

Do inpainting yourself: Generative facial inpainting guided by exemplars

Wanglong Lu^{a,b}, Hanli Zhao^{a,*}, Xianta Jiang^b, Xiaogang Jin^c, Yong-Liang Yang^d, Kaijie Shi^b

^a College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou, 325000, China

^b Department of Computer Science, Memorial University of Newfoundland, St. John's, NL A1B 3X5, Canada

^c State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, 310000, China

^d Department of Computer Science, University of Bath, Bath, BA2 7AY, United Kingdom

ARTICLE INFO

Communicated by L. Celona

Keywords:

Generative adversarial networks

Image generation

Image inpainting

Facial image inpainting

ABSTRACT

We present EXE-GAN, a novel exemplar-guided facial inpainting framework using generative adversarial networks. Our approach not only preserves the quality of the input facial image but also completes 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 the exemplar image. We introduce a novel attribute similarity metric to encourage the 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. We extensively evaluate EXE-GAN on public CelebA-HQ and FFHQ datasets with practical applications, which demonstrates the superior visual quality of facial inpainting. The source code is available at <https://github.com/LonglongaaaGo/EXE-GAN>.

1. Introduction

Faces are widely recognized as the most representative and expressive aspect of human beings [1–3]. 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 makeup 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 [4,5], semantics [6], and exemplars [7]. 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 [8]. 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 [9,10] attempt to generate diverse image inpainting results allowing users to select a desired one. However, they cannot complete missing regions with user guidance. Many recent methods try to employ additional landmarks [11], strokes [12], or sketches [13,14] to guide the inpainting of facial structures and attributes. However, these methods tend to overfit the resulting images with this 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 the input image, the stochastic style generated from the random latent code, and the exemplar style of the exemplar image to generate highly 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 regions, we devise a novel spatial

* Corresponding author.

E-mail address: hanlizhao@wzu.edu.cn (H. Zhao).



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 locally 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.

variant gradient backpropagation method for the network training. We compare our method to the state-of-the-art (SOTA) 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 with facial attributes guided by exemplars.
- A novel 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 [15,16] and patch-based methods [17–19] 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 models (e.g., adversarial generative networks (GANs) [20] or diffusion probabilistic models [21]) to achieve semantic completion, such as completion with global and local discriminators [22], reconstructive sampling [23], structure and texture fusion [3], attention-based inpainting [1,8,24–28], inpainting via multi-column CNNs [29], multi-scale CNNs [2], Transformer-based inpainting [30–33], light-weight models [34], data-efficient inpainting [35], inpainting for irregular holes [13,36], high-resolution inpainting [37], and inpainting using diffusion models [38–40]. Some recent works [9,10,30,31,41–43] try to generate pluralistic inpainted images without guidance information. By taking advantage of auxiliary information, such as landmarks [11], sketches [12–14,44], and geometries [45,46], 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. Some closely related reference-guided methods [47,48] perform image inpainting by directly inputting a reference image alongside the masked image or inserting high-level features of the guidance images [49,50]. Different from these methods, our method can effectively generate realistic inpainted face images guided by an exemplar image containing rich textural and semantic information and works well on different poses/lighting conditions.

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 [6,51]. 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 [4,5,52]. 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 [7,53–57]. Nevertheless, these methods have a common limitation as users cannot flexibly select facial regions for the local transfer of facial attributes. Re-composition [58] and StyleMapGAN [59] leverage intermediate extracted representations to perform image re-composition or 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 with natural attribute transitions while remaining unaltered regions.

Our work is also closely related to *image embedding* which enables image synthesis from latent space [60–62]. StyleGANs [61,62] and SemanticStyleGAN [63] 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 [64,65] and encoder-based embedding [66–71] methods perform image manipulation by inverting an image to the latent space [72]. Recently, EditingInStyle [73] and StyleFusion [74] show the impressive performance of image editing by semantically manipulating the style latent space. Although image embedding has a strong capability in presenting image styles, these methods may change unedited regions because of information losses in inversion process. For instance, Richardson et al. [67] presented results of inpainting using a pixel2style2pixel framework but failed to preserve the 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.

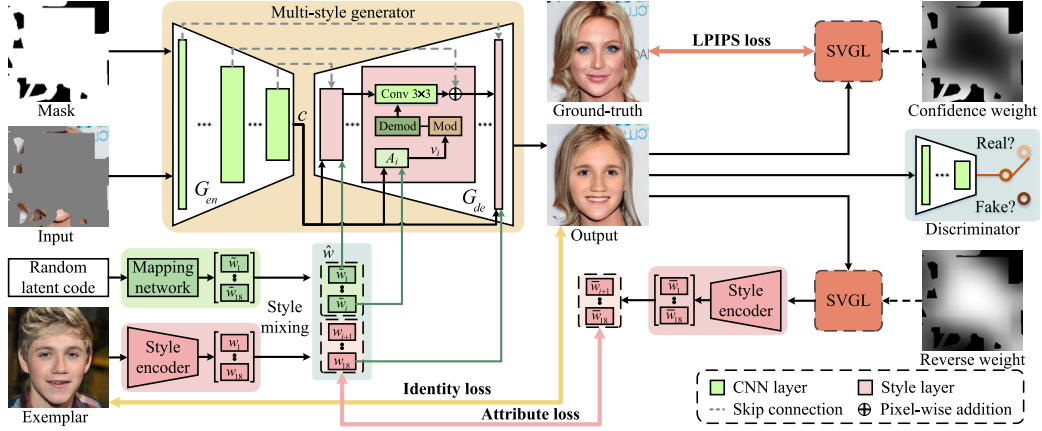


Fig. 2. Overview of our EXE-GAN framework. We employ style mixing on stochastic and exemplar style codes, and modulate them with the global style code of input image into the multi-style generator for facial inpainting. The adversarial, identity, LPIPS, and attribute losses are integrated into the overall training objective. Spatial variant gradient layers (SVGL) are utilized for natural transition across the filling boundary. Once trained, EXE-GAN can be applied to various application scenarios, such as local facial attribute transfer, guided facial style mixing, hairstyle editing, and guided facial image recovery.

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 linearly maps a random latent code $z \in \mathbb{R}^{512 \times 1}$ to a stochastic style code $\tilde{w} = \{\tilde{w}_i \in \mathbb{R}^{512 \times 1} | i \in T\} \in \tilde{W}+$, where $\tilde{W}+$ 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 style encoder E directly maps an image to a disentangled style latent space $W+$. Given the 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 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 propose a multi-style generator by also incorporating 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 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. Coarse-resolution layers correspond to high-level facial attributes and fine-resolution layers correspond to 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 [62]:

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

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 [62]. 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

EXE-GAN 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 [20] with R_1 regularization [75]. Specifically, the adversarial objective is defined as:

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

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 [76] for face recognition.

3.3.3. LPIPS loss

We employ the Learned Perceptual Image Patch Similarity (LPIPS) loss [77] 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 [78] 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:

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

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

3.3.5. Total objective

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:

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

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 the input image, and the visual contents should be naturally transitioned on the boundary. 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 [29] 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 (e.g., attribute loss) are defined in the embedding space and the 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-function’’ $P(x, M_x)$. The forward-propagation and backpropagation behaviors of SVGL are defined below:

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

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 an SVGL $P(\cdot, M_w)$ for the spatial variant LPIPS loss and an 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. 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.

Algorithm 1 Training procedure of EXE-GAN

```

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

```

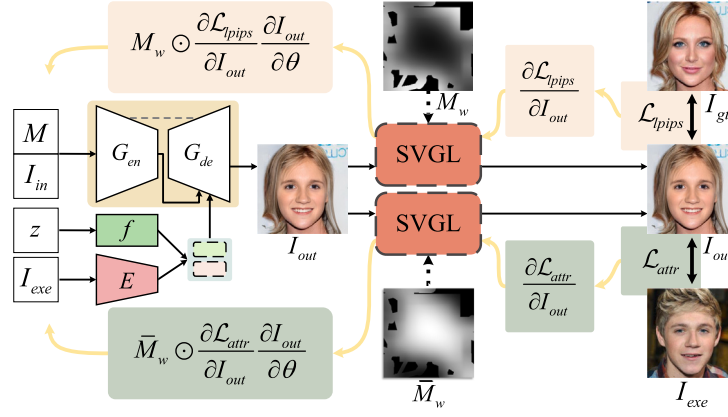


Fig. 3. Illustration of the SVGL on LPIPS and attribute losses. In forward-propagation, SVGL does not change any information for I_{out} . In backpropagation, gradients are re-weighted based on the spatial variant M_w and $\bar{M}_w$, respectively.

3.5. 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. We employed the mapping network f and discriminator D architectures from StyleGANv2 [62], and borrowed the pre-trained the style encoder E from pSp [67]. We kept E frozen for training f , G and D . All weights of f , G , and D are trained from scratch. Following the settings of StyleGANv2 [62], 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 [62] with a probability of 0.5 was employed to generate stochastic style codes during training. For computing the confidence weight mask in SVGL, the Gaussian kernel is with size 64×64 and its standard deviation is 40. 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 [13] 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.

4. Experiments

4.1. Settings

Experiments were conducted on two publicly available face image datasets CelebA-HQ [79] and FFHQ [61]. For CelebA-HQ [79], we randomly selected 28,000 images for training and remained 2,000 images for testing. For FFHQ [61], we randomly selected 60,000 images for training and the rest 10,000 images for testing. Each image was resized to 256×256 . We trained the networks for 800,000 iterations with batch size of 8. Unless specified, all experiments were conducted on the NVIDIA Tesla V100 GPU. The training time was around three weeks.

The performance was quantitatively evaluated using the Fréchet inception distance (FID) [80] and the paired/unpaired inception discriminative score (P-IDS/U-IDS) [10]. 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.

4.2. Comparisons

4.2.1. Comparison to free-form inpainting

We compared EXE-GAN on CelebA-HQ and FFHQ datasets to the SOTA free-form inpainting methods, including Contextual Attention (CA) [24], Partial Convolutions (PConv) [36], Globally & Locally (G&L) [22], Gated Convolution (GConv) [13], EdgeConnect [14], GM-CNN [29], CMOD [10], ZITS [81], TFill [82], Stable Diffusion [38], RGTSI [83], and TransRef. [84].

We used the publicly available MMEediting framework [85] for Contextual Attention (CA) [24], Partial Convolutions (PConv) [36], Globally & Locally (G&L) [22], and Gated Convolution (GConv) [13]. MMEediting is an open-source image and video editing toolbox based on PyTorch. We used the official codes for EdgeConnect [14], GM-CNN [29], ZITS [81], TFill [82], Stable Diffusion [38], RGTSI [83], and TransRef. [84]. For CMOD [10], we used the official TensorFlow-based code to implement a PyTorch-based version. Except for Stable Diffusion [38], all the compared free-form inpainting models were trained using CelebA-HQ and FFHQ datasets, respectively. We used the pre-trained Stable Diffusion [38] sd-v1.5-inpainting model for inference. The positive prompts included “best quality, extremely detailed, real human face, human face, high fidelity face, high-quality face”. To avoid low-quality images, we also employed negative prompts such as “bad anatomy, bad hands, missing fingers, extra digit, fewer digits, cropped, worst quality, low quality”. For reference-guided methods, RGTSI [83] and TransRef. [84], input-reference pairs with high similarity are crucial for inpainting results. However, the CelebA-HQ and FFHQ datasets lack paired ground-truth and exemplar images with the same attributes (identity, expression, decorative goods, etc.). To address this, we adapted the training strategy by copying the ground-truth image and applying random color jittering, translation, flipping, and bilinear scaling operations [86] to create the exemplar image for each iteration. This approach ensures effective training and reasonable results, aligning with the strategies used by RGTSI and TransRef for natural images. Our EXE-GAN does not require paired data due to the proposed self-supervised attribute similarity metric. We randomly chose exemplar images during training and used a reversed batch of input images to guide the inpainting during testing. For a fair comparison, we used the same training/testing splits for all experiments.

Fig. 4 shows the quantitative performance comparisons between our method and the SOTA free-form methods on CelebA-HQ and FFHQ datasets. For easier comparison, we also show the corresponding FID scores in Table 1. To simulate various scenarios, we used a fixed center rectangle mask of 128×128 and irregular mask templates from PConv [36] for mask ratios ranging from 10%–20% to 50%–60%. For evaluating large-scale mask inpainting, we generated mask templates using the mask generation procedure of CMOD [10] for ratios

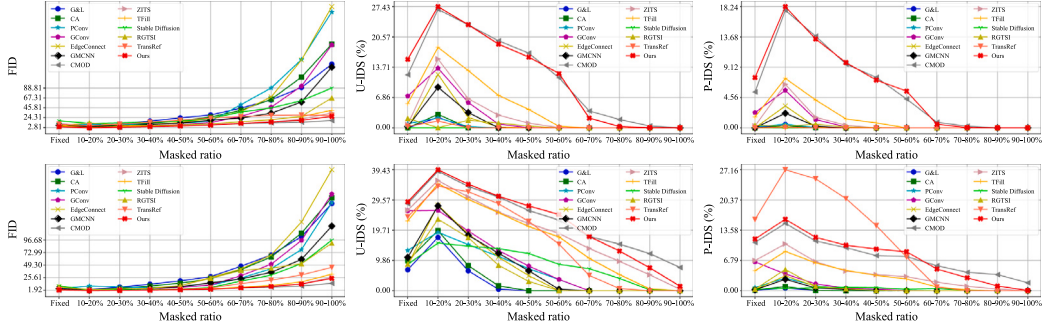


Fig. 4. Quantitative comparisons of our method to SOTA free-form inpainting methods on the CelebA-HQ (top) and FFHQ (bottom) datasets.

Table 1

FID scores for our method and SOTA’s free-form inpainting methods on the CelebA-HQ (top) and FFHQ (bottom) datasets. The best is **bolded**, while the second-best is underlined.

Methods	Fixed	10–20%	20–30%	30–40%	40–50%	50–60%	60–70%	70–80%	80–90%	90–100%
G&L	11.15	5.26	10.62	17.02	23.76	30.22	45.17	63.06	90.06	140.83
CA	7.12	4.70	8.23	12.36	17.50	25.81	40.35	64.27	112.61	184.60
PConv	6.23	10.00	9.44	10.16	12.51	16.60	52.59	89.00	151.03	254.59
GConv	<u>4.46</u>	3.46	5.45	7.12	9.52	13.61	27.08	47.17	93.10	182.71
EdgeConnect	9.53	3.96	6.76	9.84	14.07	22.51	37.53	69.44	150.17	266.11
GMCNN	7.46	4.47	6.63	9.67	12.60	18.75	23.58	34.11	58.48	134.48
CMOD	4.99	2.35	3.79	<u>5.21</u>	<u>6.82</u>	<u>8.64</u>	11.15	12.87	14.53	18.12
ZITS	4.31	2.81	4.42	5.71	7.26	9.11	11.94	15.34	21.07	34.73
TFill	6.48	3.28	5.27	6.69	8.68	11.04	15.44	20.11	26.66	40.58
Stable Diffusion	17.67	11.95	12.98	14.57	16.85	22.35	36.72	45.43	61.38	88.80
RGTSI	6.03	13.20	8.09	7.40	8.71	11.13	15.35	19.75	27.32	67.33
TransRef	10.39	6.71	10.35	14.61	18.98	23.18	26.69	29.18	30.01	29.10
Ours	4.66	2.10	3.50	4.80	6.22	8.16	<u>11.86</u>	<u>14.78</u>	<u>18.71</u>	<u>27.15</u>
G&L	6.45	3.26	7.67	13.05	19.47	27.36	47.16	68.56	103.23	166.28
CA	4.19	2.46	5.52	9.70	15.06	23.49	40.67	64.71	110.03	177.22
PConv	4.35	9.84	7.95	6.87	8.50	10.92	24.66	42.43	78.97	167.49
GConv	1.92	2.10	3.08	4.73	7.19	11.42	29.29	50.96	96.21	184.05
EdgeConnect	7.92	1.68	3.33	5.83	10.43	25.75	41.80	68.52	131.25	230.16
GMCNN	4.68	1.81	3.45	5.46	8.53	15.56	23.15	35.63	60.19	123.32
CMOD	2.19	0.82	1.41	<u>2.02</u>	<u>2.70</u>	<u>3.76</u>	6.04	7.57	9.66	14.85
ZITS	<u>2.09</u>	0.95	1.60	2.33	3.06	4.32	6.56	9.25	13.95	26.09
TFill	<u>1.92</u>	1.09	1.74	2.43	3.27	4.32	7.42	10.88	17.45	32.12
Stable Diffusion	10.03	1.98	2.38	2.89	3.77	6.87	18.97	30.33	52.16	96.67
RGTSI	7.78	3.02	4.35	5.21	10.44	24.14	39.09	43.55	51.95	90.94
TransRef	2.84	1.35	1.63	2.24	3.30	6.04	14.37	20.77	30.60	45.17
Ours	2.22	0.75	1.30	1.88	2.50	3.57	<u>6.27</u>	<u>8.77</u>	<u>13.44</u>	<u>24.18</u>

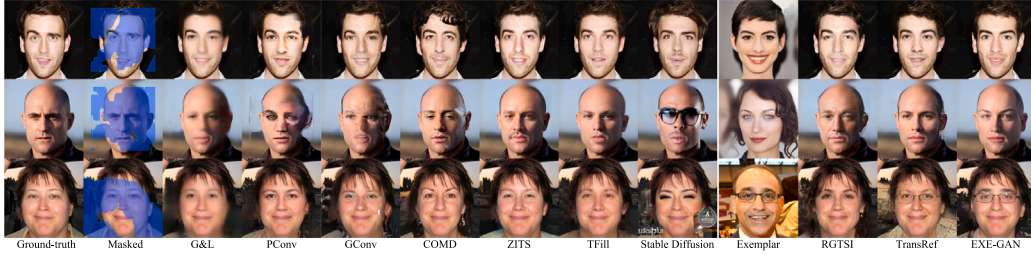


Fig. 5. A qualitative comparison of our method to current free-form inpainting methods.

ranging from 60%–70% to 90%–100%. Each range includes 2,000 mask templates. Quantitative results show that our method can compete with SOTA methods, despite the fact that its inpainting was guided by exemplars, which was thought to be difficult to maintain image quality [10]. Furthermore, when the masked ratios increase and the useful information from unmasked regions decreases, G&L [22], PConv [36], GConv [13], EdgeConnect [14], and GMCNN [29] often fail to generate high-fidelity results due to their limited generative capacity. For large-scale masks, ZITS [81], TFill [82], and CMOD [10] perform better thanks to their more powerful generative capabilities, but they still have performance limitations. Stable Diffusion [38] has the capability to produce high-quality inpainted results by benefiting from extremely

large-scale training data. However, since the highly diverse results may not align with the ground-truth images, the quantitative scores could be affected. RGTSI [83] and TransRef. [84] rely on exemplar facial features to guide inpainting, but may struggle when exemplars and ground truths differ in poses, lighting, or extreme styles. Exemplar-guided methods are expected to extract useful exemplar facial features and precisely fill the corresponding facial features to align with the inpainting boundaries. On FFHQ and CelebA-HQ with various masks in terms of FID, U-IDS, and P-IDS, our EXE-GAN achieves overall better quantitative performance than exemplar-guided methods and competitive performance than SOTA free-form inpainting methods.

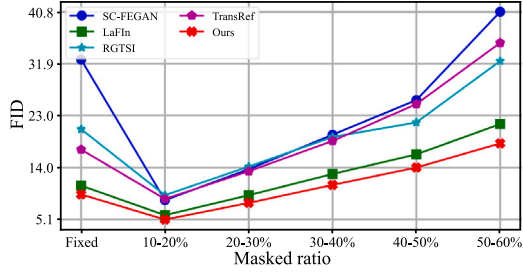


Fig. 6. On the Celeba-HQ dataset, we performed a quantitative comparison of our method to SOTA guidance-based facial inpainting methods.

Fig. 5 shows the quantitative performance of compared methods. Although all the methods can compatibly fill in the missing regions, G&L [22] tends to produce blurry inpainting while PConv [36] and GConv [13] fail to inpaint large-scale missing regions. CMOD [10], ZITS [81], and TFill [82] well handle textures and structures inpainting. However, with large masked regions, CMOD would produce a little artifact on masked boundaries (e.g., the first row). Some blurred contents could be found under close inspection in TFill’s results. Stable Diffusion [38] can control the inpainting contents via text prompts, but some facial features, such as identity, are difficult to describe using prompts. RGTSI [83] and TransRef. [84] produce plausible inpainted results. However, their results do not always accurately reflect exemplar facial features, such as the hazy glasses structure shown in the third row. Our EXE-GAN can generate high-quality inpainted results. Notably, the inpainting of facial attributes cannot be controlled with CMOD, ZITS, and TFill. In contrast, with the help of exemplar facial attributes, the inpainting of facial attributes with our EXE-GAN can be controlled easily.

4.2.2. Comparison to guidance-based inpainting

We compared EXE-GAN on Celeba-HQ dataset to the SOTA guidance-based facial inpainting methods, including sketch-and-color-based facial inpainting SC-FEGAN [12], landmark-based face inpainting LaFIn [11], reference-based inpainting RGTSI [83] and TransRef. [84]. For SC-FEGAN and LaFIn, we generated corresponding guided information from exemplar images and alternately took one facial image as the exemplar and the other one as the masked image for facial attribute inpainting. We used the exemplar image directly for inpainting in RGTSI, TransRef, and EXE-GAN. In addition, we used ground-truth structure images as input for RGTSI, following the methodology of the original paper.

In this experiment, we used the officially released pre-trained SC-FEGAN [12] and LaFIn [11] models, as well as the trained RGTSI and TransRef models described in Section 4.2.1. SC-FEGAN [12] uses sketches and color as 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 [11] relies on landmarks to fill missing regions. Therefore, we utilized the face alignment network FAN [87] 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 1100 inpainted images for each comparison method.

Fig. 6 shows the quantitative comparison of our EXE-GAN to SOTA guidance-based inpainting methods. 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. RGTSI and TransRef are primarily designed for inpainting images with natural scenes, with the assumption that each ground-truth image and its reference (exemplar) image are similar in terms of structure and texture. However, this assumption does not always hold for exemplar-guided facial image inpainting, particularly when the exemplars differ significantly in pose, expression, identity, or gender. In terms of the authenticity of inpainted images guided by exemplars, our method outperforms the compared methods.

As shown in Fig. 7, SC-FEGAN [12] 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 inpainted images with SC-FEGAN. LaFIn [11] generates facial expressions similar to exemplars but may fail to inpaint the decorative attributes, such as glasses and hairstyles. RGTSI [83] and TransRef. [84] can produce high-quality inpainted results, but closer inspection reveals some artifacts around the boundaries. In comparison, our method directly learns the style of facial attributes from the exemplar without extra input, and can produce more realistic facial inpainted results with exemplar-like facial attributes including decorative attributes.

User study. 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 [12], LaFIn [11], 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 [12], LaFIn [11], 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. We recruited 63 volunteers to subjectively evaluate the effectiveness of our method comprehensively. The results show that our EXE-GAN obtains the majority of votes (59.67%) compared to SC-FEGAN [12] (12.87%) and LaFIn [11] (27.46%). The user study validates that the output of our EXE-GAN is more realistic than the compared methods visually observed by subjects.

4.2.3. Comparison to facial attribute transfer

We compared our method on the Celeba-HQ dataset to the SOTA facial attribute transfer methods, including StarGANv2 [6], MaskGAN [4], SimSwap [55], RGTSI [83], TransRef. [84], DiffFace [88], and DiffSwap [89], where the same exemplar image was used to guide the attribute transfer.

In this experiment, we used trained RGTSI [83] and TransRef. [84] models, as described in Section 4.2.1. The pre-trained models of StarGANv2 [6], MaskGAN [4], SimSwap [55], DiffFace [88], and DiffSwap [89] provided in the official repository were used. For StarGANv2 [6], we set the input image as the “reference” image and the exemplar image as the “source” image. For MaskGAN [4], 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 [55], DiffFace [88], and DiffSwap [89] directly perform the exemplar-guided face synthesis with the input and exemplar images. RGTSI [83], TransRef. [84], and our EXE-GAN synthesize facial attributes for masked regions of input images guided by exemplar images.

As shown in Fig. 8, StarGANv2 [6] can transform an input image reflecting the identity of the exemplar. However, it leaves users little freedom to manipulate face images interactively. MaskGAN [6] 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

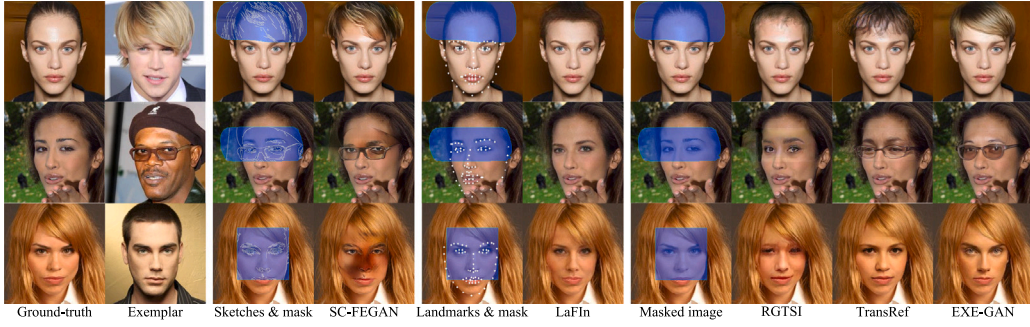


Fig. 7. A qualitative comparison of our method and the SOTA guidance-based facial inpainting methods.

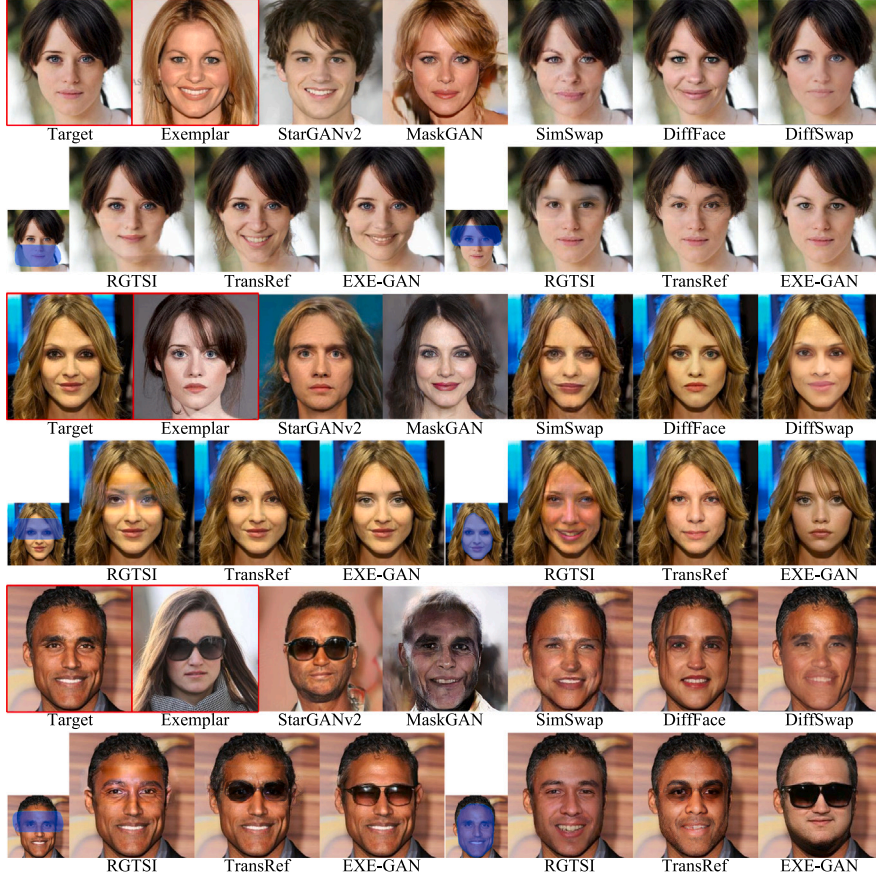


Fig. 8. A qualitative comparison of our method and the SOTA facial attribute transfer methods.

irrelevant changes to fine details in the background. SimSwap [55], DiffFace [88], and DiffSwap [89] can transfer the identity of the exemplar face to the input face and preserve the facial attributes of the input. However, they do not allow users to flexibly select regions for face editing and cannot successfully transfer decorative attributes, like sunglasses. RGTSI [83], TransRef. [84], and our method not only preserve the pixels of known regions but also allow more degrees of freedom to interactively perform facial attribute manipulation. However, because the exemplar facial features may conflict with those around mask boundaries, RGTSI and TransRef struggle to ensure a natural transition across the inpainted region boundaries. Thanks to our self-supervised attribute similarity metric for learning exemplar facial features and SVGL for enhancing natural transitions, EXE-GAN can produce high-quality results with facial attributes guided by exemplars, including gender, makeup style, hairstyle, and decorative style (e.g., glasses).

4.2.4. More comparisons to closely related works

More comparisons to closely related works are shown in Fig. 9. Although most methods produce plausible results, some visual artifacts can be seen in the details of IdentityPreserving [47] and GuidedInpainting [50], RefMatch [90], RGTSI [83], and TransRef. [84]. Color inconsistencies may occur in SemanticStyleGAN [63], ILVR [60], and RefMatch [90]. In addition, Image2StyleGAN++ [65], Re-composition [58], StyleMapGAN [59], SemanticStyleGAN, and ILVR introduce unwanted changes in background or unedited regions. RefMatch [90] hypothesizes that the exemplar and targeted inpainting images have very similar contexts, such as the same landmark building. The inpainting quality deteriorates when the exemplar exhibits different facial features, such as poses, identity, or styles, that violate assumptions, as discussed in RefMatch’s limitation section. On the other hand, our EXE-GAN can seamlessly fill in the masked pixels using exemplar-like attributes without changing unmasked areas, yielding high-quality inpainting results while avoiding the above artifacts.

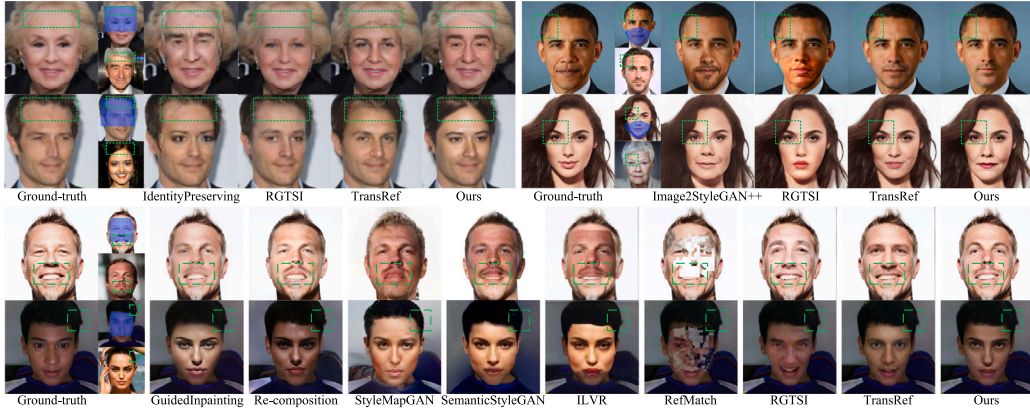


Fig. 9. Visual comparisons between our EXE-GAN and closely related works (small images in each group: masked and exemplar images). The results of IdentityPreserving [47] and Image2StyleGAN++ [65] are taken from their respective papers. The results of GuidedInpainting [50] and Re-composition [58] are from paper [58]. The results of StyleMapGAN [59], SemanticStyleGAN [63], ILVR [60], and RefMatch [90] are generated using their publicly available trained models. The results of RGTSI [83] and TransRef. [84] are generated from models re-trained from scratch.



Fig. 10. Visual results of EXE-GAN on images with low light and large pose variations (from top left to right in each group): masked image, exemplar image, inpainting result.

Table 2

Network complexity and computational efficiency of compared methods. Bold indicates the best.

Method	EdgeConnect	GMCNN	CMOD	TFill	ZITS	RGTSI	TransRef	EXE-GAN
# Parameters (M)	21.54	12.56	79.17	109.45	67.90	175.55	41.97	79.18
FLOPs (G)	122.64	67.20	90.25	45.45	182.73	146.67	7.55	90.26
FPS	73.59	121.58	77.79	43.55	11.55	34.56	81.60	77.62
Training time (H)	331.84	61.44	192.28	287.04	299.53	123.92	88.41	269.49

4.3. Performance of EXE-GAN on challenging cases

Fig. 10 illustrates the exemplar-guided inpainting of EXE-GAN on challenging cases, including low-light conditions (first row) and large pose variations (second row), using various masks. In the first row, the inpainted images retain the low-light characteristics of the masked inputs, but faces remain recognizable. Additionally, using a low-light image as an exemplar (4-th case), the inpainted result clearly exhibits the exemplar’s features while maintaining high fidelity. The second row shows that EXE-GAN effectively handles large pose variations. For instance, the second-last case demonstrates that EXE-GAN performs well even with extreme exemplar poses; the last case shows that despite using a completely different pose for inpainting, the resulting image maintains good quality.

4.4. Analysis on network complexity and computational efficiency

In Table 2, we show the model complexity and computational efficiency of our EXE-GAN and the SOTA methods in terms of the number of parameters, FLOPs, frames per second (FPS) in the inference phase, and training time (hours). All methods were evaluated on a machine with a single RTX 3090 GPU using images with the size of 256×256 . To estimate the training time, we first recorded the time taken for 1000 iterations on the CelebA-HQ dataset for each module of each method using their default hyperparameters. We then multiplied

this time by the total number of 1000-iteration chunks required and summed the times for all modules to get the estimated training time for each method. All the models show good efficiency and practical viability, with no more than 200.0M learnable parameters. Most models, including GMCNN, TransRef, and EXE-GAN, can perform real-time inference and process more than 70 images per second. For example, TransRef only needs 7.55 GFLOPs for each image, showing excellent computational efficiency with robust inpainting performance. While our model does not have the fewest parameters, the moderate capacity of our network is critical in achieving competitive results when learning to use the exemplar facial features. Furthermore, our competitive image inpainting performance on the CelebA-HQ and FFHQ datasets shows that EXE-GAN achieves a good balance between image inpainting quality and computational resources.

4.5. Ablation study

4.5.1. Ablation study on SVGL

We further investigated the effectiveness of SVGL in our EXE-GAN via an ablation study. The free-form mask sampling strategy mentioned in Section 3.5 was adopted. We present qualitative examples in Fig. 11(a-c) to express the visual effects of the ablation study. Table 3 shows the quantitative results. When SVGLs are removed from both the attribute and LPIPS losses, both quantitative measures and visual qualities drop dramatically, and completed images may not show clear

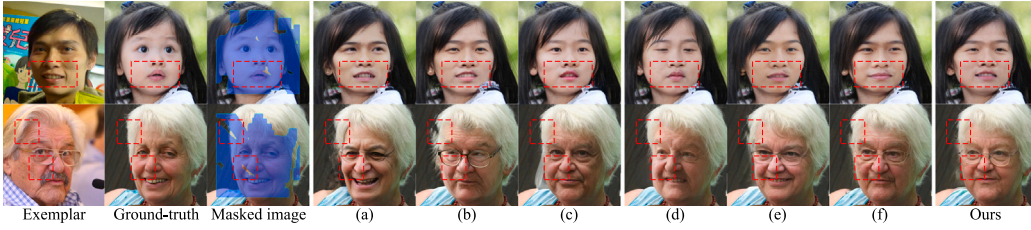


Fig. 11. Qualitative examples of the ablation study for large-scale facial inpainting by exemplars with (a) EXE-GAN without any SVGLs in $\mathcal{L}_{attr}$ and $\mathcal{L}_{lpijs}$, (b) EXE-GAN without SVGL in $\mathcal{L}_{attr}$, (c) EXE-GAN without SVGL in $\mathcal{L}_{lpijs}$, (d) EXE-GAN without $\mathcal{L}_{attr}$, (e) EXE-GAN without $\mathcal{L}_{lpijs}$, (f) EXE-GAN without $\mathcal{L}_{id}$, and (Ours) EXE-GAN.

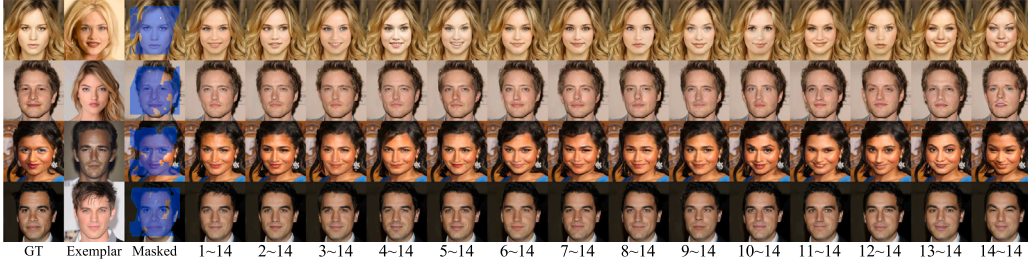


Fig. 12. Examples of facial inpainting with various subsets of the style codes: ground-truth, exemplar, masked image, and various style effects. In each row, values of the i th to j th layers in the style code are from the exemplar, and values of the remaining layers are from the stochastic style code.

Table 3

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

Method	CelebA-HQ			FFHQ		
	FID [↓]	U-IDS [↑]	P-IDS [↑]	FID [↓]	U-IDS [↑]	P-IDS [↑]
(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%

Table 4

Ablation study for large-scale facial inpainting by exemplars with (d) EXE-GAN without $\mathcal{L}_{attr}$, (e) EXE-GAN without $\mathcal{L}_{lpijs}$, and (Ours) EXE-GAN. Results are averaged over 5 runs. **Bold**: top-2 quantity.

Method	CelebA-HQ			FFHQ		
	FID [↓]	U-IDS [↑]	P-IDS [↑]	FID [↓]	U-IDS [↑]	P-IDS [↑]
(d)	9.714	11.38%	4.65%	4.453	23.84%	9.86%
(e)	10.207	7.37%	2.95%	4.726	23.50%	9.53%
Ours	9.967	9.175%	3.85%	4.353	24.33%	9.92%

manifestations of exemplar facial images in attributes, such as wearing glasses in Fig. 11(a). When replacing the SVGL-based attribute loss with the standard attribute loss without SVGL, there may be visible boundary inconsistencies in the generated results (see Fig. 11(b)), and the quantitative performance is also affected. When replacing the SVGL-based LPIPS loss with the standard LPIPS loss without SVGL, 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, as demonstrated in Fig. 11(c). 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. In comparison, our EXE-GAN is able to produce realistic facial images with facial attributes similar to exemplars while achieving competitive quantitative scores.

4.5.2. Ablation study on attribute loss, LPIPS loss, and identity loss

Another ablation study was carried out to investigate attribute loss and LPIPS loss in EXE-GAN. The free-form mask sampling strategy

mentioned in Section 3.5 was employed. Fig. 11(d-e) and Table 4 show visual comparisons as well as quantitative results. When the attribute loss in EXE-GAN is removed, the model produces comparable results but reduces the visual similarity of facial attributes between the generated result and the exemplar image, as shown in Fig. 11(d). When the LPIPS loss is removed, both the quantitative and qualitative measures decrease rapidly, and visual artifacts can be seen around the mask boundaries, as illustrated in Fig. 11(e). In contrast, EXE-GAN results demonstrate high fidelity with facial attributes similar to exemplars and competitive quantitative scores.

To further demonstrate the effectiveness of identity loss in preserving the identity information of exemplar images, we assessed the identity distance metric (IDD) in ArcFace’s feature space. We trained another model without $\mathcal{L}_{id}$ and tested it alongside our EXE-GAN on the CelebA-HQ test set with different masked ratios. Fig. 11(f) shows that without $\mathcal{L}_{id}$, the trained model can still reflect the exemplar attributes in the inpainted images with similar visual quality. However, the IDD similarity results in Table 5 indicate that the inpainted images of Model (f) lose some identity information of the given exemplar image. By incorporating identity loss, our EXE-GAN performs better in preserving the identity information guided by exemplars.

4.5.3. Ablation study on exemplar styles modulation

We performed ablation experiments on the modulation of exemplar styles by re-training our model with different vector ϕ configurations, and ϕ is a binary vector that indicates which style for each style layer is modulated. According to Fig. 12, the more exemplar style codes that are modulated, the more exemplar facial attributes will appear in the inpainted images.

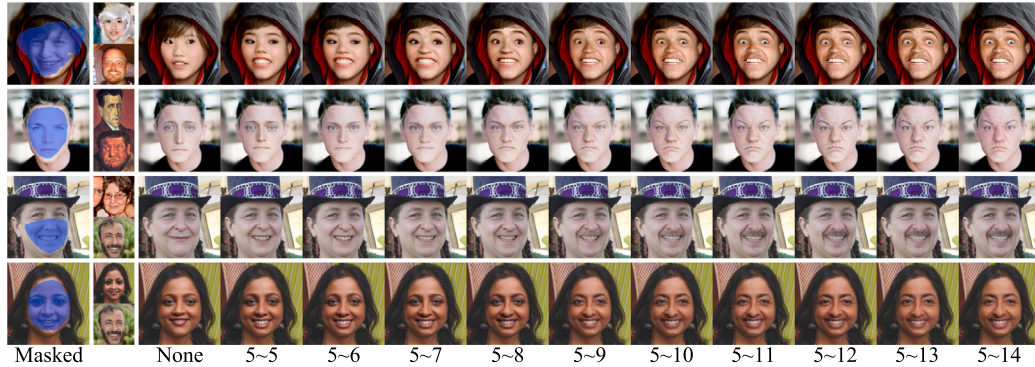
5. Applications with our facial inpainting

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 Figs. 13 and 14, for a masked input, our EXE-GAN produces high-quality local facial attribute transfer results by leveraging facial attributes of exemplars.

Table 5IDD scores for facial inpainting by exemplars with (f) EXE-GAN without $\mathcal{L}_{id}$, and (Ours) EXE-GAN. **Bold: top-1 quantity.**

Methods	Fixed	10–20%	20–30%	30–40%	40–50%	50–60%	60–70%	70–80%	80–90%	90–100%
(f)	1.277	1.449	1.430	1.407	1.376	1.314	1.292	1.251	1.211	1.193
Ours	1.264	1.447	1.425	1.401	1.368	1.303	1.283	1.240	1.197	1.161

**Fig. 13.** Examples of local facial attribute transfer.**Fig. 14.** More examples of local facial attribute transfer guided by exemplars.**Fig. 15.** Examples of facial style mixing (from top 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 channels are replaced by those of the second exemplar.

5.2. Guided facial style mixing

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 [61,62] 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. 15.

5.3. Hairstyle editing

We further fine-tuned our trained EXE-GAN with extra hand-drawn-like sketches which were produced automatically with a pencil-sketch filter [67]. The application allows users to sketch in the mask to



Fig. 16. Examples of hairstyle editing: (the first row) edited results with different sketches guided by different exemplars and (last two rows) edited results with the same sketch guided by different exemplars.



Fig. 17. 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 [91].



Fig. 18. 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.

indicate roughly the hair styles. Given a masked image with sketches and an exemplar image, EXE-GAN produces a style-edited output. As shown in Fig. 16, 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 achieved by masking them out and taking a different image from the same person as exemplar. As shown in Fig. 17, we compared our method to Lyu et al.’s eyeglasses removal method [91] on tinted eyeglasses (leftmost), sunglasses (mid-left), and myopia glasses (mid-right). The results show that our guided facial image recovery

method performs well in removing tinted eyeglasses and sunglasses which may fail with Lyu et al.’s method. Fig. 18 shows that our method can also recover faces from occluded glasses or masks effectively.

5.5. Inherent stochasticity

EXE-GAN can 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 [61,62]. Fig. 19 shows a variety of facial inpainting results with various random latent codes.



Fig. 19. 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.

6. Conclusion, limitations, and future work

In this paper, we have presented a novel interactive framework for realistic facial inpainting by taking advantage 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 back-propagation technique to address the issue of visual inconsistency on the filling boundary. Extensive experiments and applications have demonstrated the effectiveness of our 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 [67]. 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 images in the experimented training datasets [61,79] 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.

CRedit authorship contribution statement

Wanglong Lu: Writing – original draft, Validation, Software, Methodology. **Hanli Zhao:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Xianta Jiang:** Writing – review & editing, Supervision, Methodology. **Xiaogang Jin:** Writing – review & editing, Methodology, Conceptualization. **Yong-Liang Yang:** Writing – review & editing, Formal analysis, Conceptualization. **Kaijie Shi:** Investigation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments. This work was supported by the Ningbo Major Special Projects of the “Science and Technology Innovation 2025” (Grant No. 2023Z143), Zhejiang Provincial Natural Science Foundation of China (Grant No. LZ21F020001), Wenzhou Major Tackling Project of Science and Technology Innovation (Grant No. ZF2024001), Natural Sciences and Engineering Research Council of Canada (NSERC, Grant No. DGEER-2020-00296), Government of Canada’s New Frontiers in Research Fund (NFRF, Grant No. NFRFE-2022-00407), and the Digital Research Alliance of Canada.

Data availability

The data involved in this study are all public data, which can be downloaded through public channels.

References

- [1] M. Chen, Z. Liu, L. Ye, Y. Wang, Attentional coarse-and-fine generative adversarial networks for image inpainting, *Neurocomputing* 405 (2020) 259–269.
- [2] L. Jiao, H. Wu, H. Wang, R. Bie, Multi-scale semantic image inpainting with residual learning and GAN, *Neurocomputing* 331 (2019) 199–212.
- [3] L. Chen, C. Yuan, X. Qin, W. Sun, X. Zhu, Contrastive structure and texture fusion for image inpainting, *Neurocomputing* 536 (2023) 1–12.
- [4] C.-H. Lee, Z. Liu, L. Wu, P. Luo, MaskGAN: Towards diverse and interactive facial image manipulation, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5548–5557.
- [5] A. Chen, R. Liu, L. Xie, Z. Chen, H. Su, J. Yu, Sofgan: A portrait image generator with dynamic styling, *ACM Trans. Graph.* 41 (1) (2022) 1.
- [6] Y. Choi, Y. Uh, J. Yoo, J.-W. Ha, StarGAN v2: Diverse image synthesis for multiple domains, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8185–8194.
- [7] X. Li, S. Zhang, J. Hu, L. Cao, X. Hong, X. Mao, F. Huang, Y. Wu, R. Ji, Image-to-image translation via hierarchical style disentanglement, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8639–8648.
- [8] J. Liu, C. Jung, Facial image inpainting using attention-based multi-level generative network, *Neurocomputing* 437 (2021) 95–106.
- [9] L. Zhao, Q. Mo, S. Lin, Z. Wang, Z. Zuo, H. Chen, W. Xing, D. Lu, UCTGAN: Diverse image inpainting based on unsupervised cross-space translation, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5740–5749.
- [10] S. Zhao, J. Cui, Y. Sheng, Y. Dong, X. Liang, E.I. Chang, Y. Xu, Large scale image completion via co-modulated generative adversarial networks, in: *International Conference on Learning Representations*, 2021.
- [11] Y. Yang, X. Guo, Generative landmark guided face inpainting, in: *Chinese Conference on Pattern Recognition and Computer Vision*, 2020, pp. 14–26.
- [12] Y. Jo, J. Park, SC-FEGAN: Face editing generative adversarial network with user’s sketch and color, in: *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1745–1753.
- [13] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T. Huang, Free-form image inpainting with gated convolution, in: *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4470–4479.
- [14] K. Nazari, E. Ng, T. Joseph, F. Qureshi, M. Ebrahimi, EdgeConnect: Structure guided image inpainting using edge prediction, in: *IEEE/CVF International Conference on Computer Vision Workshop*, 2019, pp. 3265–3274.
- [15] M. Bertalmio, G. Sapiro, V. Caselles, C. Ballester, Image inpainting, in: *ACM Annual Conference on Computer Graphics and Interactive Techniques*, ACM, 2000, pp. 417–424.
- [16] Levin, Zomet, Weiss, Learning how to inpaint from global image statistics, in: *IEEE/CVF International Conference on Computer Vision*, 2003, pp. 305–312.
- [17] V. Kwatra, I. Essa, A. Bobick, N. Kwatra, Texture optimization for example-based synthesis, *ACM Trans. Graph.* 24 (3) (2005) 795–802.
- [18] C. Barnes, E. Shechtman, A. Finkelstein, D.B. Goldman, PatchMatch: A randomized correspondence algorithm for structural image editing, *ACM Trans. Graph.* 28 (3) (2009) 24.
- [19] H. Zhao, H. Guo, X. Jin, J. Shen, X. Mao, J. Liu, Parallel and efficient approximate nearest patch matching for image editing applications, *Neurocomputing* 305 (2018) 39–50.

- [20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: *Advances in Neural Information Processing Systems*, vol. 27, Cambridge, MA, 2014, pp. 2672–2680.
- [21] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, S. Ganguli, Deep unsupervised learning using nonequilibrium thermodynamics, in: *ICML*, 2015, pp. 2256–2265.
- [22] S. Iizuka, E. Simo-Serra, H. Ishikawa, Globally and locally consistent image completion, *ACM Trans. Graph.* 36 (4) (2017) 107.
- [23] L. Xu, X. Zeng, W. Li, Z. Huang, Multi-granularity generative adversarial nets with reconstructive sampling for image inpainting, *Neurocomputing* 402 (2020) 220–234.
- [24] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Generative image inpainting with contextual attention, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.
- [25] H. Liu, B. Jiang, Y. Xiao, C. Yang, Coherent semantic attention for image inpainting, in: *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4169–4178.
- [26] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, X. Liu, S. Wen, E. Ding, Image inpainting with learnable bidirectional attention maps, in: *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8857–8866.
- [27] C. Wang, M. Shao, D. Meng, W. Zuo, Dual-pyramidal image inpainting with dynamic normalization, *IEEE Trans. Circuits Syst. Video Technol.* 32 (9) (2022) 5975–5988.
- [28] H. Wu, J. Zhou, IID-net: Image inpainting detection network via neural architecture search and attention, *IEEE Trans. Circuits Syst. Video Technol.* 32 (3) (2022) 1172–1185.
- [29] Y. Wang, X. Tao, X. Qi, X. Shen, J. Jia, Image inpainting via generative multi-column convolutional neural networks, in: *Advances in Neural Information Processing Systems*, 2018, pp. 331–340.
- [30] Z. Wan, J. Zhang, D. Chen, J. Liao, High-fidelity pluralistic image completion with transformers, in: *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4672–4681.
- [31] Q. Liu, Z. Tan, D. Chen, Q. Chu, X. Dai, Y. Chen, M. Liu, L. Yuan, N. Yu, Reduce information loss in transformers for pluralistic image inpainting, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11347–11357.
- [32] S. Zhou, K.C. Chan, C. Li, C.C. Loy, Towards robust blind face restoration with codebook lookup TransFormer, in: *Advances in Neural Information Processing Systems*, 2022.
- [33] M. Wang, W. Lu, J. Lyu, K. Shi, H. Zhao, Generative image inpainting with enhanced gated convolution and transformers, *Displays* 75 (2022) 102321.
- [34] S.S. Phutke, S. Murala, Pseudo decoder guided light-weight architecture for image inpainting, *IEEE Trans. Image Process.* 31 (2022) 6577–6590.
- [35] W. Lu, X. Jiang, X. Jin, Y.-L. Yang, M. Gong, T. Wang, K. Shi, H. Zhao, GRIG: Few-shot generative residual image inpainting, 2023, arXiv preprint arXiv: 2304.12035.
- [36] G. Liu, F.A. Reda, K.J. Shih, T.-C. Wang, A. Tao, B. Catanzaro, Image inpainting for irregular holes using partial convolutions, in: *European Conference on Computer Vision*, 2018, pp. 89–105.
- [37] Z. Yi, Q. Tang, S. Azizi, D. Jang, Z. Xu, Contextual residual aggregation for ultra high-resolution image inpainting, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7508–7517.
- [38] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: *CVPR*, 2022, pp. 10684–10695.
- [39] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, M. Norouzi, Palette: Image-to-image diffusion models, in: *ACM SIGGRAPH*, 2022.
- [40] A. Lugmayr, M. Daneljanić, A. Romero, F. Yu, R. Timofte, L. Van Gool, Repaint: Inpainting using denoising diffusion probabilistic models, in: *CVPR*, 2022.
- [41] C. Zheng, T.-J. Cham, J. Cai, Pluralistic image completion, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1438–1447.
- [42] Y. Yu, F. Zhan, R. Wu, J. Pan, K. Cui, S. Lu, F. Ma, X. Xie, C. Miao, Diverse image inpainting with bidirectional and autoregressive transformers, in: *ACM International Conference on Multimedia*, ACM, 2021, pp. 69–78.
- [43] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, J. Jia, MAT: Mask-aware transformer for large hole image inpainting, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [44] S. Xu, D. Liu, Z. Xiong, E2I: Generative inpainting from edge to image, *IEEE Trans. Circuits Syst. Video Technol.* 31 (4) (2021) 1308–1322.
- [45] Y. Ren, X. Yu, R. Zhang, T.H. Li, S. Liu, G. Li, StructureFlow: Image inpainting via structure-aware appearance flow, in: *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 181–190.
- [46] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, J. Luo, Foreground-aware image inpainting, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5833–5841.
- [47] Y. Zhao, W. Chen, J. Xing, X. Li, Z. Bessinger, F. Liu, W. Zuo, R. Yang, Identity preserving face completion for Large Ocular Region occlusion, in: *British Machine Vision Conference*, BMVA Press, 2018, 109.
- [48] J. Yu, K. Li, J. Peng, Reference-guided face inpainting with reference attention network, *Neural Comput. Appl.* 34 (12) (2022) 9717–9731.
- [49] H. Li, W. Wang, C. Yu, S. Zhang, SwapInpaint: Identity-specific face inpainting with identity swapping, *IEEE Trans. Circuits Syst. Video Technol.* 32 (7) (2022) 4271–4281.
- [50] Y. Zhao, B. Price, S. Cohen, D. Gurari, Guided image inpainting: Replacing an image region by pulling content from another image, in: *2019 IEEE Winter Conference on Applications of Computer Vision*, IEEE, Los Alamitos, 2019, pp. 1514–1523.
- [51] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.
- [52] T. Portenier, Q. Hu, A. Szabó, S.A. Bigdeli, P. Favaro, M. Zwicker, Faceshop: Deep sketch-based face image editing, *ACM Trans. Graph.* 37 (4) (2018).
- [53] T. Xiao, J. Hong, J. Ma, ELEGANT: Exchanging latent encodings with GAN for transferring multiple face attributes, in: *European Conference on Computer Vision*, 2018, pp. 172–187.
- [54] L. Li, J. Bao, H. Yang, D. Chen, F. Wen, Advancing high fidelity identity swapping for forgery detection, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5073–5082.
- [55] R. Chen, X. Chen, B. Ni, Y. Ge, SimSwap: An efficient framework for high fidelity face swapping, in: *28th ACM International Conference on Multimedia*, ACM, 2020, pp. 2003–2011.
- [56] X. Liu, Y. Yang, P. Hall, Learning to warp for style transfer, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3701–3710.
- [57] Z. Liu, M. Li, Y. Zhang, C. Wang, Q. Zhang, J. Wang, Y. Nie, Fine-grained face swapping via regional GAN inversion, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8578–8587.
- [58] R. Zhang, W. Li, Y. Zhang, H. Zhang, J. Yu, R. Yang, W. Xu, Image re-composition via regional content-style decoupling, in: *ACM International Conference on Multimedia*, ACM, Virtual Event, China, 2021, pp. 3–11.
- [59] H. Kim, Y. Choi, J. Kim, S. Yoo, Y. Uh, Exploiting spatial dimensions of latent in GAN for real-time image editing, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 852–861.
- [60] J. Choi, S. Kim, Y. Jeong, Y. Gwon, S. Yoon, ILVR: Conditioning method for denoising diffusion probabilistic models, in: *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14347–14356.
- [61] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [62] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [63] Y. Shi, X. Yang, Y. Wan, X. Shen, SemanticStyleGAN: Learning compositional generative priors for controllable image synthesis and editing, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [64] R. Abdal, Y. Qin, P. Wonka, Image2StyleGAN: How to embed images into the stylegan latent space? in: *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4431–4440.
- [65] R. Abdal, Y. Qin, P. Wonka, Image2StyleGAN++: How to edit the embedded images? in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8293–8302.
- [66] J. Zhu, Y. Shen, D. Zhao, B. Zhou, In-domain GAN inversion for real image editing, in: *European Conference on Computer Vision*, 2020, pp. 592–608.
- [67] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shaprio, D. Cohen-Or, Encoding in style: a stylegan encoder for image-to-image translation, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2287–2296.
- [68] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, D. Cohen-Or, Designing an encoder for stylegan image manipulation, *ACM Trans. Graph.* 40 (4) (2021) 133.
- [69] Y. Wu, Y. Yang, Q. Xiao, X. Jin, Coarse-to-fine: facial structure editing of portrait images via latent space classifications, *ACM Trans. Graph.* 40 (4) (2021) 46.
- [70] Y. Wu, Y. Yang, X. Jin, HairMapper: Removing hair from portraits using GANs, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4227–4236.
- [71] B. Li, T. Ma, P. Zhang, M. Hua, W. Liu, Q. He, Z. Yi, ReGANIE: Rectifying GAN inversion errors for accurate real image editing, *Proc. AAAI Conf. Artif. Intell.* (AAAI) 37 (1) (2023) 1269–1277.
- [72] J. Zhu, P. Krähenbühl, E. Shechtman, A.A. Efros, Generative visual manipulation on the natural image manifold, in: *European Conference on Computer Vision*, 2016, pp. 597–613.
- [73] E. Collins, R. Bala, B. Price, S. Süsstrunk, Editing in style: Uncovering the local semantics of GANs, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5770–5779.
- [74] O. Kafri, O. Patashnik, Y. Alaluf, D. Cohen-Or, StyleFusion: Disentangling spatial segments in StyleGAN-generated images, *ACM Trans. Graph.* 41 (5) (2022) 179.
- [75] L. Mescheder, S. Nowozin, A. Geiger, Which training methods for GANs do actually converge? in: *International Conference on Machine Learning*, vol. 80, PMLR, 2018, pp. 3481–3490.
- [76] J. Deng, J. Guo, N. Xue, S. Zafeiriou, ArcFace: Additive angular margin loss for deep face recognition, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4685–4694.
- [77] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

- [78] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations, 2015.
- [79] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: International Conference on Learning Representations, 2018.
- [80] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, 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, 2017, pp. 6629–6640.
- [81] Q. Dong, C. Cao, Y. Fu, Incremental transformer structure enhanced image inpainting with masking positional encoding, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11358–11368.
- [82] C. Zheng, T.-J. Cham, J. Cai, D. Phung, Bridging global context interactions for high-fidelity image completion, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11512–11522.
- [83] T. Liu, L. Liao, Z. Wang, S. Satoh, Reference-guided texture and structure inference for image inpainting, in: 2022 IEEE International Conference on Image Processing, ICIP, IEEE, 2022, pp. 1996–2000.
- [84] L. Liao, T. Liu, D. Chen, J. Xiao, Z. Wang, C.-W. Lin, S. Satoh, TransRef: Multi-scale reference embedding transformer for reference-guided image inpainting, 2023, arXiv preprint arXiv:2306.11528.
- [85] MMEEditing Contributors, MMEEditing: OpenMMLab image and video editing toolbox, 2022, <https://github.com/open-mmlab/mmediting>.
- [86] S. Zhao, Z. Liu, J. Lin, J.-Y. Zhu, S. Han, Differentiable augmentation for data-efficient GAN training, in: NeurIPS, vol. 33, 2020, pp. 7559–7570.
- [87] A. Bulat, G. Tzimiropoulos, How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks), in: IEEE/CVF International Conference on Computer Vision, 2017, pp. 1021–1030.
- [88] K. Kim, Y. Kim, S. Cho, J. Seo, J. Nam, K. Lee, S. Kim, K. Lee, DiffFace: Diffusion-based face swapping with facial guidance, 2022, arXiv preprint arXiv:2212.13344.
- [89] W. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, J. Lu, DiffSwap: High-fidelity and controllable face swapping via 3D-aware masked diffusion, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8568–8577.
- [90] J. Li, Z. Xiong, D. Liu, Reference-guided landmark image inpainting with deep feature matching, IEEE Trans. Circuits Syst. Video Technol. 32 (12) (2022) 8422–8435.
- [91] J. Lyu, Z. Wang, F. Xu, Portrait eyeglasses and shadow removal by leveraging 3D synthetic data, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3429–3439.



Wanglong Lu is a Ph.D. student in the Department of Computer Science at Memorial University of Newfoundland, Canada. He received his B.Sc. degree in digital media technology from the Communication University of Zhejiang, China, in 2018, and his M.Sc. degree in computer software and theory from Wenzhou University, China, in 2021. His research interests include image inpainting, image editing, and image recognition.



Hanli Zhao is a Professor in the College of Computer Science and Artificial Intelligence, Wenzhou University, China. He received his B.Sc. degree in software engineering from Sichuan University, China, in 2004, and his Ph.D. degree in computer science from the State Key Lab of CAD&CG, Zhejiang University in 2009. His current research interests include computer graphics, computer vision, pattern recognition, medical image analysis, and deep learning.



Xianta Jiang is currently an Associate Professor in the Department of Computer Science at Memorial University of Newfoundland, Canada. He received a dual Ph.D. in computer science from Simon Fraser University and Zhejiang University (2015). He then took a post-doctoral fellowship in Engineering Science, Simon Fraser University and worked as a senior research associate in the Department of Surgery at the University of Alberta. His research interests include machine intelligence, HCI, and wearables.



Xiaogang Jin is a Professor in the State Key Laboratory of CAD&CG, Zhejiang University. He received a B.Sc. degree in computer science and M.Sc. and Ph.D. degrees in applied mathematics from Zhejiang University, P.R. China, in 1989, 1992, and 1995, respectively. His current research interests include image processing, digital humans, traffic simulation, collective behavior simulation, cloth animation, virtual try-on, digital faces, implicit surface modeling and applications, creative modeling, computer-generated marbling, sketch-based modeling, and virtual reality. He received an ACM Recognition of Service Award in 2015 and two Best Paper Awards from CASA 2017 and CASA 2018. He is a member of the IEEE and ACM.



Yong-Liang Yang is a Senior Lecturer in the Department of Computer Science at the University of Bath, United Kingdom. He received B.Sc. and Ph.D. degrees in computer science from Tsinghua University, China. His research area is broadly in visual computing, with particular interests in shape modeling, computational design, and interactive techniques.



Kaijie Shi is a Ph.D. student in the Department of Computer Science at Memorial University of Newfoundland, Canada. He received his B.Sc. degree in geographic information science from Nanjing Xiaozhuang University, China, in 2019, and his M.Sc. degree in computer science and technology from Wenzhou University, China, in 2022. His research interests include computer graphics and computer vision.