with the Tencent AI Lab. His research interests include computer vision and machine learning.

**Tieru Wu** is a professor in the Institute of Mathematics at Jilin University. His research interests include techniques of computer graphics, geometry processing, and machine learning. His research is supported in part by the National Natural Science Foundation of China.

**Weiming Dong** is a professor in the Sino-European Lab in Computer Science, Automation and Applied Mathematics (LIAMA) and the National Laboratory of Pattern Recognition (NLPR) at the Institute of Automation, Chinese Academy of Sciences. He received his B.Sc. and M.Sc. degrees in computer science in 2001 and 2004 respectively, both from Tsinghua University. He received his Ph.D. degree in computer science from the University of Lorraine, France, in 2007. His research interests include image synthesis and image recognition. He is a member of the ACM and IEEE.