

Do Inpainting Yourself: Generative Facial Inpainting Guided by Exemplars

Wanglong Lu, Hanli Zhao*, Xianta Jiang, Xiaogang Jin, Yongliang Yang,
Min Wang, Jiankai Lyu, and Kaijie Shi

Abstract—We present EXE-GAN, a novel exemplar-guided facial inpainting framework using generative adversarial networks. Our approach can not only preserve the quality of the input facial image but also complete the image with exemplar-like facial attributes. We achieve this by simultaneously leveraging the global style of the input image, the stochastic style generated from the random latent code, and the exemplar style of exemplar image. We introduce a novel attribute similarity metric to encourage networks to learn the style of facial attributes from the exemplar in a self-supervised way. To guarantee the natural transition across the boundaries of inpainted regions, we introduce a novel spatial variant gradient backpropagation technique to adjust the loss gradients based on the spatial location. Extensive evaluations and practical applications on public CelebA-HQ and FFHQ datasets validate the superiority of EXE-GAN in terms of the visual quality in facial inpainting.

Index Terms—Image editing, image inpainting, facial image inpainting, facial attribute transfer, generative adversarial networks.

1 INTRODUCTION

FACES are widely recognized as the most representative and expressive aspect of human beings. With the advancement of digital imaging and mobile computing techniques, facial photographs may now be readily collected and distributed. This increases the need for effective and fast facial image altering in a convenient manner while keeping authenticity.

In this paper, we aim to solve a new face image manipulation problem. The goal is to seamlessly fill in the missing region of an input image by referring to the corresponding content of an exemplar image. This can largely help to generate a satisfactory face image that would favor various application scenarios, including recovering faces occluded by face masks, sunglasses, etc.; synthesizing faces of interest for person identification; designing personalized hairstyles according to existing examples; and generating face makeups for visual effects, to name just a few.

Many face image manipulation methods can achieve impressive manipulation of facial attributes based on guidance information, such as geometries [1], [2], semantics [3], and exemplars [4]. However, these methods often introduce unwanted changes to unedited regions and thus cannot guarantee visual information of known regions unchanged.

Facial inpainting plays an important role in facial image editing for filling missing or masked regions. To achieve realistic facial inpainting guided by exemplar images, there are two main challenges: how to learn the style of facial attributes from the exemplar and how to guarantee natural transition on the mask boundary. Some works [5], [6] attempt to generate diverse image inpainting results allowing users to select a desired one. But they cannot complete missing regions with user

guidance. Many recent methods try to employ additional landmarks [7], strokes [8], or sketches [9], [10] to guide the inpainting of facial structures and attributes. However, these methods tend to overfit the resulting images with these limited guidance information. As a result, these methods still require considerable professional skills in order to generate satisfactory target facial attributes, such as identity, expression, and gender.

To this end, we propose EXE-GAN, a novel interactive facial inpainting framework, which enables high-quality generative facial inpainting guided by exemplars. Our framework consists of four main components, including a mapping network, a style encoder, a multi-style generator, and a discriminator. Our method mixes the global style of input image, the stochastic style generated from the random latent code, and the exemplar style of exemplar image to generate high realistic images. We impose a perceptual similarity constraint to preserve the global visual consistency of the image. To enable the completion of exemplar-like facial attributes, we further employ facial identity and attribute constraints on the output result. To guarantee natural transition across the boundary of inpainted region, we devise a novel spatial variant gradient backpropagation method for the network training. We compare our method to the state-of-the-art methods to validate its advantages. Experimental results show that our method outperforms competitive methods in terms of visual quality. We also demonstrate several applications that could benefit from our framework, including local facial attribute transfer, guided facial style mixing, hairstyle editing, and guided facial image recovery (see Fig. 1).

In summary, our paper makes the following contributions:

- A novel interactive facial inpainting framework for high-quality generative inpainting of facial images



Fig. 1. Facial inpainting examples using our method. Top two rows: starting with the input image (the top-left sub-image with mask), our method gradually edits the eye style (left), the mouth style (middle left), the hair style (middle), and the facial styles (right) from exemplars. Hairstyles can be edited with the insertion of basic sketches (middle). Real-world and artistic face photos can both be used to direct the inpainting of (blended) facial features in the local edited regions without affecting the visual content of the rest of the image. Bottom row: For occluded portraits with eyeglasses and masks, we perform guided facial image recovery from exemplars.

with facial attributes guided by exemplars.

- A self-supervised attribute similarity metric to encourage the generative network to learn the style of facial attributes from exemplars.
- A novel spatial variant gradient backpropagation method for network training to guarantee realistic inpainting with natural transition on the boundary.
- Several applications benefiting from the proposed facial inpainting approach, including local facial attribute transfer, guided facial style mixing, hairstyle editing, and guided facial image recovery.

2 RELATED WORK

Traditional image inpainting techniques, such as diffusion-based methods [11], [12] and patch-based methods [13], [14], [15] mainly leverage low-level features to inpaint missing regions. These methods can perform well for small and narrow missing regions, but lack semantic understanding of the image.

Deep-learning-based inpainting methods employ deep neural networks and generative adversarial networks (GANs) [16] to achieve semantic completion, such as completion with global and local discriminators [17], attention-based inpainting [18], [19], [20], inpainting via multi-column CNNs [21], inpainting for irregular holes [22], [9], and high-resolution inpainting [23]. Some recent works [24], [5], [6], [25], [26], [27] try to generate pluralistic inpainted images without guidance information. For instance, Co-Mod-GAN [6] succeeds in completing large-scale missing regions and obtains diverse results by introducing inherent stochasticity. By taking advantage of auxiliary information, such as landmarks [7], sketches [9], [10], [8], and geometries [28], [29], structures and attributes of inpainted regions can be

generated accordingly. Since facial attributes contain rich visual information such as color, geometry and texture, these methods tend to overfit with simple geometric information. Thus, the variety of completed solutions is quite limited. Different from these methods, our method can effectively generate realistic inpainted face images guided by an exemplar image containing rich textural and semantic information.

Facial attribute transfer can be achieved by face manipulation methods using latent-guided codes or reference images. Semantic-level face manipulation methods can control a set of attributes, e.g., with or without glasses, mainly using a domain label to index the mapped style codes [30], [3]. However, these methods only operate on a set of pre-defined attributes and leave users little freedom for face manipulation. Geometry-based face manipulation methods implement exemplar-based facial transfer based on semantic geometry [1], [2]. Since information loss occurs in the projection and reconstruction between real-captured photos and corresponding representations, these methods tend to change fine details in the background. Exemplar-based face manipulation methods transfer facial attributes from exemplars at the instance level [31], [32], [33], [34], [4], [35]. In particular, SimSwap [34] extends the identity-specific face swapping to arbitrary face swapping. Nevertheless, these methods have a common limitation as users cannot flexibly select facial regions for the local transfer of facial attributes. Kim et al. [36] proposed an intermediate representation with spatial dimensions to perform local editing, but their method may fail when the poses in the original and reference images differ. In contrast, our method transfers local facial attributes from exemplars interactively to produce realistic inpainted face images

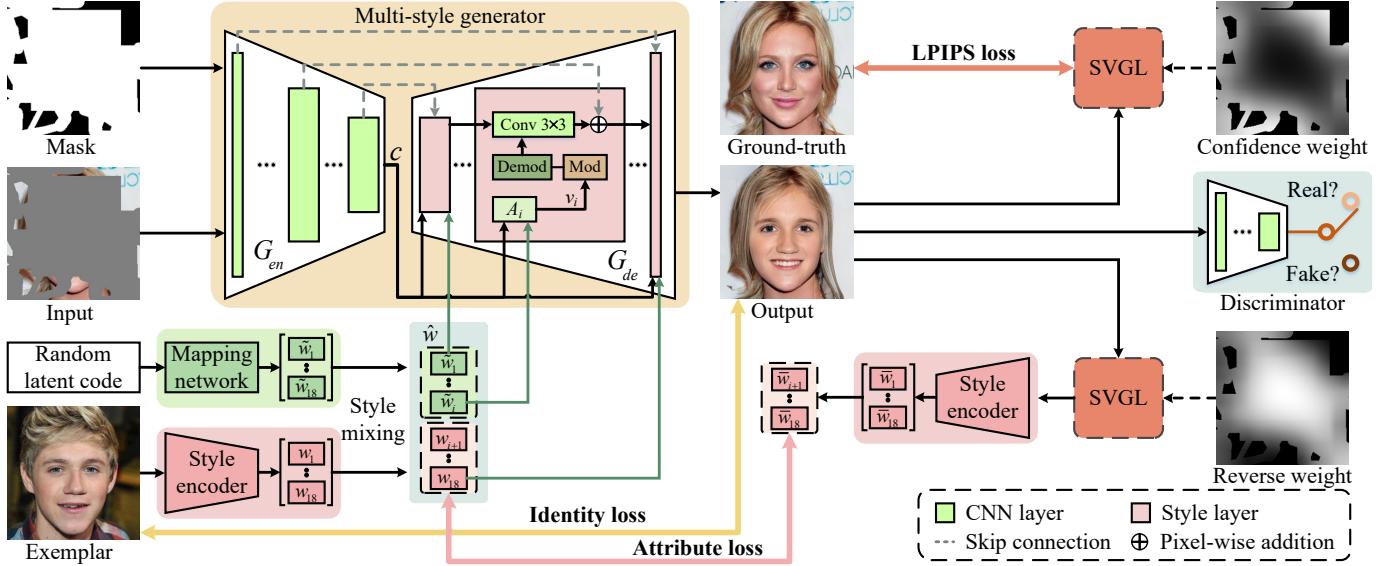


Fig. 2. Overview of our EXE-GAN framework. We employ style mixing on stochastic and exemplar style codes, and modulate them with the global style code of input image into the multi-style generator for facial inpainting. The adversarial, identity, LPIPS, and attribute losses are integrated as the overall training objective. Spatial variant gradient layers (SVGL) are utilized for natural transition across the filling boundary.

with natural attribute transitions while remaining unaltered regions.

Our work is also closely related to *image embedding* which enables image synthesis from latent space. StyleGANs [37], [38] enable direct scale-specific control of image synthesis with disentangled intermediate latent style space and can produce plausible results for unconditional face image synthesis. Optimization-based embedding [39], [40] and encoder-based embedding [41], [42], [43], [44], [45] methods perform image manipulation by inverting an image to the latent space [46]. Recently, EditingInStyle [47] and StyleFusion [48] show the impressive performance of image editing by semantically manipulating the style latent space. Although image embedding has the strong capability in presenting image styles, these methods may change unedited regions because of information losses in GAN inversion. For instance, Richardson et al. [42] presented results of inpainting using a pixel2style2pixel framework but failed to preserve visual contents of unmasked parts. In this work, we propose a novel facial inpainting framework by taking advantage of style latent codes while keeping the unmasked region.

3 METHOD

3.1 Overview

The overall structure of our EXE-GAN framework is shown in Fig. 2. Given a ground-truth face image $I_{gt} \in \mathbb{R}^{h \times w \times 3}$, an exemplar image $I_{exe} \in \mathbb{R}^{h \times w \times 3}$, and a binary mask $M \in \mathbb{R}^{h \times w \times 1}$ (with value 1 for unknown and 0 for known pixels), the input image $I_{in} \in \mathbb{R}^{h \times w \times 3}$ is obtained by $I_{in} = I_{gt} \odot (1 - M)$, where \odot denotes the Hadamard product. The goal of our EXE-GAN framework is to

automatically generate a realistic face image I_{out} , where the inpainting of the masked regions in I_{in} is guided by the facial attributes of I_{exe} while the known regions remain unchanged. The proposed EXE-GAN consists of four main components, including a mapping network, a style encoder, a multi-style generator, and a discriminator.

3.1.1 Mapping network

A multi-layer fully-connected neural network f [38], [6] linearly maps a random latent code $z \in \mathbb{R}^{512 \times 1}$ to a stochastic style code $\tilde{w} = \{\tilde{w}_i \in \mathbb{R}^{512 \times 1} | i \in T\} \in \widetilde{W}^+$, where \widetilde{W}^+ denotes the extended stochastic style latent space and $T = \{1, 2, \dots, 18\}$ denotes the index set. Let θ_f be the learnable network parameters in f , we have $\tilde{w} = f(z; \theta_f)$.

3.1.2 Style encoder

A pre-trained pixel2style2pixel style encoder E [42], [43] directly maps an image to a disentangled style latent space W^+ . Given the pre-trained network parameters θ_e , the style encoder extracts the exemplar style code $w = \{w_i \in \mathbb{R}^{512 \times 1} | i \in T\} = E(I_{exe}; \theta_e)$ and the inpainted style code $\bar{w} = \{\bar{w}_i \in \mathbb{R}^{512 \times 1} | i \in T\} = E(I_{out}; \theta_e)$. Therefore, $w \in W^+$ and $\bar{w} \in W^+$.

3.1.3 Multi-style generator

A generative network G that leverages multiple representations (i.e., I_{in} , M , and \hat{w}) to generate an intermediate result $I_{pred} \in \mathbb{R}^{h \times w \times 3}$, where $\hat{w} = \{\hat{w}_i \in \mathbb{R}^{512 \times 1} | i \in T\}$ is the mixed style code of w and \tilde{w} . Let θ_g be the learnable network parameters of G , we have $I_{pred} = G(I_{in}, M, \hat{w}; \theta_g)$. The multi-style generator can be

further divided into an encoder G_{en} and a decoder G_{de} , i. e., $G = \{G_{en}, G_{de}\}$.

3.1.4 Discriminator

A discriminative network D [38], [6] learns to judge whether an image is a real or fake image. Let θ_d be the learnable network parameters of D , the discriminative network maps the inpainted image I_{out} to a scalar $D(I_{out}; \theta_d) \in \mathbb{R}^{1 \times 1}$.

3.2 Multi-style modulation

To leverage the global style of the input image, the stochastic style generated from the random latent code, and the exemplar style of exemplar image to perform generative facial inpainting, we build upon the generator of Co-mod-GAN [6] and extend it to our multi-style generator G . This is done by incorporating also the exemplar style, and mixing it with other styles based on carefully designed style modulation. The proposed multi-style generator can not only preserve the global visual consistency of the input image, but also embed exemplar facial attributes to the local facial inpainting. In addition, it has the good property of inherent stochasticity with the stochastic style latent code.

First of all, the mixed style code \hat{w} is obtained by style mixing [37], [38] of the stochastic and exemplar styles. Specifically, each layer of \hat{w} is defined as:

$$\hat{w}_i = \begin{cases} w_i, & \text{if } \phi_i = 1, \\ \tilde{w}_i, & \text{otherwise,} \end{cases} \quad (1)$$

where $i \in T = \{1, 2, \dots, 18\}$ and $\phi \in \mathbb{R}^{18 \times 1}$ is a binary vector to indicate which style is modulated for each layer. As demonstrated in StyleGAN [37], coarse-resolution layers correspond to high-level facial attributes and fine-resolution layers could change small-scale features. We empirically set $\phi = [0, 0, 0, 0, 1, 1, \dots, 1]$ by balancing the stochastic and exemplar styles in this paper.

Secondly, the encoder G_{en} takes I_{in} and M as input, and outputs a global style code $c \in \mathbb{R}^{2 \times 512 \times 1}$ as well as the corresponding multi-resolution feature maps.

Then, as illustrated in Fig. 2, the global style code c and the mixed style code \hat{w} are transformed to multi-style vectors v for subsequent modulation within the style layers of the decoder G_{de} . For each i -th style layer, the transformation is defined as [38]:

$$v_i = A_i([c, \hat{w}_i]), \quad (2)$$

where $[\cdot]$ refers to the concatenation operator, A_i is a learned affine transformation within the i -th style layer, and v_i is a linearly learned style representation conditioned on the input style representations.

Next, the decoder G_{de} utilizes the multi-style vectors v and the multi-resolution feature maps output by G_{en} to generate the intermediate inpainting I_{pred} . The decoder contains two style layers in each resolution. In each i -th style layer, the multi-style vector v_i is then used

for weight modulation and demodulation [38], [6]. As shown in Fig. 2, skip connections are used for collecting the multi-resolution feature maps in the decoder G_{de} .

Finally, the inpainted image I_{out} is generated as follows:

$$I_{out} = I_{in} \odot (1 - M) + I_{pred} \odot M. \quad (3)$$

3.3 Training objectives

Our framework is trained to optimize the learnable network parameters θ_g , θ_f , and θ_d using the following objectives.

3.3.1 Adversarial loss

We use the adversarial non-saturating logistic loss [16] with R_1 regularization [49]. Specifically, the adversarial objective is defined as:

$$\begin{aligned} \mathcal{L}_{adv}(I_{out}, I_{gt}) = & \mathbb{E}_{I_{out}}[\log(1 - D(I_{out}))] \\ & + \mathbb{E}_{I_{gt}}[\log(D(I_{gt}))] - \frac{\gamma}{2} \mathbb{E}_{I_{gt}}[\|\nabla_{I_{gt}} D(I_{gt})\|_2^2], \end{aligned} \quad (4)$$

where γ is used to balance the R_1 regularization term. We empirically set $\gamma = 10$. The generative network G learns to generate a visually realistic image I_{out} while the discriminative network D tries to distinguish between the ground-truth I_{gt} and the generated image I_{out} . G and D are trained in an alternating manner.

3.3.2 Identity loss

We constrain identity similarity between the output image I_{out} and the exemplar image I_{exe} in the embedding space. The identity loss is formulated as follows:

$$\mathcal{L}_{id}(I_{out}, I_{exe}) = 1 - \cos(R(I_{out}), R(I_{exe})), \quad (5)$$

where $R(\cdot)$ is a pre-trained ArcFace network [50] for face recognition.

3.3.3 LPIPS loss

We employ the Learned Perceptual Image Patch Similarity (LPIPS) loss [51] to constrain the perceptual similarity between the output image I_{out} and the ground-truth I_{gt} :

$$\mathcal{L}_{lpips}(I_{out}, I_{gt}) = \begin{cases} \|F(I_{out}) - F(I_{gt})\|_2 & \text{if } I_{gt} = I_{exe}, \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

where $F(\cdot)$ is the pre-trained perceptual feature extractor and we adopt VGG [52] in our work. Note that \mathcal{L}_{lpips} is applied only when I_{gt} and I_{exe} are sampled from the same image (see Section 4.1 for the detailed settings).

3.3.4 Attribute loss

In order to learn the style of facial attributes from the exemplar image, we introduce a novel self-supervised attribute similarity metric to measure the consistency

between facial attributes of the inpainted result I_{out} and the exemplar I_{exe} in the style latent space:

$$\mathcal{L}_{attr}(I_{out}, I_{exe}) = \frac{1}{\|\phi\|_0} \sum_{i \in T} \phi_i \cdot \|\bar{w}_i - \hat{w}_i\|_2, \quad (7)$$

where the L0 norm $\|\cdot\|_0$ indicates the number of non-zeros.

3.3.5 Total objective

After defining loss functions above, the total training objective can be expressed as:

$$O(\theta_g, \theta_f, \theta_d, \hat{\theta}_e) = \mathcal{L}_{adv}(I_{out}, I_{gt}) + \lambda_{id}\mathcal{L}_{id}(I_{out}, I_{exe}) \\ + \lambda_{lpips}\mathcal{L}_{lpips}(I_{out}, I_{gt}) + \lambda_{attr}\mathcal{L}_{attr}(I_{out}, I_{exe}), \quad (8)$$

where λ_{id} , λ_{lpips} , and λ_{attr} are weights of corresponding losses, respectively. We empirically set $\lambda_{id} = 0.1$, $\lambda_{lpips} = 0.5$, and $\lambda_{attr} = 0.1$ in this work. During training, we can obtain the optimized parameters θ_g , θ_f , and θ_d via the minimax game iteratively:

$$(\theta_g, \theta_f) = \arg \min_{\theta_g, \theta_f} O(\theta_g, \theta_f, \theta_d, \hat{\theta}_e), \\ (\theta_d) = \arg \max_{\theta_d} O(\theta_g, \theta_f, \theta_d, \hat{\theta}_e). \quad (9)$$

3.4 Spatial variant gradient backpropagation

It is expected that the inpainted facial attributes close to the filling center are more similar to those of the exemplar image. Moreover, the inpainted values close to the boundary should be perceptually more similar to those of input image and the visual contents should be naturally transited on the boundary. Therefore, in order to generate naturally looking inpainting, we further exert constraint based on spatial location.

From Eqs. 6 and 7, we can find that the LPIPS loss and attribute loss are defined over the entire inpainted image. GMCNN [21] applies the spatial constraint to the pixel-wise reconstruction loss. However, we cannot directly impose GMCNN's spatial constraint on our loss functions. The reason is that our losses are defined in embedding space and dimensions of embedding features do not match those of the spatial space. In our work, a novel spatial variant gradient layer (SVGL) is designed to impose the spatial constraint on loss gradients in backpropagation.

As illustrated in Fig. 3, SVGL has no parameter but relies on a spatial weight mask. During forward-propagation, SVGL acts as an identity transform, which does not change any information from the input. During backpropagation, it collects gradients from subsequent layers, re-weights the gradients based on the spatial weight mask, and passes the re-weighted gradients to the preceding layers.

Mathematically, given an input feature x and a spatial weight mask M_x , we can treat SVGL as a "pseudo-

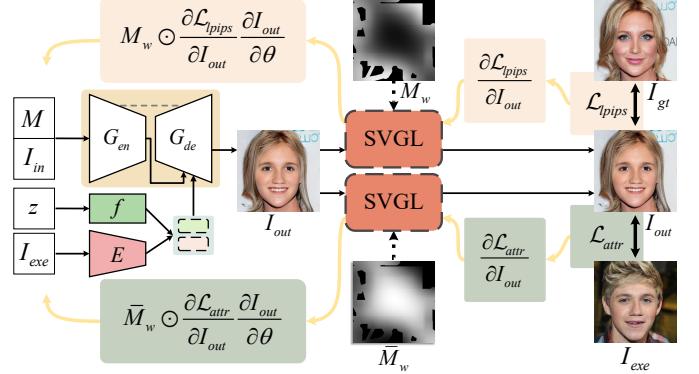


Fig. 3. Illustration of the SVGL on LPIPS and attribute losses. In forward-propagation, SVGL does not change any information for I_{out} . In backpropagation, gradients are re-weighted based on the spatial variant M_w and \bar{M}_w , respectively.

function" $P(x, M_x)$. The forward-propagation and back-propagation behaviors of SVGL are defined below:

$$P(x, M_x) = x, \\ \frac{\partial P(x, M_x)}{\partial x} = M_x \odot \mathbf{I}, \quad (10)$$

where \mathbf{I} represents an identity matrix.

Then, we apply SVGL to the spatial variant LPIPS and attribute losses. Specifically, we equipped the network with a SVGL $P(\cdot, M_w)$ for the spatial variant LPIPS loss and a SVGL $P(\cdot, \bar{M}_w)$ for the spatial variant attribute loss, respectively, where the confidence weight mask $M_w \in \mathbb{R}^{h \times w \times 1}$ is obtained with Gaussian smoothing on the masked region of M and the reverse weight mask $\bar{M}_w = (1 - M_w) \odot M \in \mathbb{R}^{h \times w \times 1}$. As shown in Fig. 3, both SVGLs are added right after the layer of generating I_{out} . Our SVGL is general and can be used to apply spatial constraints to any loss functions with spatial variant backpropagation. Note that the values of non-masked regions are zeros and the weights of the pre-trained style encoder are frozen during training. With the spatial variant gradient layers, the training objective is computed with Eq. 8 during forward-propagation while its gradients are computed in a spatial variant manner.

4 EXPERIMENTS

4.1 Settings

4.1.1 Datasets

Experiments were conducted on two publicly available face image datasets CelebA-HQ [53] and FFHQ [37]. For CelebA-HQ [53], we randomly selected 28,000 images for training and remained 2,000 images for testing. For FFHQ [37], we randomly selected 60,000 images for training and the rest 10,000 images for testing. Each image was resized to 256×256 .

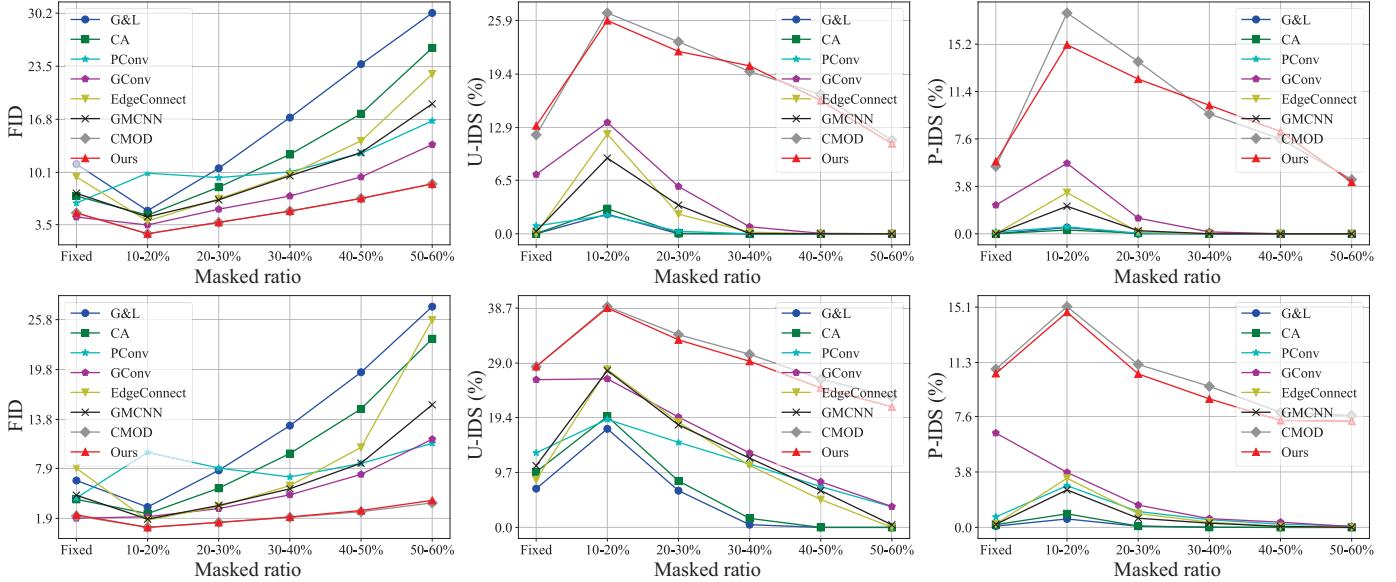


Fig. 4. Quantitative comparisons between our method and the state-of-the-art free-form inpainting methods on the CelebA-HQ (top row) and FFHQ (bottom row) datasets, respectively.

Algorithm 1 Training procedure of EXE-GAN

```

1: while  $f$ ,  $G$ , and  $D$  have not converged do
2:   Sample batch images  $\mathcal{I}_{gt}$  from training data
3:   Sample random latent vectors  $\mathcal{Z}$ 
4:   Sample a random number  $r \in [0, 1]$ 
5:   if  $r >$  threshold  $\tau$  then
6:     Sample batch exemplars  $\mathcal{I}_{exe}$  from training data
7:   else
8:     Set batch exemplars from ground-truth
 $\mathcal{I}_{exe} \leftarrow \mathcal{I}_{gt}$ 
9:   Create random masks  $\mathcal{M}$  for  $\mathcal{I}_{in}$ 
10:  Get confidence weight masks  $\mathcal{M}_w$  for  $\mathcal{L}_{lpips}$ 
11:  Get reverse weight masks  $\bar{\mathcal{M}}_w$  for  $\mathcal{L}_{attr}$ 
12:  Get inputs  $\mathcal{I}_{in} \leftarrow \mathcal{I}_{gt} \odot (1 - \mathcal{M})$ 
13:  Get  $\hat{w} \leftarrow mixing(E(\mathcal{I}_{exe}), f(\mathcal{Z}))$ 
14:  Get  $\mathcal{I}_{pred} \leftarrow G(\mathcal{I}_{in}, \mathcal{M}, \hat{w})$ 
15:  Get outputs  $\mathcal{I}_{out} \leftarrow \mathcal{I}_{in} \odot (1 - \mathcal{M}) + \mathcal{I}_{pred} \odot \mathcal{M}$ 
16:  Update  $f$  and  $G$  with  $\mathcal{L}_{adv}$ ,  $\mathcal{L}_{id}$ ,  $\mathcal{L}_{lpips}$ , and  $\mathcal{L}_{attr}$ 
17:  Update  $D$  with  $\mathcal{L}_{adv}$ 

```

4.1.2 Evaluation metrics

The performance was quantitatively evaluated using the Fréchet inception distance (FID) [54] and the paired/unpaired inception discriminative score (P-IDS/U-IDS) [6]. FID has been proven to correlate well with human perception for the visual quality of generated images. P-IDS and U-IDS are robust assessment measures for the perceptual fidelity of generative models.

4.1.3 Implementation details

Algorithm 1 lists the pseudo-code for our EXE-GAN framework's training procedure. The threshold $\tau \in [0, 1]$ was used to control the probability that the sampled ground-truth image and exemplar image were the same. We set threshold $\tau = 0.1$ in this paper. Our framework was implemented using Python and PyTorch. Following the settings of StyleGANv2 [38] and Co-mod-GAN [6], we employed the Adam optimizer with the first momentum coefficient of 0.5, the second momentum coefficient of 0.99, and the learning rate of 0.002. Mixing regularization [37] with a probability of 0.5 was employed to generate stochastic style codes during training. The free-form mask sampling strategy was adopted for training by simulating random brush strokes and rectangles. The brush strokes were generated using the algorithm presented in GConv [9] with maxVertex of 20, maxLength of 100, maxBrushWidth of 24, and maxAngle of 360. Multiple up-to-half-size rectangles and up-to-quarter-size rectangles were generated randomly. The numbers of up-to-half-size rectangles and up-to-quarter-size rectangles were uniformly sampled within [0, 5] and [0, 10], respectively. We trained the networks for 800,000 iterations with batch size of 8. All experiments were conducted on the NVIDIA Tesla V100 GPU. The training time was around three weeks.

4.2 Comparison to free-form inpainting

We compared EXE-GAN on CelebA-HQ and FFHQ datasets to the state-of-the-art free-form inpainting methods, including Contextual Attention (CA) [18], Partial Convolutions (PConv) [22], Globally & Locally (G&L) [17], Gated Convolution (GConv) [9], EdgeConnect [10], GMCNN [21], and Co-mod-GAN (CMOD) [6].

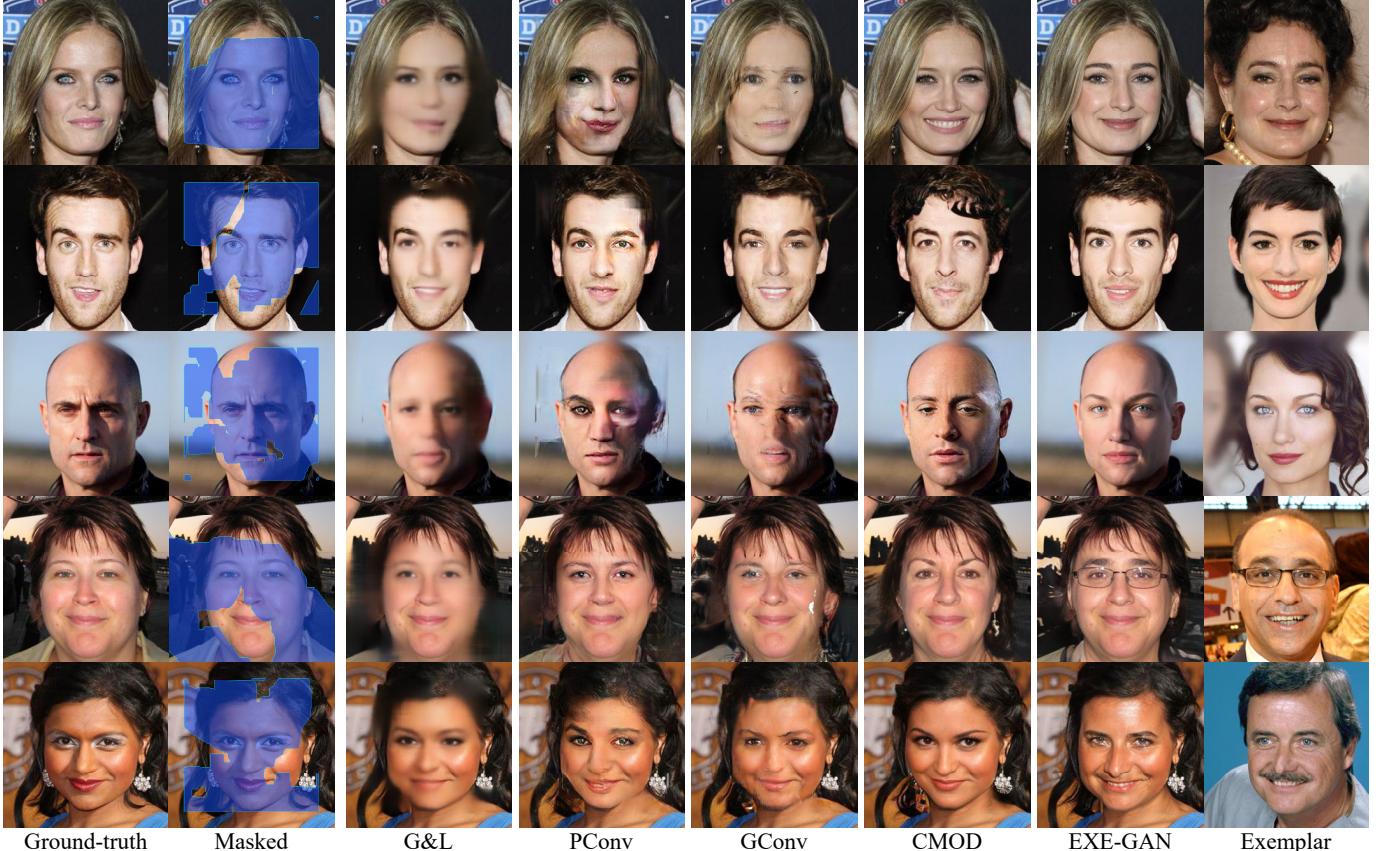


Fig. 5. Qualitative comparison between our method and the state-of-the-art free-form inpainting methods.

4.2.1 Experiment settings

We used the publicly available MMEditing framework [55] for Contextual Attention (CA) [18], Partial Convolutions (PConv) [22], Globally & Locally (G&L) [17], and Gated Convolution (GConv) [9]. MMEditing is an open-source image and video editing toolbox based on PyTorch. We used the official codes for EdgeConnect [10] and GMCNN [21]. For Co-modGAN (CMOD) [6], we used the official TensorFlow-based code to implement a PyTorch-based version. All the compared free-form inpainting models were trained using CelebA-HQ and FFHQ datasets, respectively. For fair comparison, we used the same training/testing splits for all experiments. We randomly took exemplar images during training while using reverse indices to the input image batch during testing.

4.2.2 Quantitative comparison

Fig. 4 shows the quantitative performance comparisons between our method and the state-of-the-art free-form methods on CelebA-HQ and FFHQ datasets, respectively. Various irregular masks with different masked ratios as well as a fixed rectangle center mask with the size of 128×128 were employed to simulate various situations for facial image inpainting. Quantitative results show that our method performs better than most of the compared free-form inpainting methods, even though

the inpainting of our method was guided by exemplars, which was considered challenging to keep image quality [6]. Moreover, our EXE-GAN can achieve comparable performance to CMOD [6] with various kinds of masks in terms of FID [54], U-IDS and P-IDS [6].

4.2.3 Qualitative comparison

As shown in Fig. 5, although all the methods can compatibly fill in the missing regions, G&L [17] tends to produce blurry inpainting while PConv [22] and GConv [9] fail to inpaint large-scale missing regions. Our method and CMOD [6] can generate competitive inpainted results. However, the inpainting of facial attributes cannot be controlled with CMOD [6]. In contrast, with the help of exemplar facial attributes, the inpainting of facial attributes with our EXE-GAN can be controlled easily.

4.3 Comparison to guidance-based inpainting

We compared our method on the Celeba-HQ dataset to the state-of-the-art guidance-based facial inpainting methods, including sketch-and-color-based facial inpainting SC-FEGAN [8] and landmark-based face inpainting LaFIn [7]. For the compared methods, we generated corresponding guided information from exemplar images and alternately took one facial image as the



Fig. 6. Qualitative comparison between our method and the state-of-the-art guidance-based facial inpainting methods.

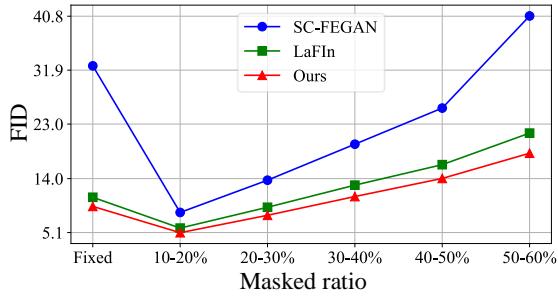


Fig. 7. Quantitative comparison between our method and the state-of-the-art guidance-based facial inpainting methods on the Celeba-HQ dataest.

exemplar and the other one as the masked image for facial attribute inpainting.

4.3.1 Experiment settings

The officially released pre-trained SC-FEGAN [8] and LaFIn [7] models were used in this experiment. SC-FEGAN [8] uses sketches and color as the guidance to generate missing pixels. Therefore, we leveraged the Canny edge detector to automatically generate sketches from the exemplar image. To avoid inconsistency of color in the inpainted pixels, we did not introduce color information of the exemplar into missing regions. LaFIn [7] relies on landmarks to fill missing regions. Therefore,

we utilized the face alignment network FAN [56] to generate landmarks for the exemplar image. To avoid the misalignment between the guidance information (i.e., sketches and landmarks) and unmasked regions in the masked image, we first extracted the angles of roll, pitch, and yaw from the CelebAMask-HQ dataset, then selected 550 pairs with similar poses from the testing set. For each pair, we alternately took one facial image as the exemplar and the other one as the masked image to perform facial attribute inpainting. As a consequence, we obtained 1,100 inpainted images for each comparison method.

4.3.2 Quantitative comparison

Fig. 7 shows the quantitative comparison of our EXE-GAN to SC-FEGAN [8] and LaFIn [7]. FID scores with various masked ratios were compared. Experimental results show that our EXE-GAN is able to achieve the best FID scores for all kinds of masks. In terms of the authenticity of inpainted images guided by exemplars, our method outperforms the compared methods.

4.3.3 Qualitative comparison

As shown in Fig. 6, SC-FEGAN [8] effectively generates facial attributes with shapes guided by sketches but requires more information for high-quality facial attributes inpainting. Moreover, there may be visual artifacts in the

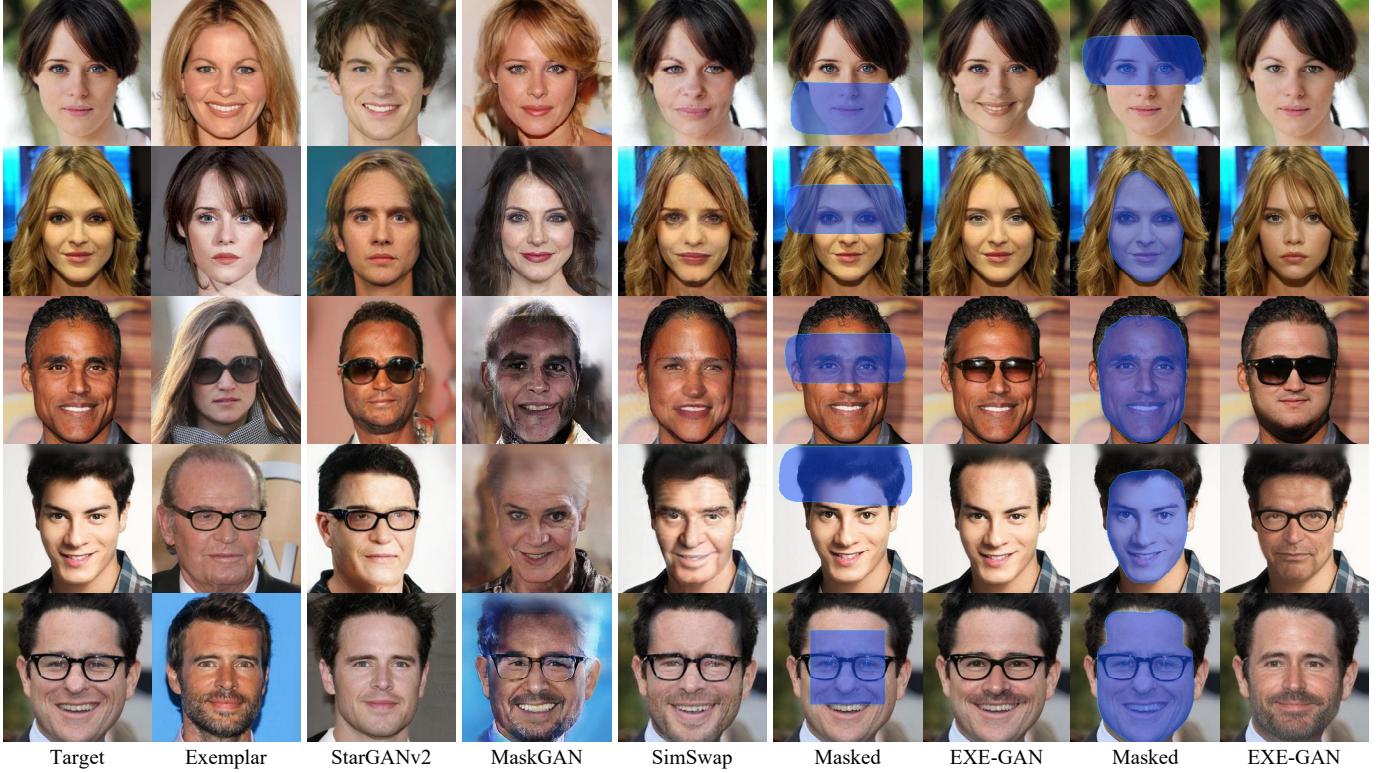


Fig. 8. Qualitative comparison between our method and the state-of-the-art facial attribute transfer methods.

inpainted images with SC-FEGAN in Fig. 6. LaFIn [7] generates facial expressions similar to exemplars but may fail to inpaint the decorative attributes, such as glasses and hairstyles in Fig. 6. In comparison, our method directly learns the style of facial attributes from the exemplar without extra input, and can successfully generate exemplar-like facial attributes including decorative attributes. As shown in Fig. 6, our EXE-GAN is able to produce more realistic facial inpainted results with facial attributes similar to exemplars.

4.3.4 User study

We further recruited 63 volunteers to subjectively evaluate the effectiveness of our method comprehensively. For the user study, we randomly selected 100 pairs of images from the 550 pairs of images with similar poses mentioned above. Then we randomly divided these selected 100 pairs into 5 groups, and each group consists of 20 pairs with different types of inpainting masks. For each pair, we set one image as the masked input and the other as the exemplar to generate inpainting results using the methods of SC-FEGAN [8], LaFIn [7], and our EXE-GAN, respectively. We then randomly ordered one of these 5 groups to a user. For each round, an exemplar image and its corresponding three inpainting images of SC-FEGAN [8], LaFIn [7], and our EXE-GAN were provided. The participants were asked to select the best image based on the visual quality of inpainting and the perceptual similarity to the exemplar. The results show that our EXE-GAN obtains the majority of votes (59.67%)

compared to SC-FEGAN [8] (12.87%) and LaFIn [7] (27.46%). The user study validates that the output of our EXE-GAN is more realistic than the compared methods visually observed by subjects.

4.4 Comparison to facial attribute transfer

We compared our method on the Celeba-HQ dataset to the state-of-the-art facial attribute transfer methods, including StarGANv2 [3], MaskGAN [1], and SimSwap [34], where the same exemplar image was used to guide the attribute transfer.

4.4.1 Experiment settings

The pre-trained models of StarGANv2 [3], MaskGAN [1], and SimSwap [34] provided in the official online repository were used in this experiment. For StarGANv2 [3], we set the input image as the “reference” image and the exemplar image as the “source” image. For MaskGAN [1], we extracted semantic masks of input images from the CelebAMask-HQ dataset and obtained the style transferred results based on semantic masks and exemplars. SimSwap [34] directly performs the exemplar-guided face synthesis with the input and exemplar images. Our EXE-GAN synthesizes facial attributes for masked regions of input images guided by exemplar images.

4.4.2 Qualitative comparison

As shown in Fig. 8, StarGANv2 [3] can transform an input image reflecting the identity of the exemplar. How-

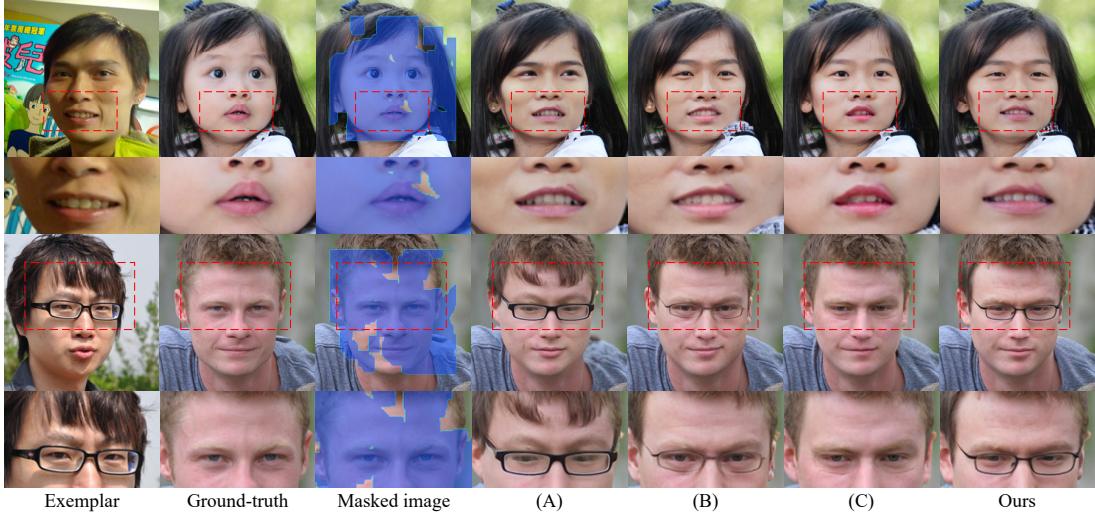


Fig. 9. Qualitative examples of the ablation study for large-scale facial inpainting by exemplars with (A) EXE-GAN without any SVGLs in \mathcal{L}_{attr} and \mathcal{L}_{lpips} , (B) EXE-GAN without SVGL in \mathcal{L}_{attr} , (C) EXE-GAN without SVGL in \mathcal{L}_{lpips} , and (Ours) EXE-GAN.

TABLE 1

Ablation study for large-scale facial inpainting with (A) baseline CMOD [6], (A) CMOD + standard \mathcal{L}_{lpips} , and (B) CMOD + SVGL-based \mathcal{L}_{lpips} . Results are averaged over 5 runs. **Bold**: top-2 quantity.

Method	CelebA-HQ			FFHQ		
	FID \downarrow	U-IDS \uparrow	P-IDS \uparrow	FID \downarrow	U-IDS \uparrow	P-IDS \uparrow
CMOD	9.733	9.68%	4.00%	4.000	26.19%	12.88%
(A)	8.993	10.87%	4.70%	3.432	29.08%	13.52%
(B)	8.985	11.25%	4.85%	3.318	29.97%	14.88%

ever, it leaves users little freedom to manipulate face images interactively. MaskGAN [3] transfers the style of exemplar to the input face image using the semantic mask. It requires projecting images into semantic masks and reconstructing images from the mask manifold. As a result, it may introduce irrelevant changes to fine details in the background. SimSwap [34] can transfer the identity of the exemplar face to the input face and preserve the facial attributes of the input. Nevertheless, it does not allow users to flexibly select regions for face editing. In comparison, our method not only preserves the pixels of known regions but also allows more degrees of freedom to interactively perform the facial attribute manipulation. Our EXE-GAN is able to produce high-quality results with facial attributes guided by exemplars, including gender, makeup style, hairstyle, and decorative style (e.g., glasses).

4.5 Ablation study

4.5.1 Ablation study on SVGL in CMOD [6]

To demonstrate the effectiveness of the proposed spatial variant gradient layer (SVGL), we experimented to apply our SVGL to CMOD for the image inpainting task.

TABLE 2

Ablation study for large-scale facial inpainting by exemplars with (A) EXE-GAN without any SVGLs in \mathcal{L}_{attr} and \mathcal{L}_{lpips} , (B) EXE-GAN without SVGL in \mathcal{L}_{attr} , (C) EXE-GAN without SVGL in \mathcal{L}_{lpips} , and (Ours) EXE-GAN. Results are averaged over 5 runs. **Bold**: top-2 quantity.

Method	CelebA-HQ			FFHQ		
	FID \downarrow	U-IDS \uparrow	P-IDS \uparrow	FID \downarrow	U-IDS \uparrow	P-IDS \uparrow
(A)	12.853	4.024%	1.25%	7.298	17.29%	6.55%
(B)	10.433	8.875%	3.35%	4.909	22.46%	8.75%
(C)	9.804	8.875%	4.25%	4.408	24.61%	10.04%
Ours	9.967	9.175%	3.85%	4.353	24.33%	9.92%

As shown in Table 1, the performance is improved by adding the LPIPS loss into the baseline CMOD. The quantitative scores are considerably improved by further introducing SVGL in the LPIPS loss. The SVGL-based LPIPS loss helps the generator focus more on pixels close to the hole boundary to avoid visual inconsistency and still encourage inherent stochasticity with less constraints to pixels away from the boundary. This ablation study shows that the proposed SVGL can effectively boost the performance of image inpainting.

4.5.2 Ablation study on SVGL in EXE-GAN

We further investigated the effectiveness of each component in our EXE-GAN by performing ablation study. Table 2 shows the quantitative results. We also present qualitative examples in Fig. 9 to better express visual effects of the ablation study. When removing SVGLs in both the attribute loss and the LPIPS loss (A), both the quantitative measures and visual qualities drop dramatically. When replacing the SVGL-based attribute loss with the standard attribute loss without SVGL (B), there may be visible boundary inconsistencies in the generated

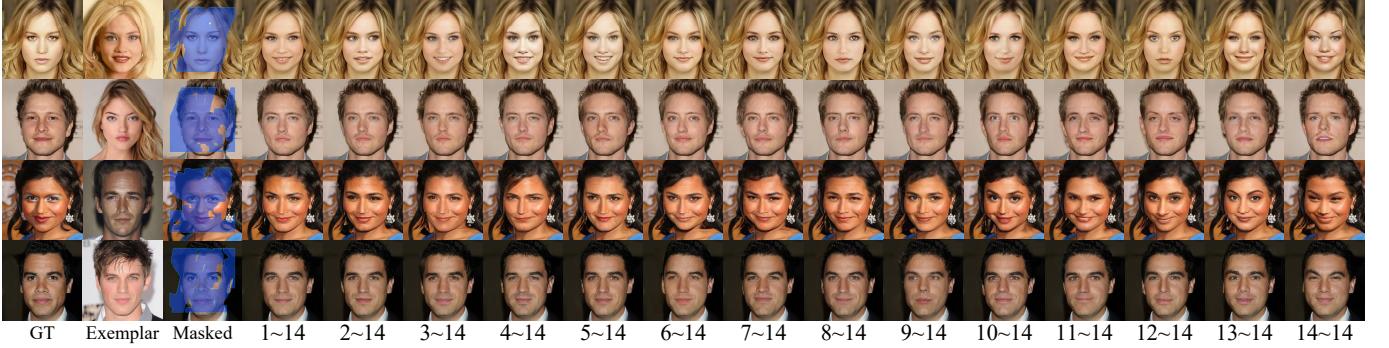


Fig. 10. Examples of facial inpainting with various subsets of the style codes: ground-truth, exemplar, masked image, and various style effects. In each row, values of i -th to j -th layers in the style code are from the exemplar and values of remaining layers are from the stochastic style code.

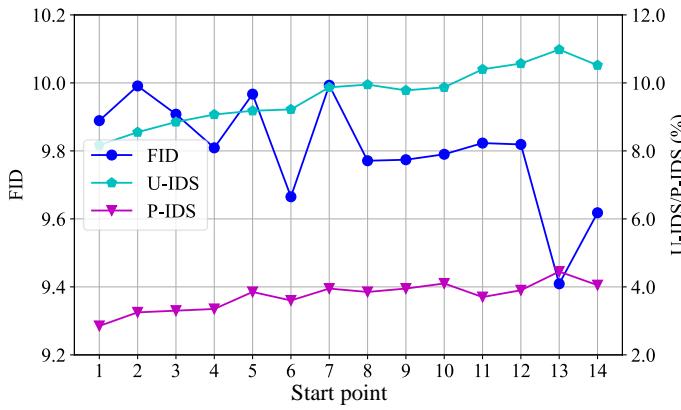


Fig. 11. Ablation study for large-scale facial inpainting by exemplars on modulated exemplar style codes on the CelebA-HQ dataset. The start point i indicates i -th to 14-th style layers are modulated with exemplar styles while the other layers are with random styles.

results, and the quantitative performance is also affected. When replacing the SVGL-based LPIPS loss with the standard LPIPS loss without SVGL (C), the visual similarities of facial attributes (e.g., facial expression, wearing glasses) between the generated result and the exemplar image decrease, while the quantitative scores are comparable to EXE-GAN for both testing datasets. In this case, the standard LPIPS loss is applied to all pixels of the image to enforce the generator to reconstruct the contents of ground-truth instead of exemplar attributes, as demonstrated in Fig. 9. In comparison, our EXE-GAN is able to produce realistic facial images with facial attributes similar to exemplars and competitive quantitative scores.

4.5.3 Ablation study on exemplar styles modulation

We conducted ablation experiments on the exemplar styles modulation by re-training our model with various configurations of the vector ϕ . As shown in Eq. 1, ϕ is a binary vector to indicate which style is modulated for each style layer. From Fig. 10 we can find that the more

exemplar style codes are modulated the more exemplar facial attributes will present in the inpainted images. We recorded the quantitative scores for various subsets of the style codes in Fig 11. The quantitative scores gradually get better with slight fluctuations when fewer exemplar styles are modulated. This ablation study validated that more exemplar styles lead to more exemplar facial attributes in inpainted images while the quantitative performance decreases.

5 APPLICATIONS WITH OUR FACIAL INPAINTING

In this section, we demonstrate a number of applications by equipping our generative facial inpainting method.

5.1 Local facial attribute transfer

Since our EXE-GAN helps the generator learn the mapping between injected exemplar representations and corresponding facial attributes, our method can be used to produce vivid facial attribute transfer effects guided by various exemplars, such as real-world facial attributes and artistic expressions. As shown in Fig. 12 and Fig. 13, for a masked input, our EXE-GAN produces high-quality local facial attribute transfer results by leveraging facial attributes of exemplars.

5.2 Guided facial style mixing

This application shows that our EXE-GAN can be used to generate facial inpainting effects by mixing two exemplar style latent codes. We first employ the style encoder E to obtain two exemplar style codes from two exemplars, respectively. Then, we apply the style mixing [37], [38] on the two latent codes with a crossover point. By simply changing the crossover point, we can obtain multiple mixed latent codes. Therefore, guided facial style mixing effects can be obtained by moving the crossover point over the vector ϕ in Eq. 1, as shown in Fig. 14.



Fig. 12. Examples of local facial attribute transfer guided by different cartoon exemplar images (left group) and real-world exemplar images (right group).



Fig. 13. More examples of local facial attribute transfer guided by exemplars.

5.3 Hairstyle editing

In this application, we further fine-tuned our trained EXE-GAN with extra hand-drawn-like sketches which were produced automatically with a pencil-sketch filter [42]. The application allows users to sketch in the mask to indicate roughly the hair styles. Given a masked image with sketches and an exemplar image, our EXE-GAN produces a style edited output. As shown in Fig. 15, various hairstyle editing results are obtained by changing the user-edited style sketches. It is easy even for a novice to obtain various styles by simple sketch editing.

5.4 Guided facial image recovery

The guided facial image recovery for occluded portrait eyeglasses and masks was implemented by masking them out and taking a different image from the same person as exemplar. As shown in Fig. 16, we compared our method to Lyu et al.'s eyeglasses removal method [57] on tinted eyeglasses (leftmost), sunglasses (mid-left), and myopia glasses (mid-right). The results show that our

guided facial image recovery method performs well on removing tinted eyeglasses and sunglasses which may fail with Lyu et al.'s method. As shown in Fig. 17, our method can also recover faces from occluded masks effectively.

5.5 Inherent stochasticity

Our method can be extended to produce multiple diverse facial inpainting results for an input masked facial image and an exemplar image by leveraging the inherent stochasticity. Users can easily select the preferred one among these results. The inherent stochasticity is achieved by adding per-pixel noise after each convolutional layer, leveraging the injected random latent code, and applying a truncation trick to tune the stochastic style representations [37], [38]. Fig. 18 shows a variety of facial inpainting results with various random latent codes.

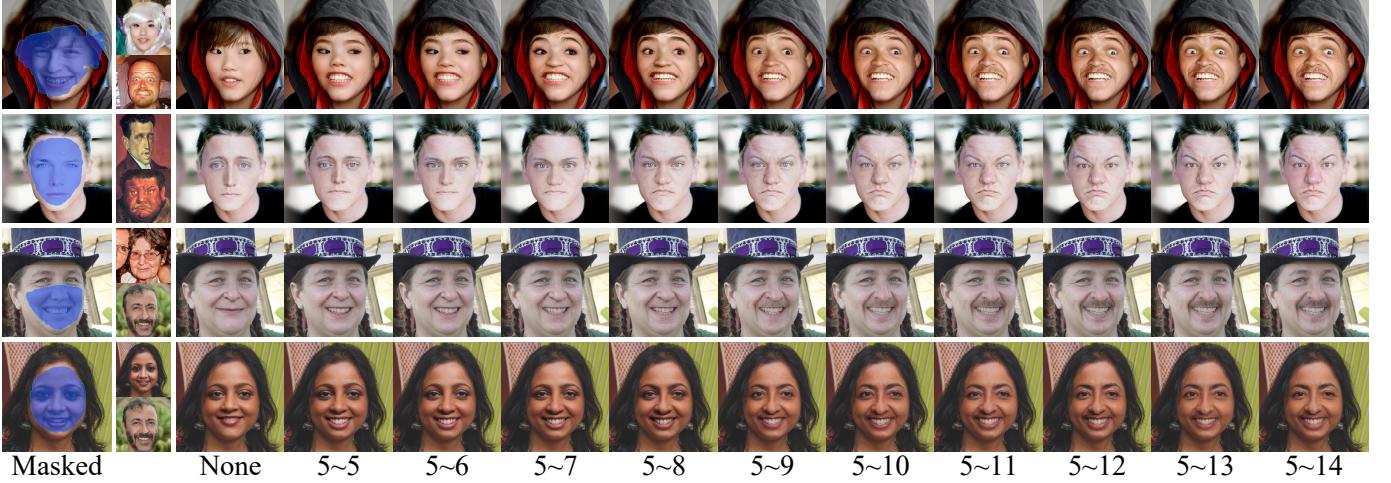


Fig. 14. Examples of guided facial style mixing (from left to right in each group): masked image, pairs of exemplars, and style-mixing effects. In each row, values of the style code of the first exemplar from i -th to j -th layers are replaced by those of the second exemplar.

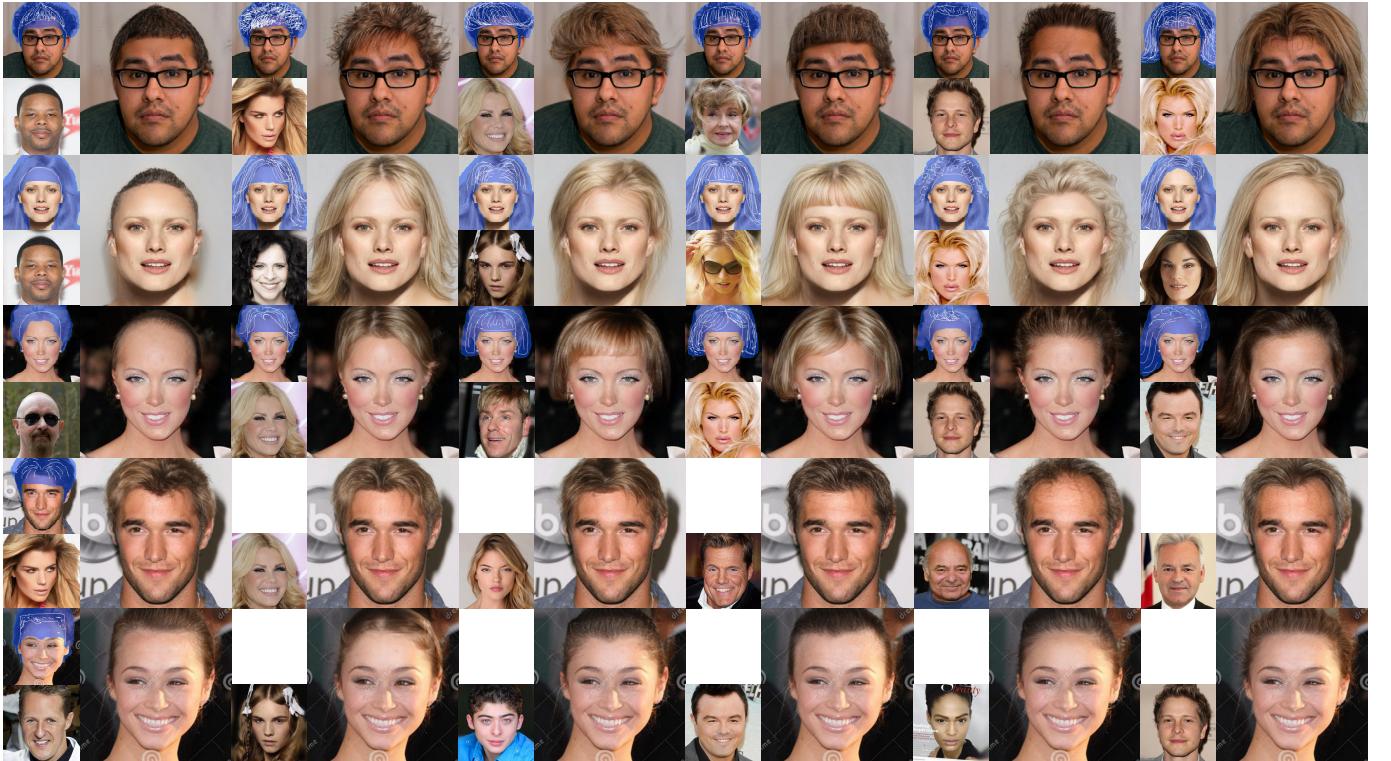


Fig. 15. Examples of hairstyle editing: (top three rows) edited results with different sketches guided by different exemplars and (last two rows) edited results with the same sketch guided by different exemplars.

6 CONCLUSION, LIMITATIONS, AND FUTURE WORK

In this paper, we have presented a novel interactive framework for realistic facial inpainting by taking advantages of exemplar facial attributes. An attribute similarity metric was introduced to help the generative network learn the style of facial attributes from the exemplar. We further proposed a novel spatial variant gradient backpropagation technique to address the issue of visual inconsistency on the filling boundary. Extensive

experiments and applications have demonstrated the effectiveness of the proposed method.

Our method has some limitations. Using the embedded style codes, we successfully transfer the facial attribute styles from the exemplar image. The explicit mapping between the facial attribute and the embedded style codes, on the other hand, is still unknown [42]. Incorporating a more advanced embedding algorithm into our pipeline could be a good next step. The trained model works well for aligned images because the facial



Fig. 16. Comparison on portrait eyeglasses removal (from top left to right in each group): masked image, exemplar image, our recovered result, and Lyu et al.'s result [57].

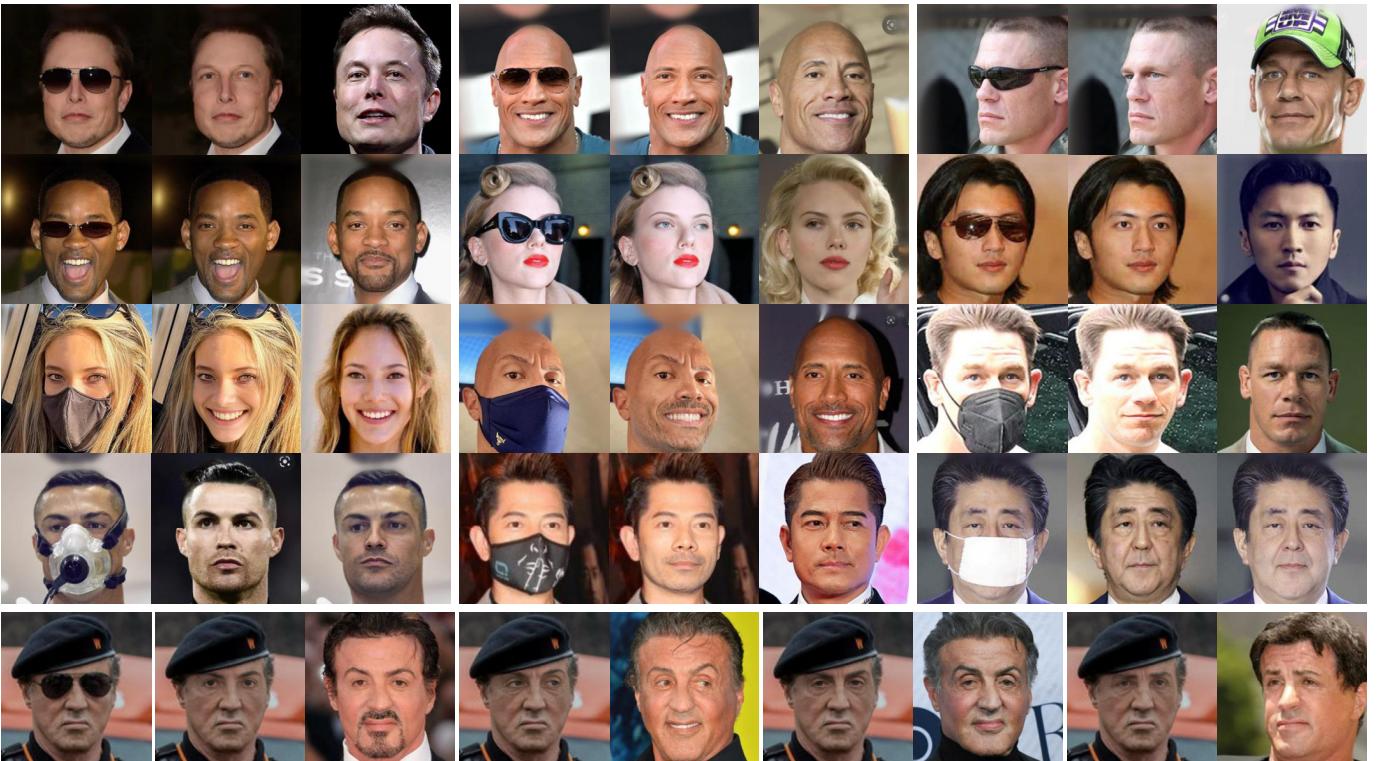


Fig. 17. More examples of guided facial image recovery: (top four rows, from left to right in each group) occluded image, recovered face image, and exemplar; (bottom row) diverse recovered results guided by different exemplars.

images in the experimented training datasets [53], [37] are highly aligned. It is necessary to align and crop the inputs before inpainting nonaligned images. It would be preferable to train models on unstructured datasets to create a more sophisticated algorithm.

REFERENCES

- [1] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington: IEEE, 2020, pp. 5548–5557.
- [2] A. Chen, R. Liu, L. Xie, Z. Chen, H. Su, and J. Yu, "Sofgan: A portrait image generator with dynamic styling," *ACM Transactions on Graphics*, vol. 41, no. 1, p. Article 1, 2022.
- [3] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington: IEEE, 2020, pp. 8185–8194.
- [4] X. Li, S. Zhang, J. Hu, L. Cao, X. Hong, X. Mao, F. Huang, Y. Wu, and R. Ji, "Image-to-image translation via hierarchical style disentanglement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington: IEEE, 2021, pp. 8639–8648.
- [5] L. Zhao, Q. Mo, S. Lin, Z. Wang, Z. Zuo, H. Chen, W. Xing, and D. Lu, "Uctgan: Diverse image inpainting based on unsupervised cross-space translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington: IEEE, 2020, pp. 5740–5749.
- [6] S. Zhao, J. Cui, Y. Sheng, Y. Dong, X. Liang, E. I. Chang, and Y. Xu, "Large scale image completion via co-modulated generative adversarial networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [7] Y. Yang and X. Guo, "Generative landmark guided face inpainting," in *Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Cham: Springer, 2020, pp. 14–26.
- [8] Y. Jo and J. Park, "Sc-fegan: Face editing generative adversarial network with user's sketch and color," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Washington: IEEE, 2019, pp. 1745–1753.
- [9] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Washington: IEEE, 2019, pp. 4470–4479.
- [10] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edge-connect: Structure guided image inpainting using edge prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. Washington: IEEE, 2019, pp. 3265–3274.
- [11] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proceedings of the 27th Annual Conference on Computer*

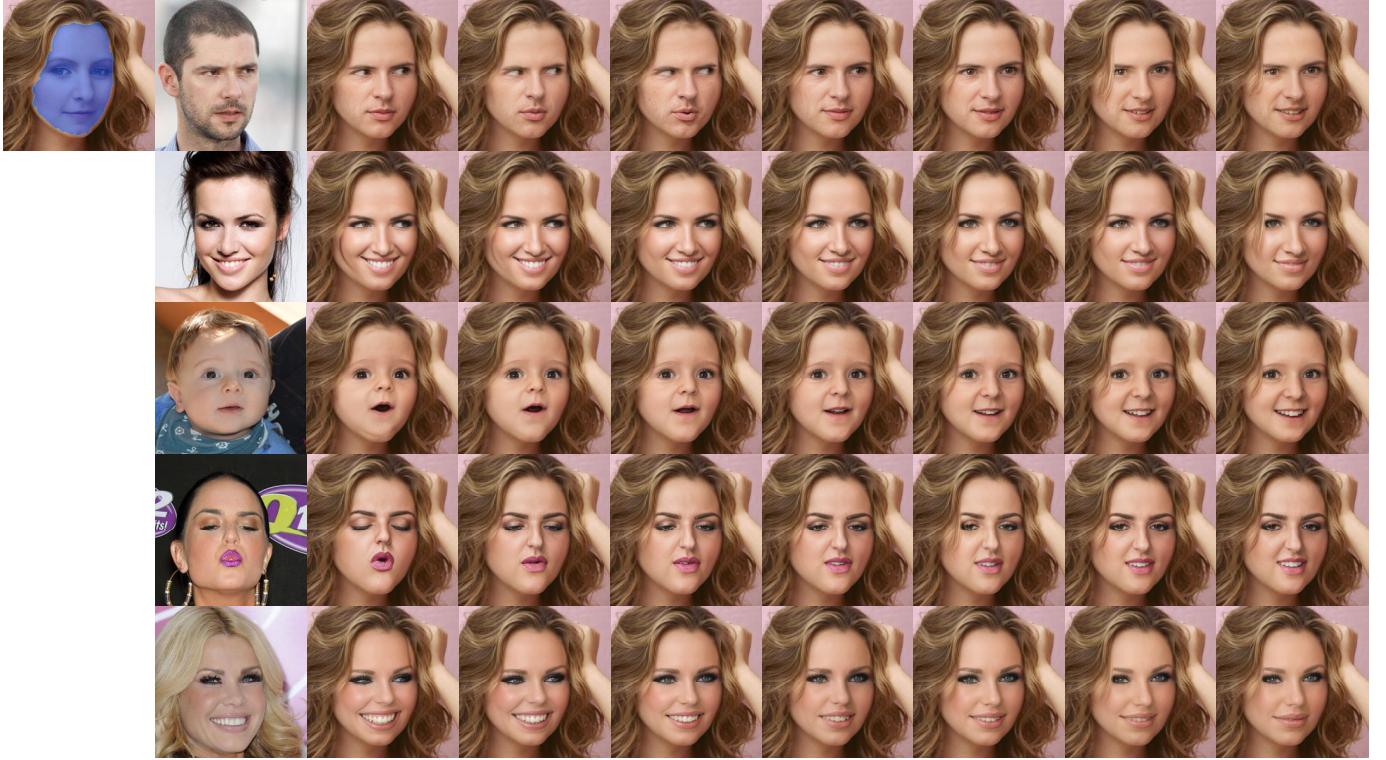


Fig. 18. Examples of diverse facial inpainting with inherent stochasticity. Based on the same masked image, we use different exemplars to guide the generation of various results.

- Graphics and Interactive Techniques (SIGGRAPH)*. New York: ACM, 2000, pp. 417–424.
- [12] Levin, Zomet, and Weiss, “Learning how to inpaint from global image statistics,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, vol. 1. Washington: IEEE, 2003, pp. 305–312.
 - [13] V. Kwatra, I. Essa, A. Bobick, and N. Kwatra, “Texture optimization for example-based synthesis,” *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 795–802, 2005.
 - [14] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, “Patchmatch: A randomized correspondence algorithm for structural image editing,” *ACM Transactions on Graphics*, vol. 28, no. 3, p. Article 24, 2009.
 - [15] H. Zhao, H. Guo, X. Jin, J. Shen, X. Mao, and J. Liu, “Parallel and efficient approximate nearest patch matching for image editing applications,” *Neurocomputing*, vol. 305, pp. 39–50, 2018.
 - [16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, vol. 27. Cambridge, MA.: MIT Press, 2014, pp. 2672–2680.
 - [17] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Globally and locally consistent image completion,” *ACM Transactions on Graphics*, vol. 36, no. 4, p. Article 107, 2017.
 - [18] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, “Generative image inpainting with contextual attention,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington: IEEE, 2018, pp. 5505–5514.
 - [19] H. Liu, B. Jiang, Y. Xiao, and C. Yang, “Coherent semantic attention for image inpainting,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Washington: IEEE, 2019, pp. 4169–4178.
 - [20] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, X. Liu, S. Wen, and E. Ding, “Image inpainting with learnable bidirectional attention maps,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Washington: IEEE, 2019, pp. 8857–8866.
 - [21] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, “Image inpainting via generative multi-column convolutional neural networks,” in *Advances in Neural Information Processing Systems*. MIT Press, 2018, pp. 331–340.
 - [22] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, “Image inpainting for irregular holes using partial convolutions,” in *Proceedings of the European Conference on Computer Vision (ECCV)*. Cham: Springer International Publishing, 2018, pp. 89–105.
 - [23] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, “Contextual residual aggregation for ultra high-resolution image inpainting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington: IEEE, 2020, pp. 7508–7517.
 - [24] C. Zheng, T.-J. Cham, and J. Cai, “Pluralistic image completion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Washington: IEEE, 2019, pp. 1438–1447.
 - [25] Z. Wan, J. Zhang, D. Chen, and J. Liao, “High-fidelity pluralistic image completion with transformers,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Washington: IEEE, 2021, pp. 4672–4681.
 - [26] Y. Yu, F. Zhan, R. Wu, J. Pan, K. Cui, S. Lu, F. Ma, X. Xie, and C. Miao, “Diverse image inpainting with bidirectional and autoregressive transformers,” in *Proceedings of the ACM International Conference on Multimedia (MM)*. New York: ACM, 2021, pp. 69–78.
 - [27] Q. Liu, Z. Tan, D. Chen, Q. Chu, X. Dai, Y. Chen, M. Liu, L. Yuan, and N. Yu, “Reduce information loss in transformers for pluralistic image inpainting,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington: IEEE, 2022, pp. 11347–11357.
 - [28] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, “Structure-flow: Image inpainting via structure-aware appearance flow,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Washington: IEEE, 2019, pp. 181–190.
 - [29] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo, “Foreground-aware image inpainting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Washington: IEEE, 2019, pp. 5833–5841.
 - [30] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “Stargan: Unified generative adversarial networks for multi-domain image-to-image translation,” in *Proceedings of the IEEE/CVF Conference*

- on Computer Vision and Pattern Recognition (CVPR).* Washington: IEEE, 2018, pp. 8789–8797.
- [31] T. Xiao, J. Hong, and J. Ma, “Elegant: Exchanging latent encodings with gan for transferring multiple face attributes,” in *Proceedings of the European Conference on Computer Vision (ECCV).* Cham: Springer International Publishing, 2018, pp. 172–187.
- [32] J. Guo, Z. Qian, Z. Zhou, and Y. Liu, “Mulgan: Facial attribute editing by exemplar,” *arXiv preprint*, vol. arXiv:1912.12396, 2019.
- [33] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, “Faceshifter: Towards high fidelity and occlusion aware face swapping,” *arXiv preprint*, vol. arXiv:1912.13457, 2019.
- [34] R. Chen, X. Chen, B. Ni, and Y. Ge, “Simswap: An efficient framework for high fidelity face swapping,” in *Proceedings of the 28th ACM International Conference on Multimedia (MM’20).* New York: ACM, 2020, pp. 2003–2011.
- [35] X. Liu, Y. Yang, and P. Hall, “Learning to warp for style transfer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* Washington: IEEE, 2021, pp. 3701–3710.
- [36] H. Kim, Y. Choi, J. Kim, S. Yoo, and Y. Uh, “Exploiting spatial dimensions of latent in gan for real-time image editing,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Washington: IEEE, 2021, pp. 852–861.
- [37] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, 2021.
- [38] 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).* Washington: IEEE, 2020, pp. 8107–8116.
- [39] R. Abdal, Y. Qin, and P. Wonka, “Image2stylegan: How to embed images into the stylegan latent space?” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* Washington: IEEE, 2019, pp. 4431–4440.
- [40] ———, “Image2stylegan++: How to edit the embedded images?” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* Washington: IEEE, 2020, pp. 8293–8302.
- [41] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, “In-domain gan inversion for real image editing,” in *Proceedings of the European Conference on Computer Vision (ECCV).* Cham: Springer International Publishing, 2020, pp. 592–608.
- [42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* Washington: IEEE, 2021, pp. 2287–2296.
- [43] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, “Designing an encoder for stylegan image manipulation,” *ACM Transactions on Graphics*, vol. 40, no. 4, p. Article 133, 2021.
- [44] Y. Wu, Y. Yang, Q. Xiao, and X. Jin, “Coarse-to-fine: facial structure editing of portrait images via latent space classifications,” *ACM Transactions on Graphics*, vol. 40, no. 4, p. Article 46, 2021.
- [45] Y. Wu, Y. Yang, and X. Jin, “Hairmapper: Removing hair from portraits using gans,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* Washington: IEEE, 2022, pp. 4227–4236.
- [46] J. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, “Generative visual manipulation on the natural image manifold,” in *Proceedings of the European Conference on Computer Vision (ECCV).* Cham: Springer International Publishing, 2016, pp. 597–613.
- [47] E. Collins, R. Bala, B. Price, and S. Süsstrunk, “Editing in style: Uncovering the local semantics of gans,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 2020, pp. 5770–5779.
- [48] O. Kafri, O. Patashnik, Y. Alaluf, and D. Cohen-Or, “Stylefusion: Disentangling spatial segments in stylegan-generated images,” *ACM Transactions on Graphics*, 2022.
- [49] L. Mescheder, S. Nowozin, and A. Geiger, “Which training methods for gans do actually converge?” in *Proceedings of the International Conference on Machine Learning (ICML),* vol. 80. PMLR, 2018, pp. 3481–3490.
- [50] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* Washington: IEEE, 2019, pp. 4685–4694.
- [51] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* Washington: IEEE, 2018, pp. 586–595.
- [52] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proceedings of the 3rd International Conference on Learning Representations (ICLR),* 2015.
- [53] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” in *Proceedings of the International Conference on Learning Representations (ICLR),* 2018.
- [54] 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*, vol. 30. MIT Press, 2017, pp. 6629–6640.
- [55] MMEditioning Contributors, “MMEditioning: OpenMMLab image and video editing toolbox,” <https://github.com/open-mmlab/mmediting>, 2022.
- [56] A. Bulat and G. Tzimiropoulos, “How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks),” in *Proceedings of the International Conference on Computer Vision (ICCV).* Washington: IEEE, 2017, pp. 1021–1030.
- [57] J. Lyu, Z. Wang, and F. Xu, “Portrait eyeglasses and shadow removal by leveraging 3d synthetic data,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* Washington: IEEE, 2022, pp. 3429–3439.