The Visual Computer (2024) 40:699–715
https://doi.org/10.1007/s00371-023-02810-4

ORIGINAL ARTICLE

Image-to-image translation using an offset-based multi-scale codes GAN encoder

Zihao Guo[1] · Mingwen Shao[1] · Shunhang Li[1]

Accepted: 12 February 2023 / Published online: 4 March 2023

Abstract

Despite the remarkable achievements of generative adversarial networks (GANs) in high-quality image synthesis, applying pre-trained GAN models to image-to-image translation is still challenging. Previous approaches typically map the conditional image into the latent spaces of GANs by per-image optimization or learning a GAN encoder. However, neither of these two methods can ideally perform image-to-image translation tasks. In this work, we propose a novel learning-based framework which can complete common image-to-image translation tasks with high quality in real-time based on pre-trained GANs. Specifically, to mitigate the semantic misalignment between conditional and synthesized images, we propose an offset-based image synthesis method that allows our encoder to use multiple rather than one forward propagation to predict the latent codes. During the multiple forward passes, the final latent codes are adjusted continuously according to the semantic difference between the conditional image and the current synthesized image. To further reduce the loss of details during encoding, we extract multiple latent codes at multiple scales from input instead of a single code to synthesize the image. Moreover, we propose an optional multiple feature maps fusion module that combines our encoder with different generators to implement our multiple latent codes strategies. Finally, we analyze the performance and demonstrate the effectiveness of our framework by comparing it with state-of-the-art works on super-resolution and conditional face synthesis tasks.

Keywords Generative adversarial networks · GAN inversion · Image-to-image translation · Super-resolution · Conditional face synthesis

1 Introduction

Recently, Generative Adversarial Networks (GANs) [1] have evolved considerably [2] in synthesizing high-quality and diverse images and have contributed many pre-trained models for unconditional image generation [3–5]. The ability to synthesize high-quality images enables GANs to be competent for many image-to-image translation tasks, including super-resolution [6–9] and conditional image synthesis [10–16]. However, most existing GAN-based methods have to design their network structures or loss functions specially for a particular task and train their networks from scratch. This paradigm limits their generalization ability and prevents them from enjoying the phenomenal generative power of well-trained large-scale GANs (e.g., ProGAN [3] and StyleGAN [4,5]). To apply these excellent pre-trained GANs to image-to-image translation, one faces the challenge that standard GANs are initially designed for synthesizing images from random noises. Therefore, these pre-trained GANs are unable to take conditional images for any post-processing. To address this issue, GAN inversion [17] is the common approach, which means the given image is first mapped into a latent code and then the latent code is fed into the generator to reconstruct the given image. In this way, the inverted code can be used for further processing.

Current GAN inversion methods are mainly divided into two categories: one is optimization-based [18–20], which processes each image separately and iterates thousands of times to reconstruct a single image. And the other is learning-based [21–23], which requires training a deep neural network (encoder) to map the given image back into a latent code.

✉ Mingwen Shao
smw278@126.com

Zihao Guo
guozh980422@163.com

Shunhang Li
z20070036@s.upc.edu.cn

[1] College of Computer Science and Technology, China University of Petroleum, Qingdao, China

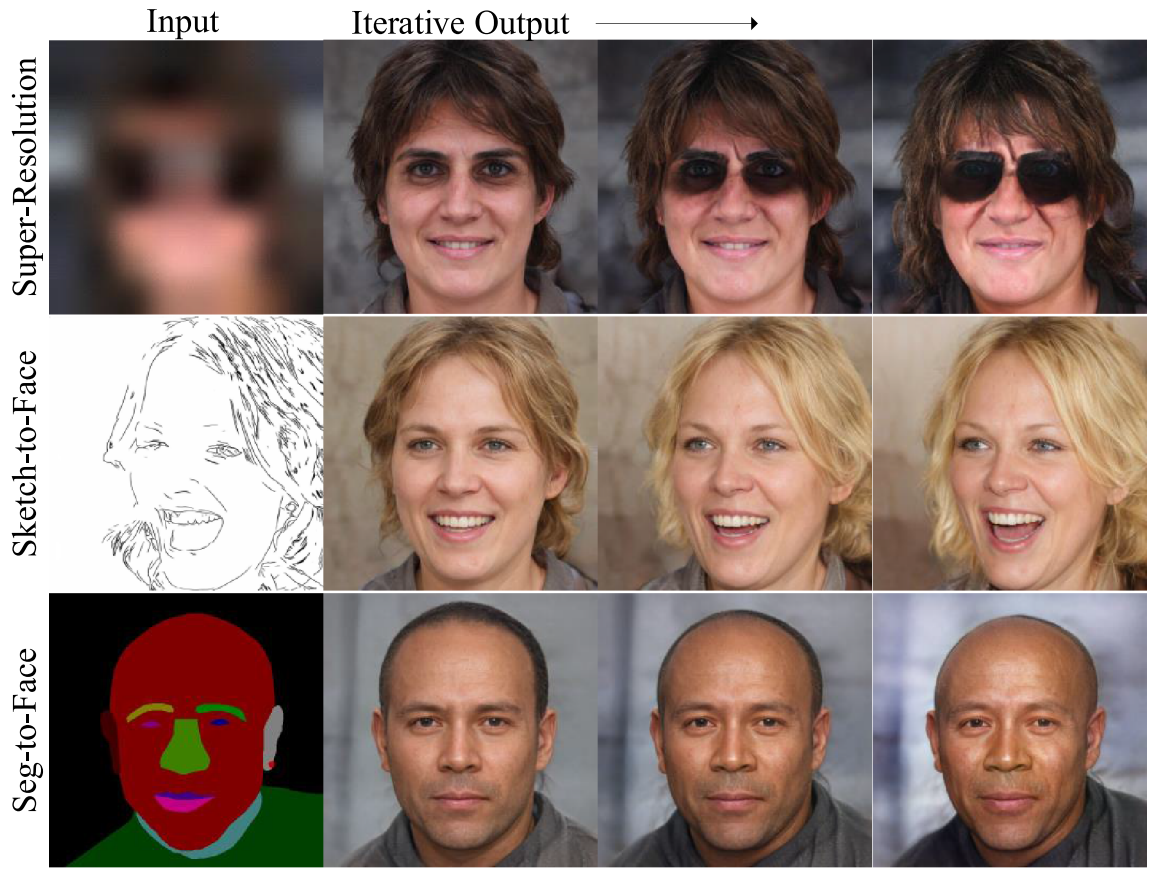

Fig. 1 The iterative output of OME on different tasks. For each task, we show the conditional input image on the left followed by intermediate synthesized images. As seen, the output image is gradually aligned with the semantics of the conditional input image

However, neither of the two mentioned methods can ideally achieve image-to-image translation tasks. Specifically, the optimization-based methods will not work when the conditional and synthesized images cannot directly form the reconstruction loss at the pixel level (e.g., the sketch-to-face image translation task). Meanwhile, the long inference time is overwhelming, although thousands of iterations usually produce better synthesized images. In contrast, learning-based approaches can achieve image-to-image translation by performing forward propagation only once. However, it is difficult for the encoder to fully exploit the generative power of the pre-trained GANs with only a single forward propagation. As a result, learning-based approaches usually lead to unsatisfactory image quality, such as semantic misalignment between conditional and synthesized images, and loss of details in the synthesized images.

In principle, it is hard for the encoder to align semantic information with a single forward propagation. After all, optimization-based methods forward propagate thousands of times but still cannot align semantic information ideally. In view of this, we propose an image synthesis method for image-to-image translation based on encoder iterations. Specifically, the encoder performs several forward passes instead of once to achieve image-to-image tasks. In each forward propagation, the encoder takes the conditional image and the last synthesized image as input and predicts offset codes based on the difference between the two images. During the iterative translation process, the latent code will be continuously updated by the offset codes. And correspondingly, the synthesized image will gradually align with the semantics of the given image, as shown in Fig. 1. In this way, we can alleviate the issue of semantic misalignment at the cost of a negligible increase in inference time.

Furthermore, one of the reasons for the loss of detail in synthetic images is the limited expressiveness of individual latent codes. Consider that the higher the resolution of the target image, the more information and details it contains. Therefore, it is difficult to synthesize a high-resolution image with a single latent code, which inevitably results in the loss of details. To alleviate this problem, we synthesize an image using multiple latent codes rather than a single one. Specifically, we use an encoder based on a Feature Pyramid Network (FPN) [24] to extract multiple multi-scale latent codes from input, which significantly improves the quality of synthesized images. Moreover, to combine our encoder with common generators with different structures and implement

our multiple latent codes strategies, we propose an optional multiple feature maps fusion module based on grouped channel attention weights. This lightweight but powerful module also helps accelerate the training convergence of our encoder and improves the performance of our framework. We analyze the performance of our framework on GAN inversion and demonstrate the versatility and effectiveness of our framework by comparing it with state-of-the-art tasks on common image-to-image translation works, including GAN inversion, super-resolution, sketch-to-face synthesis, and segmentation map-to-face synthesis.

In conclusion, it is of theoretical and practical application meaning to solve common image-to-image translation tasks with high quality in real-time. And recently, the advent of GAN inversion method has made it possible to use excellent pre-trained GAN models for conditional image generation. However, current GAN inversion methods cannot ideally perform image-to-image translation. Therefore, our primary motivation is to improve the GAN inversion-based method's performance for image-to-image translation tasks.

The main contributions of this paper are as follows:

- Based on pre-trained GANs, we propose a generic framework OME, short for offset-based multi-scale codes GAN encoder, that can complete common image-to-image tasks with high quality in real-time.
- We propose an offset-based image synthesis method for image-to-image translation that mitigates the semantic misalignment between conditional and synthesized images. To the best of our knowledge, we are the first to apply an iterative update method to image-to-image translation tasks based on pre-trained GANs.
- We propose a multiple latent codes strategy to further improve the quality of synthesized images. We also design a novel optional multiple feature maps dynamic fusion module that allows our encoder to be combined with different GANs (e.g., ProGAN and StyleGANs) to implement our multiple latent codes strategy.

2 Related work

2.1 Latent code injection of GANs

A reasonable injection of sampled noise (latent code) into the generator is one of the prerequisites for the powerful image generation capability of GANs. There are two popular methods of latent code injection for GANs. One can be represented by ProGAN [3], which injects latent codes from the top layer, and the other can be represented by StyleGAN [4,5], which injects latent codes from the global layers. Specifically, ProGAN injects latent code from the top of the generator (i.e., the layer with the smallest resolution) and employs a growth strategy during training. That is, as ProGAN training proceeds, new layers are gradually added to generate finer details, and eventually, high-resolution images are obtained. Unlike ProGAN, StyleGAN first maps the latent code to the multiple style codes through a fully connected network and then manipulates the mean and variance of the global layers (i.e., every block of the generator) using the style codes to control the content of the generated images better. Due to the structural differences between the above two kinds of GANs, previous image processing works based on pre-trained GANs are usually only applicable to one of them. In contrast, using an optional feature fusion module, our OME achieves compatibility with both GANs.

2.2 GAN inversion

The purpose of GAN inversion [21] is to find latent codes in the latent space of the pre-trained GAN model that can reconstruct given images. It has recently attracted increasing attention as an essential step in applying pre-trained GAN models to the real world. There are two main techniques for GAN inversion: optimization-based [18–20,25] and learning-based [21–23]. The optimization-based methods process each image separately and invoke the generator thousands of times to reconstruct a single image, which requires several minutes but typically achieves higher reconstruction quality. The learning-based methods typically train a neural network to encode a given image to the latent code. After training, the image reconstruction can be done quickly by invoking the generator only once in the inference phase, but it typically achieves inferior reconstruction quality. Based on the GAN inversion, we aim to further apply the powerful generative capabilities of pre-trained GANs to a wide range of image-to-image translation tasks.

2.3 Image-to-image translation

Image-to-image translation aims at mapping a conditional input image of a source domain to a corresponding image of a target domain while keeping the semantic concepts unchanged. Pix2Pix [26] is the first work to use a conditional GAN to solve a variety of image translation tasks. Inspired by Pix2Pix, much of the work expands image-to-image translation to a broader range of domains, such as high-resolution synthesis [6–8], conditional image synthesis [10–16], and unsupervised learning [27–29]. The aforementioned works have to design dedicated loss functions and network structures for their tasks and train their network from sketch. This paradigm allows them to have better control over the generative power of their network. In other words, they can synthesize images that have good semantic alignment with the conditional inputs. However, the realism of their synthetic images is far from that of well-trained large-scale GANs. In

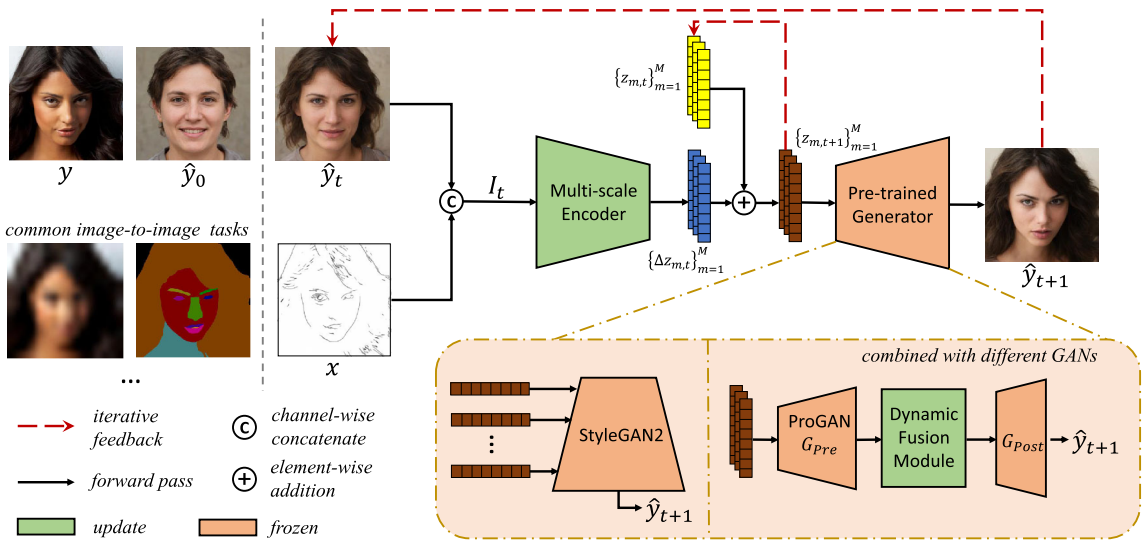

Fig. 2 The pipeline of our proposed OME. Given a conditional image $\mathbf{x}$, an offset-based iterative process is initialized with the average latent codes $\{\mathbf{z}_{m,0}\}_{m=1}^{M}$ and the corresponding image $\hat{\mathbf{y}}_0$. Take step t as an example. OME takes an extended input $\mathbf{I}_t$ obtained by concatenating $\mathbf{x}$ with the currently synthesized image $\hat{\mathbf{y}}_t$ corresponding to the current latent codes $\{\mathbf{z}_{m,t}\}_{m=1}^{M}$. The multi-scale encoder is then tasked with predicting a set of offset latent codes $\{\Delta\mathbf{z}_{m,t}\}_{m=1}^{M}$. Then the current latent codes $\{\mathbf{z}_{m,t}\}_{m=1}^{M}$ are updated to $\{\mathbf{z}_{m,t+1}\}_{m=1}^{M}$ by addition with $\{\Delta\mathbf{z}_{m,t}\}_{m=1}^{M}$. Finally, a newly computed image $\hat{\mathbf{y}}_{t+1}$ is obtained, corresponding to $\{\mathbf{z}_{m,t+1}\}_{m=1}^{M}$. Then $\hat{\mathbf{y}}_{t+1}$ is passed as input at the next step. The iterative process consists of N steps, and N is a small number. OME supports multiple forms of conditional input. Moreover, with an optional Dynamic fusion module, OME can be combined with different pre-trained GANs

order to apply the generative power of pre-trained GANs to image-to-image translation, many attempts have been made based on GAN inversion. For instance, the optimization-based mGANprior [25] controls the generative power of ProGAN well, but it suffers from excessive inference time and poor task generalization. And the learning-based pSp [30] is a state-of-the-art work based on StyleGAN2, but it suffers from semantic misalignment between the synthesized and conditional images since only one forward propagation is performed in their encoder. In contrast to them, our OME can better utilize and control the generative power of the pre-trained GANs to perform common image-to-image translation tasks with high quality in real-time.

3 Method

To apply the generative power of pre-trained GANs to image-to-image translation tasks, previous approaches typically map the conditional image into the latent spaces of GANs by per-image optimization or learning a GAN encoder. However, the optimization-based methods suffer from excessive inference time and poor task generalization, while the learning-based approaches cannot ideally exploit and control the generative power of pre-trained GANs, resulting in poor-quality synthetic images. To complete common image-to-image translation tasks with high quality in real-time, we propose OME, a generic image-to-image translation framework using an offset-based multi-scale code GAN encoder and an optional multiple feature maps dynamic fusion module, as illustrated in Fig. 2. In the following, we describe our OME framework in detail.

3.1 Offset-based image synthesis method

Let $O(\cdot)$, $G(\cdot)$, and $E(\cdot)$ denote our offset-based image synthesis method, a well-trained unconditional generator, and a GAN encoder, respectively. We aim at synthesizing an image $\hat{\mathbf{y}} = O(G(E(\mathbf{x})))$ such that $\hat{\mathbf{y}} \approx \mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ represent the conditional image and the target image, respectively. Unlike traditional learning-based methods, which simply use a single forward pass through the encoder $E(\cdot)$ and generator $G(\cdot)$ to get $\hat{\mathbf{y}}$, our offset-based method $O(\cdot)$ performs N steps to map the input into a corresponding latent code $\mathbf{z}$. Note that the N is a small number ($N <= 5$).

Specifically, let step t denotes a single forward pass through $E(\cdot)$ and $G(\cdot)$. At each step t, the conditional input $\mathbf{x}$ is concatenated with currently synthesized image $\hat{\mathbf{y}}_t$ and fed into the encoder $E(\cdot)$ as complete input $\mathbf{I}_t$. The job of the encoder $E(\cdot)$ is to predict an offset latent code $\Delta\mathbf{z}_t$, which indirectly represents differences between the conditional input $\mathbf{I}_t$ and current output $\hat{\mathbf{y}}_t$. That is,

$$\Delta\mathbf{z}_t = E\left(\mathbf{I}_t\right) = E\left(\mathbf{x} \| \hat{\mathbf{y}}_t\right), \tag{1}$$

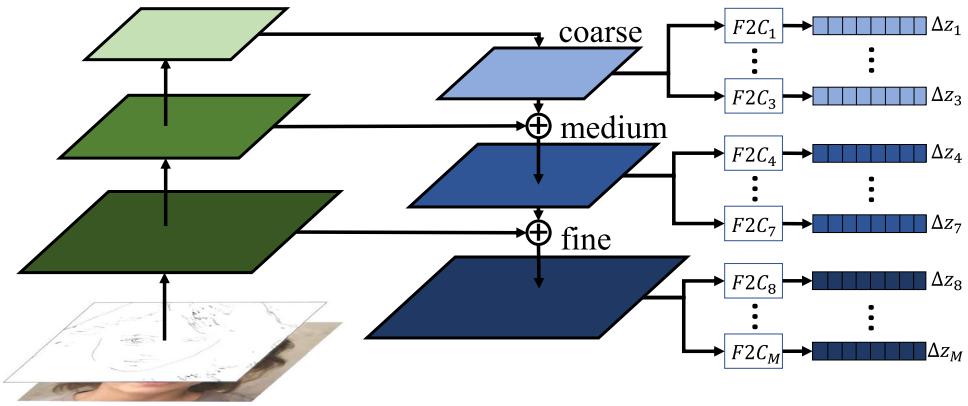

Fig. 3 The architecture of our multi-scale encoder. Our encoder extracts multi-scale features from coarse to fine using a standard FPN over a ResNet-50 backbone. The features are then mapped to offset latent codes $\{\Delta\mathbf{z}_{m,t}\}_{m=1}^{M}$ by feature-to-code networks $\{\mathbf{F2C}_m\}_{m=1}^{M}$. $\mathbf{F2C}_m$ consists of several 2-strided convolutions, aiming at reducing the spatial size of the feature map gradually

where the number of channels of $\mathbf{x}$ depends on the type of image-to-image translation task, while $\hat{\mathbf{y}}_t$ is an RGB image. Then the latent code $\mathbf{z}_t$ is updated to $\mathbf{z}_{t+1}$ by the offset latent code $\Delta\mathbf{z}_t$, and accordingly a new image $\hat{\mathbf{y}}_{t+1}$ is generated. It can be formulated as:

$$\hat{\mathbf{y}}_{t+1} = G\left(\mathbf{z}_{t+1}\right) = G\left(\Delta\mathbf{z}_t + \mathbf{z}_t\right). \tag{2}$$

Then the newly generated image $\mathbf{y}_{t+1}$ is concatenated with the conditional input $\mathbf{x}$ as the next complete input $\mathbf{I}_{t+1}$. This iterative process starts with an average latent code $\mathbf{z}_0$ sampled from the latent space of a pre-trained GAN and its corresponding image $\hat{\mathbf{y}}_0 = G(\mathbf{z}_0)$, and repeats until the preset number of steps N is reached.

In a sense, this offset-based image synthesis method enables the encoder to conduct several-step explorations in the latent space guided by differences between the conditional input and the currently synthesized image. This loosens the constraints on the encoder compared to the previous single-pass approaches and gives the encoder more opportunities to align the semantics of the synthesized image with the conditional image.

3.2 Multi-scale codes GAN encoder

The higher the resolution of the image, the more information and details it contains. In principle, it is difficult to synthesize a high-resolution image (e.g., 1024×1024) using a single latent code. Otherwise, we would have an unparalleled image compression technology. Therefore, using a single code to synthesize high-resolution images inevitably results in the loss of details as well as poor image quality. To alleviate this issue, we synthesize an image using multiple latent codes instead of a single one. Specifically, as shown in Fig. 3, we use a multi-scale encoder $ME(\cdot)$ based on a Feature Pyramid Network (FPN) [24] to extract multiple offset latent codes from multiple scales at each step. Accordingly, Eq. (1) is reformulated as:

$$\left\{\Delta\mathbf{z}_{m,t}\right\}_{m=1}^{M} = ME\left(\mathbf{I}_t\right), \tag{3}$$

where M and m denote the number and the index of latent codes, separately. And accordingly, Eq. (2) is reformulated as:

$$\begin{aligned}\hat{\mathbf{y}}_{t+1} &= G\left(\left\{\mathbf{z}_{m,t+1}\right\}_{m=1}^{M}\right) \\ &= G\left(\left\{\Delta\mathbf{z}_{m,t}\right\}_{m=1}^{M} + \left\{\mathbf{z}_{m,t}\right\}_{m=1}^{M}\right).\end{aligned} \tag{4}$$

Up to this point, we can combine our encoder $ME(\cdot)$ with the pre-trained GANs that inject latent codes from the global layers (as mentioned in Sect. 2.1) to synthesize high-quality images. Take StyleGAN2 as an example, the multiple latent codes $\{\mathbf{z}_m\}_{m=1}^{M}$ are directly used as style codes to control the generation of images. According to the structure of StyleGAN2, our encoder extracts 18 vectors, each with a dimension of 512 (i.e., M=18 and $\{\mathbf{z}_m\}_{m=1}^{M} \in \mathbb{R}^{18\times512}$ for StyleGAN2). And accordingly, the average latent code $\mathbf{z}_0$ is replaced with the average style codes $\left\{\mathbf{z}_{m,0}\right\}_{m=1}^{M}$.

In summary, when combined with a GAN that injects latent codes from the global layers, the process of synthesizing images by OME is as follows:

$$\hat{\mathbf{y}} = OME(\mathbf{x}) = O\left(G\left(ME\left(\mathbf{x}\right)\right)\right). \tag{5}$$

In order to combine with GANs that inject latent codes from the top (e.g., ProGAN) and implement our multiple codes strategy, we design a dynamic feature fusion module, which is described in detail below.

3.3 Multiple feature maps dynamic fusion module

Due to the structure of ProGAN injecting latent codes from the top layer (as mentioned in Sect. 2.1), we cannot directly synthesize a single image using multiple latent codes. Therefore, we need to integrate the information from multiple latent codes to produce a single image when combined with ProGAN. mGANprior [25] achieves this by optimizing the

channel fusion weights of the intermediate feature maps. This optimization-based strategy is effective but too time-consuming. Inspired by mGANprior, we further propose the multiple feature maps dynamic fusion module $DF(\cdot)$ based on channel attention weights of grouped feature maps. Specifically, the generator $G(\cdot)$ is divided into two parts, i.e., $G_{pre}(\cdot)$ and $G_{post}(\cdot)$, as mGANprior does. In this way, for any latent code $\mathbf{z}_m$, $G_{pre}(\cdot)$ can extract its corresponding intermediate feature map. This process can be described as:

$$\mathbf{F}_m = G_{pre}\left(\mathbf{z}_m\right). \tag{6}$$

To take full advantage of the information embedded in each latent code, we up-dimension the latent codes as much as possible and separate the generators before the final feature extraction layer (the eighth layer in ProGAN).

Our dynamic fusion module then obtains the channel attention weights [31] for each set of the feature maps:

$$\boldsymbol{\alpha}_m = SE\left(\mathbf{F}_m\right) = \sigma\left(W_1 \delta\left(W_0\left(GAP\left(\mathbf{F}_m\right)\right)\right)\right), \tag{7}$$

where $SE(\cdot)$ represents the SE module [31]. σ and δ represent the Sigmoid function and the ReLU operation [31] separately. $W_0 \in \mathbb{R}^{C \times \frac{C}{r}}$ and $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ represent the fully connected (FC) layers, while $r = 8$ represents the channel compression ratio as in [31]. And $GAP(\cdot)$ denotes the global average pooling as in [32]. After that, our dynamic fusion module realizes the dynamic information interaction between multiple groups of features by normalizing the channel attention weights:

$$\boldsymbol{\beta}_c = softmax\left(\left\{\boldsymbol{\alpha}_{m,c}\right\}_{m=1}^{M}\right), \tag{8}$$

where c denotes the channel index. Finally, we modulate and fuse the features using the normalized channel weights and feed them to the $G_{post}(\cdot)$ to obtain the processed image:

$$\hat{\mathbf{y}} = G_{post}\left(\sum_{m=1}^{M} F_m \odot \boldsymbol{\beta}_m\right), \tag{9}$$

where $\odot$ denotes channel-wise product as:

$$\left\{\mathbf{F}_m \odot \boldsymbol{\beta}_m\right\}_{i,j,c} = \left\{\mathbf{F}_m\right\}_{i,j,c} \times \left\{\boldsymbol{\beta}_m\right\}_c, \tag{10}$$

where i and j stand for the spatial location index.

In this way, we can combine our encoder $ME(\cdot)$ with the pre-trained ProGAN to synthesize high-quality images. In the setting of the number of latent codes, mGANprior has conducted a large number of experiments, and we follow their research by setting M as 20 (i.e., $\{\mathbf{z}_m\}_{m=1}^{M} \in \mathbb{R}^{20\times512}$ for ProGAN). And we repeat the average latent code $\mathbf{z}_0$ for M times to obtain the average latent codes $\left\{\mathbf{z}_{m,0}\right\}_{m=1}^{M}$ for ProGAN.

In summary, when combined with a GAN that injects latent codes from the top layer, the process of synthesizing images by OME can be represented by the following equation:

$$\hat{\mathbf{y}} = OME(\mathbf{x}) = O\left(G_{post}\left(DF\left(G_{pre}\left(ME\left(\mathbf{x}\right)\right)\right)\right)\right). \tag{11}$$

For better understanding, we describe our method concisely in Algorithm 1.

3.4 Loss functions

To competently perform a wide range of image processing tasks, it is necessary to design a uniform combination of simple but powerful loss functions. We first introduce the $\mathcal{L}_2$ loss on pixel-wise:

$$\mathcal{L}_2(\mathbf{x}, \mathbf{y}) = \|\mathbf{y} - OME(\mathbf{x})\|_2, \tag{12}$$

where $\mathbf{x}$ and $\mathbf{y}$ denote the conditional and target images, respectively. This loss function limits the generated image $OME(\mathbf{x})$ to be close to its corresponding target image $\mathbf{y}$ in the pixel dimension.

The second part is the LPIPS [33] loss, which constrains the similarity of the output image and its corresponding label image at the feature level:

$$\mathcal{L}_{\text{LPIPS}}\left(\mathbf{x}, \mathbf{y}\right) = \|F(\mathbf{y}) - F(OME(\mathbf{x}))\|_2, \tag{13}$$

where F represents the pre-trained network to extract perceptual features [34]. To further ensure the retention of face identity information in face processing tasks, we further introduce the Identity loss [30], which measures the cosine similarity of the face recognition scores between the synthesized image and the corresponding target image:

$$\mathcal{L}_{\text{ID}}(\mathbf{x}, \mathbf{y}) = 1 - \langle ArcFace(\mathbf{y}), ArcFace(OME(\mathbf{x}))\rangle, \tag{14}$$

where $ArcFace(\cdot)$ [35] is a pre-trained network for face recognition. Note that Identity loss is only counted in the face synthesizing tasks.

Therefore, our total loss function is as follows:

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = \lambda_1 \mathcal{L}_2(\mathbf{x}, \mathbf{y}) + \lambda_2 \mathcal{L}_{\text{LPIPS}}(\mathbf{x}, \mathbf{y}) + \lambda_3 \mathcal{L}_{\text{ID}}(\mathbf{x}, \mathbf{y}), \tag{15}$$

where λ_1, λ_2 and λ_3 are the hyperparameters representing the loss weights.

3.5 Implementation details

We aim to design a general framework competent for common image translation tasks, so we use the same hyperparameters of loss functions for different tasks. We assume that the finer the granularity of the loss function, the more important it is in the training process. Specifically, the $\mathcal{L}_2$ loss computes the distance on the image's pixel level with the finest granularity. The LIPIS loss computes the distance on the image's feature space with medium granularity. The Identity loss calculates the cosine similarity of the image's feature vectors with the coarsest granularity. Therefore, based on experience, we set $\lambda_1 = 1$ and $\lambda_2 = 0.8$. And for face synthesizing tasks, we set $\lambda_3 = 0.1$, otherwise $\lambda_3 = 0$. These settings make the weighted values of loss functions roughly ordered by their granularity's fineness while in a similar order of magnitude. Experiments show that our framework performs well on different tasks even when combined with different GANs, which can illustrate the robustness of our hyperparameters. Theoretically, further careful tuning of the hyperparameters can improve the performance of the framework. The resolution of the input images is 256×256, while the resolution of the synthesized face images is 1024×1024, and the resolution of the synthesized scene images is 256×256. We use ResNet [36] as the backbone of our encoder (ResNet-50 for faces and ResNet-34 for churches and bedrooms). During training, all synthesized images are resized to 256×256 before the loss objectives are computed. And we use the Ranger optimizer [37] with a constant learning rate of 0.001. During training, loss objectives are back-propagated according to each forward propagation. Note that in all experiments, our method does not involve the training of the GANs, and all pre-trained GAN models are obtained from the corresponding official open-source resources. We perform all experiments using a single NVIDIA GeForce RTX 3090.

Algorithm 1 Image-to-Image translation using our OME

Input: conditional image $\mathbf{x}$; **Output:** corresponding synthesized images $\hat{\mathbf{y}}$; **Initialization:** average latent codes $\{\mathbf{z}_{m,0}\}_{m=1}^{M}$; corresponding average image $\hat{\mathbf{y}}_0$;

$t \Leftarrow 0$;
while $t < N$ **do**
  $\mathbf{I}_t \Leftarrow x \| \hat{\mathbf{y}}_t$;
  $\{\Delta\mathbf{z}_{m,t}\}_{m=1}^{M} \Leftarrow ME(\mathbf{I}_t)$;
  $\{\mathbf{z}_{m,t+1}\}_{m=1}^{M} \Leftarrow \{\Delta\mathbf{z}_{m,t}\}_{m=1}^{M} + \{\mathbf{z}_{m,t}\}_{m=1}^{M}$;
  if combined with StyleGAN2 **then**
    $\hat{\mathbf{y}}_t \Leftarrow G(\{\mathbf{z}_{m,t}\}_{m=1}^{M})$;
  else if combined with ProGAN **then**
    $\hat{\mathbf{y}}_t \Leftarrow G_{post}\left(DF\left(G_{pre}\left(\{\mathbf{z}_{m,t}\}_{m=1}^{M}\right)\right)\right)$;
  end if
end while
return $\hat{\mathbf{y}}_t$;

4 Experiments and applications

To verify the effectiveness and generality of our approach, we conduct extensive experiments on common image-to-image translation tasks, including image inversion, super-resolution, and conditional face synthesis based on pre-trained models of state-of-the-art GANs, e.g., ProGAN [3] and StyleGAN2 [5].

Datasets We conduct experiments on various datasets, including CelebA-HQ [38] for faces and LSUN [39] for scenes. CelebA-HQ contains 30,000 high-quality images. We use about 24,000 images for training obtained by a standard train-test split. In the LSUN dataset, there are about 150,000 images for each scene, of which we use 300 images as the test set according to the official division. The ProGAN models are pre-trained on CelebA-HQ for faces, and LSUN for scenarios. And the StyleGAN2 is trained on FFHQ [4] for faces.

Baselines We compare our OME to the general image-to-image translation frameworks, including pix2pixHD [40], and pSp [30]. Pix2PixHD can synthesize higher resolution images than pix2pix, and pSp is the state-of-the-art image-to-image translation work based on StyleGAN2. For each task, we also compare specialized state-of-the-art methods. For the tasks where the given image and the target image can form pixel-level constraints (e.g., super-resolution), we also compare our method to state-of-the-art optimization-based methods, mGANprior [25] and Pulse [7]. The results show that our OME has comparable or better performance.

4.1 GAN inversion

The ability to inverse the image into the latent space of pre-trained GANs is the basis for subsequent image-to-image translation tasks. That is, our OME should at least have the ability to reconstruct images. Therefore, it is necessary to evaluate the performance of our OME in the GAN inversion task through comparison and ablation experiments. To analyze the effectiveness of all our modules, we first conduct extensive experiments on ProGAN and then verified the performance of our OME based on StyleGAN2. We first compare our OME with the following methods: (a) reconstructing an image by optimizing over a single latent code [41], (b) reconstructing a target image by optimizing over multiple latent codes [25], (c) training an encoding neural network to invert the given image into a single latent code [42]. Note that, for a fair comparison, we set the same backbone (ResNet without FPN) and loss function as OME for method (c), and we consider it as the baseline of the ablation experiment.

Results Table 1 presents a quantitative comparison measuring the Peak Signal-to-Noise Ratio (PSNR) [43], LPIPS [33], and inference time (Time) of the different methods exam-

Table 1 Quantitative comparison of GAN inversion methods

Method	Face			Church			Bedroom		
	PSNR↑	LPIPS↓	Time(s)↓	PSNR↑	LPIPS↓	Time(s)↓	PSNR↑	LPIPS↓	Time(s)↓
(a) Ma et al. [30]	20.95	0.3366	162	16.54	0.5173	93	17.10	0.6061	93
(b) Gu et al. [25]	**29.31**	*0.1619*	655	**24.29**	**0.1569**	587	**26.10**	**0.2213**	587
(c) Zhu et al. [42]	17.11	0.3127	**0.02**	13.49	0.3055	**0.02**	14.24	0.3929	**0.02**
(d) Ours	*21.43*	**0.1600**	*0.51*	*19.24*	*0.1582*	*0.50*	*17.79*	*0.2661*	*0.50*

Bold means the best, italic means the second best. ↑ means the higher the better, ↓ means the lower the better

ined above. As can be seen from the table, our learning-based method has a comparable image reconstruction quality to the optimizing-based mGANprior, which is one of the state-of-the-art methods for GAN inversion. However, mGANprior takes nearly a thousand times longer to infer than our method. A noteworthy observation is that thanks to the powerful feature extraction capability of our multi-scale codes encoder and the full utilization of the GAN generation capability by multiple encoder-based iterations, we approach or even surpass the mGANprior in the LPIPS metric.

Figure 4 shows the qualitative comparison results. We can see that images reconstructed by a single latent code (i.e., (a) and (c)) lose detailed information and face identity. Although mGANprior performs well in reconstructing the details of images, it is unbearably time-consuming. Combining Table 1, we can say the image reconstruction quality of our method is comparable to that of mGANprior, while the inference time of our method is much less than that of mGANprior, which is meaningful for high quality real-time image-to-image translation.

4.2 Ablation experiments

To further verify the effectiveness of OME, we conduct ablation experiments for each characteristic, including the offset-based inversion method (OB), extracting multiple latent codes from a single scale without FPN (MC), extracting multi-scale latent codes using an FPN encoder (MS), and the dynamic fusion module (DF). Note that when dynamic fusion is not used, all feature maps are fused using static aver-

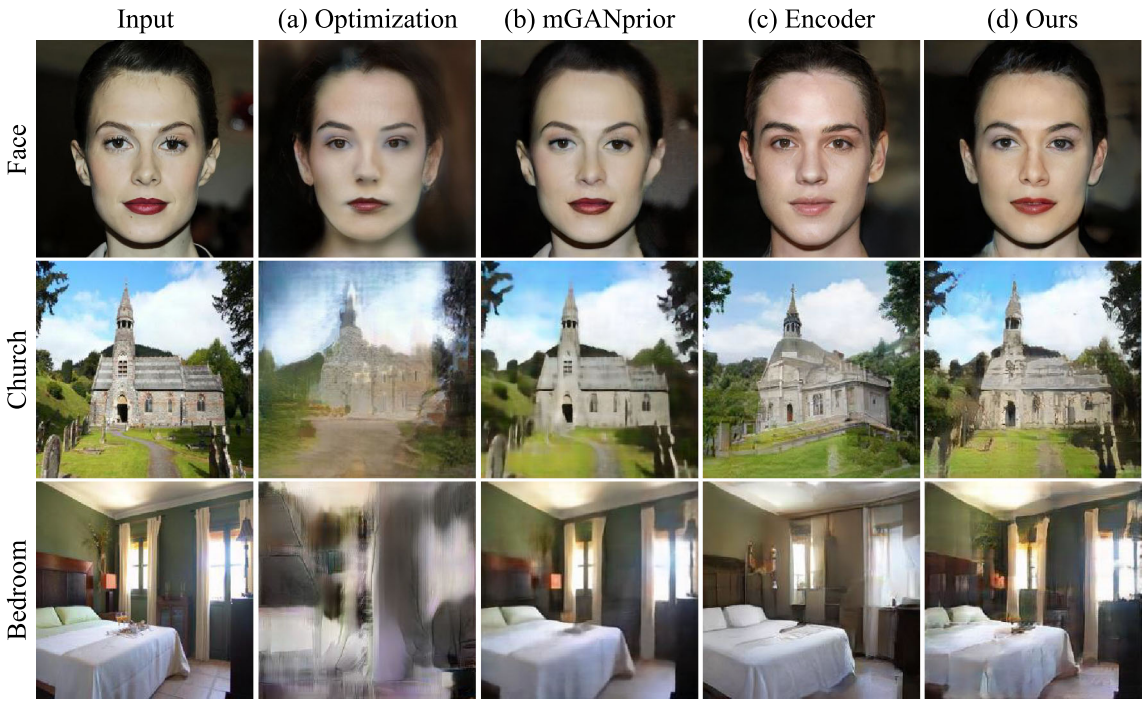

Fig. 4 Qualitative comparisons of different GAN inversion methods, including **a** optimizing a single latent code [41], **b** optimizing multiple latent codes [25], **c** learning a single latent code encoder [42], and **d** our proposed method OME. From top to bottom are the inversion results on the CelebA-HQ, LSUN church, and LSUN bedroom datasets

Table 2 Quantitative comparisons of ablation experiments on ProGAN

OB	MC	MS	DF	PSNR↑	LPIPS↓	SSIM↑	Time(s)↓
				17.1075	0.3127	0.5142	**0.0219**
	✓			20.0149	0.2234	0.5873	0.0822
		✓		20.2618	0.2189	0.5960	0.1009
		✓	✓	**20.8705**	**0.1865**	**0.6153**	0.1031
✓				18.8219	0.2394	0.5641	**0.0981**
✓	✓			20.5240	0.1993	0.5959	0.4023
✓		✓		20.5681	0.1978	0.5999	0.4988
✓		✓	✓	**21.4278**	**0.1600**	**0.6171**	0.5090

Bold means the best

age weights, which is a simple and brutal strategy. We also analyze the changes of the multi-scale latent codes during offset-based inference.

We conduct ablation experiments mainly on the CelebA-HQ dataset. Table 2 shows the quantitative comparison results of the ablation experiments. We can see that using multiple codes from a single scale (MC) can improve the quality of the synthesized images, and using codes extracted from multiple scales (MS) further extends this improvement. We can also see that the dynamic fusion module (DF) improves the performance of OME considerably with almost no increase in inference time. By comparing the top and bottom parts of the table, it can be seen that the offset-based synthesis method can steadily improve the quality of synthesized images. Figure 5 shows the qualitative comparison results of the ablation experiments. We can see that the use of OB improves the semantic alignment of synthesized images, and the use of DF reduces artifacts in synthesized images, which proves that dynamic fusion is more reasonable than average fusion. In summary, we verified the validity of each feature (OB, MS, and DF) of OME.

We further explore the impact of each feature on OME during training. Figure 6 shows the ID loss during the training process. We can see that the use of multi-scale codes boosts the upper limit of OME's ability to synthesize images. In addition, the dynamic fusion module significantly speeds up the decrease of the loss function, as expected.

Another issue to consider is how many steps we need to iterate in order to get the final latent codes. To explore this, we calculate the average change value of the latent codes for each scale between step t and $t-1$:

$$\mathbf{v}_{m,t} = \frac{1}{QD}\sum_{i=1}^{Q}\sum_{j=1}^{D}|\mathbf{z}_{m,t,i,j} - \mathbf{z}_{m,t-1,i,j}|, \tag{16}$$

where $\mathbf{z}$ means the latent code to be fed to the generator, m and M denote the index and the number of multi-scale latent codes, t denotes the index of inference steps, Q and D denote the number of test images and the number of elements of each latent code, separately.

Then we group the average change of each code at each step by the scale level as shown in Fig. 7. We can see that OME mainly adjusts fine and medium-level codes and stabilizes after a few steps. The reason is that the finer the scale, the more attention is paid to the detailed information of the image (e.g., texture), which is often difficult to synthesize. This phenomenon implies the need to extract codes from multiple scales since a single scale code can only be extracted from the feature map at the coarsest scale. We can also see that the learned offset codes decrease with inference, as we desire. Correspondingly, as shown in Fig. 8, the difference

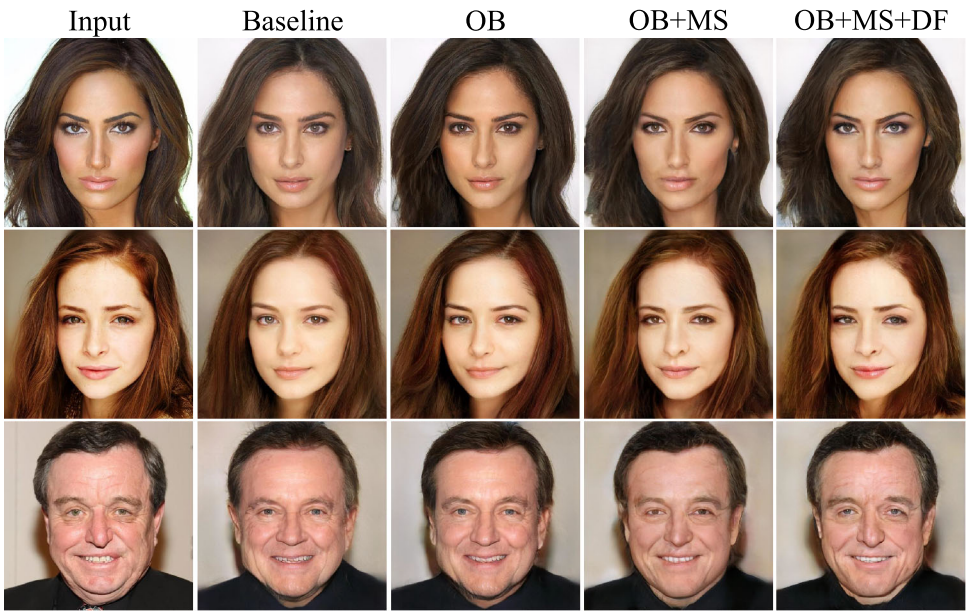

Fig. 5 Qualitative comparisons of ablation experiments on ProGAN

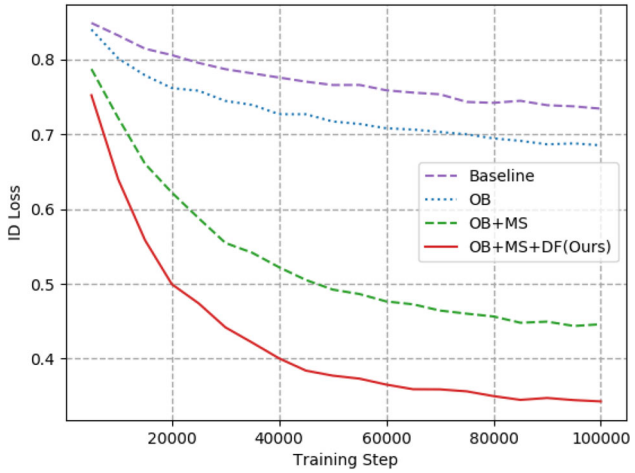

Fig. 6 We add the characteristics of OME one by one and then plot the corresponding ID loss function curves during the training process

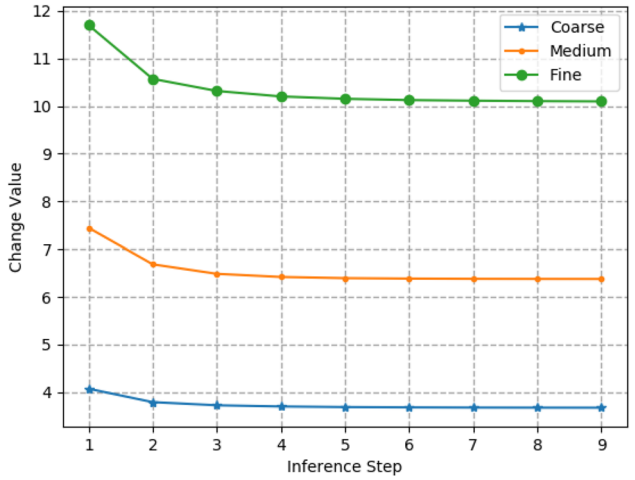

Fig. 7 Average change value of latent codes at different scales during offset-based inference

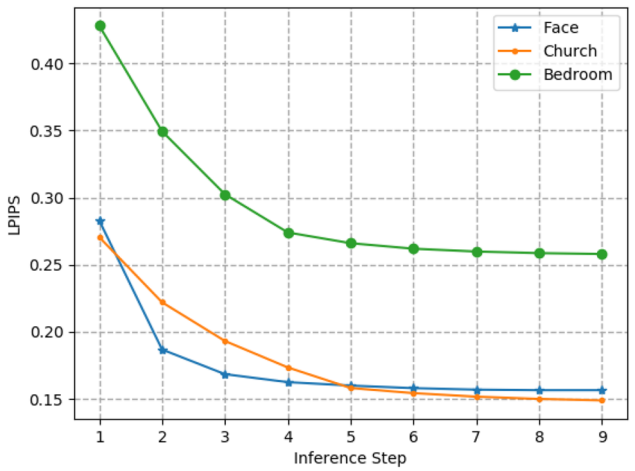

Fig. 8 LPIPS loss curves of different datasets during offset-based inference

between the synthesized image and the target image also decreases as the inference proceeds. Combining Figs. 7 and 8, we set the number of iterative steps $N = 5$ in all tasks.

We also verified the performance of OME based on StyleGAN2. Figure 9 and Table 3 show the qualitative and quantitative comparisons. We can also find that the upper limit of our OME's performance increases based on StyleGAN2's better image generation capabilities.

In this subsection, we verify the performance of OME on the fundamental and important task of GAN inversion. In the following, we demonstrate the effectiveness of OME on the more challenging image-to-image translation tasks of super-resolution as well as conditional face synthesis. To compare with the state-of-the-art StyleGAN2-based pSp, we mainly show the results of OME based on StyleGAN2. Some of the results based on ProGAN are shown in Fig. 13.

4.3 Super-resolution

In this subsection, we verify the performance of our framework on the super-resolution task by comparing it with generic image-to-image translation frameworks (i.e., Pix2PixHD, mGANprior, and pSp) and one of the specialized state-of-the-art approaches, Pulse [7]. During training, we conduct bi-cubic down-sampling for the input image with a random down-sampling factor chosen from 1, 2, 4, 8, 16, and 32. And we set the original image as the target image. To be fair, we use this supervised fashion to train all learning-based models (i.e., Pix2PixHD, pSp and OME).

Results Considering that super-resolution tasks also require consistency between the target image and the synthesized image, we quantitatively and qualitatively validate our approach. Table 4 shows the results of the quantitative comparison measuring PSNR, LPIPS, and PIQE [44]. PIQE is a metric for blind image quality evaluation. As shown, benefiting from the full utilization of pre-trained GANs, our frame performs well, especially when images are down-sampled at high factors. Figure 10 shows the qualitative comparison results, from which it can be seen that thanks to the offset-based image processing method, our results perform better in semantic alignment (e.g., smile degree, hairstyles, sunglasses reconstruction). And surpass the state-of-the-art methods (pSp, mGANprior and Pulse) in image photo-realistic.

4.4 Conditional image synthesis

The purpose of conditional image synthesis is to generate realistic images based on specific input types. In this subsection, we evaluate the performance of our OME framework in two conditional image synthesis tasks: segmentation map-to-face image translation and sketch-to-face translation. The following experiments demonstrate that our method is able to

Fig. 9 Qualitative comparisons of ablation experiments on StyleGAN2

Table 3 Quantitative comparisons of ablation experiments on StyleGAN2

MS	OB	PSNR↑	LPIPS↓	SSIM↑	Time(s)↓
		18.3693	0.2362	0.5611	**0.0643**
✓		20.9662	0.1620	0.6191	0.0805
✓	✓	**22.4731**	**0.1130**	**0.6502**	0.5807

Bold means the best

generate realistic images with higher aligned semantic information than pSp.

4.4.1 Segmentation map-to-face image translation

In this subsection, we verify the performance of OME on the segmentation map-to-face image Translation task. In addition to Pix2PixHD and pSp, we compare our method to ASAP [45], one of the state-of-the-art methods for segmentation map-to-image translation.

Results Thanks to the offset-based image processing method and the use of multi-scale codes, our approach better exploits and controls the generative power of the pre-trained StyleGAN2. Figure 11 provides a comparison on the CelebAMask-HQ dataset. We can see that compared to pSp, our results have better semantic alignment. For instance, hair in columns (a)–(f), clothes in columns (f)–(g), and shoulders in column (h). And we synthesize more realistic images compared to scratch training ASAP and pix2pixHD.

Table 4 Quantitative comparisons of super-resolution tasks using different methods

Factor	Method	PSNR↑	LPIPS↓	PIQE↓
×8	Pix2PixHD	25.0831	**0.1593**	24.1864
	mGANprior	**26.9686**	0.1965	44.2474
	Pulse	22.1668	0.2875	23.5917
	pSp	19.4106	0.2128	25.7223
	Ours	20.1552	0.2103	**19.3761**
×16	Pix2PixHD	22.5389	0.2301	24.5360
	mGANprior	**25.0774**	0.2263	39.7751
	Pulse	20.6444	0.3274	22.6326
	pSp	19.3405	0.2291	25.7913
	Ours	19.9642	**0.2242**	**19.1270**
×32	Pix2PixHD	**22.5389**	0.2970	25.7628
	mGANprior	22.0454	0.2701	36.3191
	Pulse	18.2372	0.3882	21.6511
	pSp	18.8439	0.2611	25.6459
	Ours	18.9831	**0.2568**	**19.8668**

Bold means the best

4.4.2 Sketch-to-face image translation

In this subsection, we verify the performance of OME on the sketch-to-face image translation task, whose input images are sketches and target images are face images. In practical use, only professional painters can construct highly detailed sketches [46]. Therefore, we focus on generating

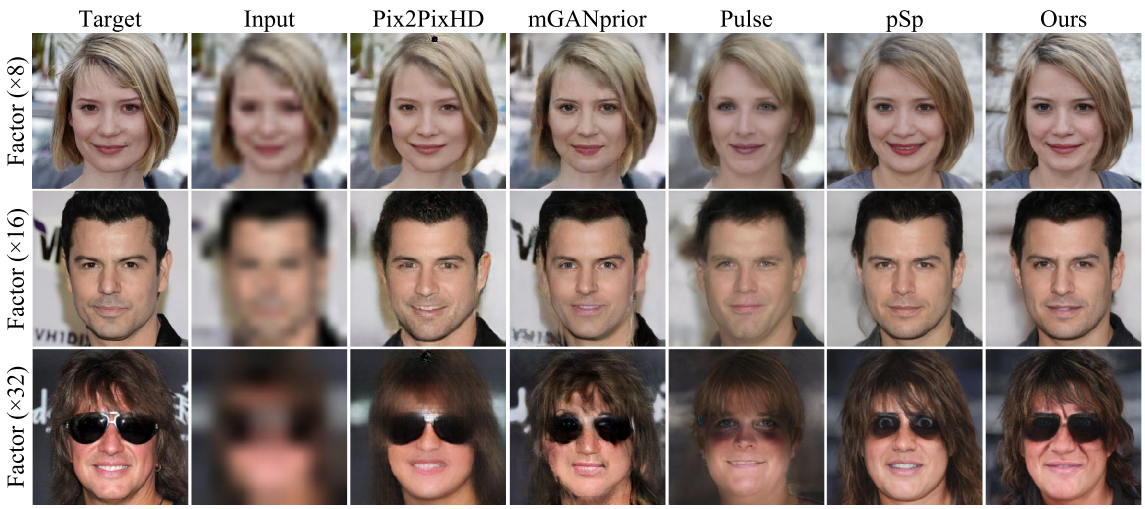

Fig. 10 Qualitative comparisons of super-resolution using different approaches. Please zoom in for details

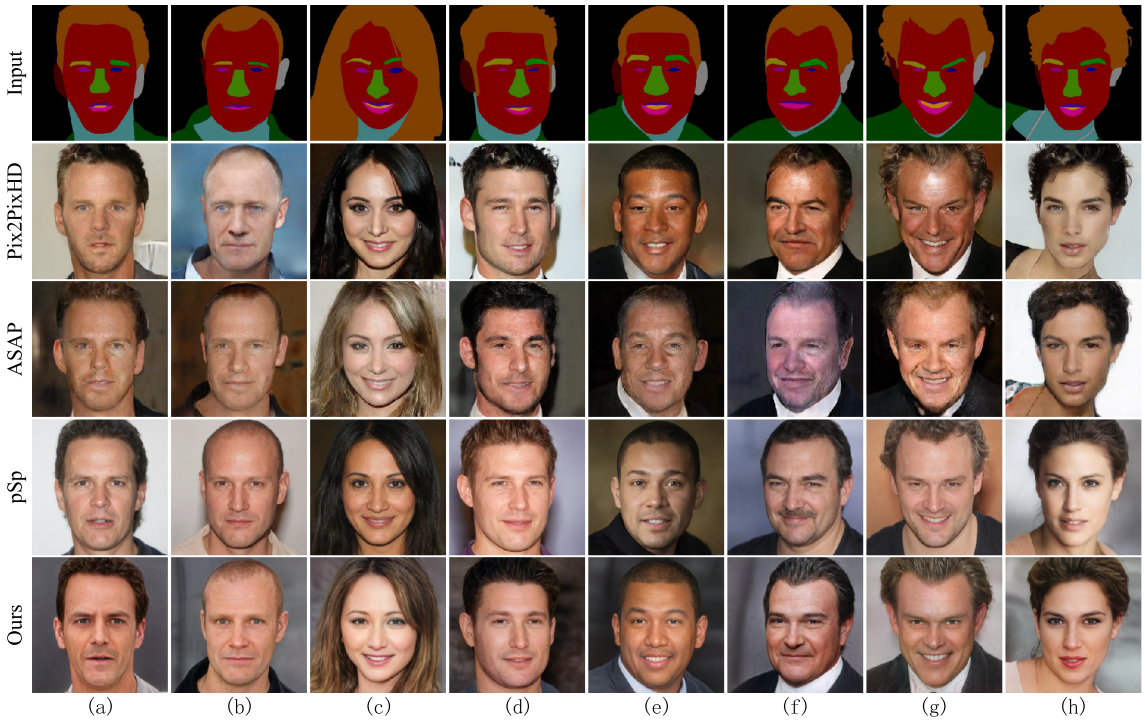

Fig. 11 Comparisons of different segmentation map-to-image methods. Please zoom-in for details

high-quality face images using sparse sketches, which is a more challenging task. We select to construct a dataset based on the CelebA-HQ dataset using the sketch-simplification method by SimoSerra et al. [47]. In addition to pix2pixHD and pSp, we also compare with DeepFaceDrawing (DeepFace) [48], one of the specially designed state-of-the-art methods, which designs specialized mapping networks and loss functions for each organ region to synthesize high-quality faces from sparse sketches.

Results As shown in Fig. 12, we can see that thanks to the excellent generative power of the pre-trained StyleGAN2, pSp and our OME can synthesize more realistic images from sparse sketches than pix2pixHD and DeepFace. Note that DeepFace only works on composing faces from frontal sketches. We can also see that our OME achieves better semantic alignment than pSp. For instance, hair in columns (b) to (d), smile level in columns (e) to (f), glasses in column (g), and eyes in column (h).

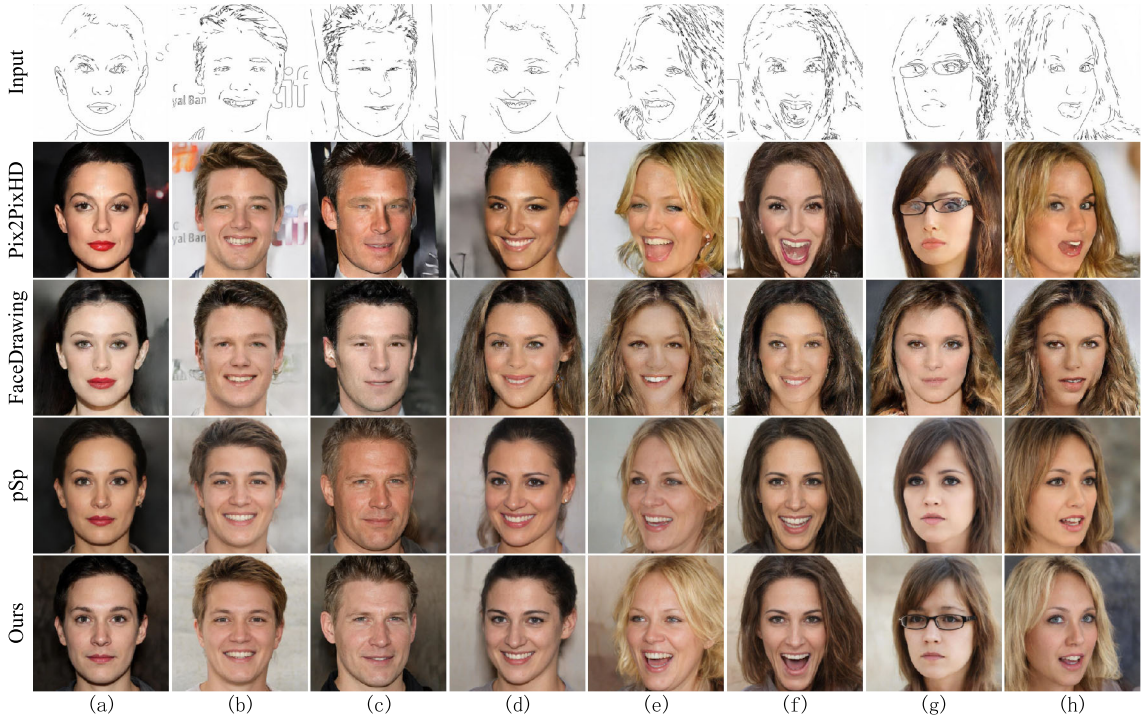

Fig. 12 Comparisons of different Sketch-to-image methods. Please zoom in for details

5 Discussion

Limitations Although our proposed framework has achieved compelling performance in common image-to-image translation tasks, it still has some limitations that are worth considering. First, our framework is limited by the generative power of the GAN since it utilizes a pre-trained GAN to generate high-quality images. Take StyleGAN as an example. As shown in Fig. 14, despite the improved semantic alignment of our OME compared to pSp, neither OME nor pSp can generate hats ideally. This is because only a small number of images in the dataset FFHQ where StyleGAN is trained are hatted, thus causing StyleGAN itself to tend to generate images that are not hatted. From another point of view, during the training period, the decoder of our OME (i.e., StyleGAN) is fixed, and in the latent space of StyleGAN, there are nearly no latent codes to generate hats. Therefore it is difficult for our OME to find such latent codes. In contrast, pix2pixHD can generate hats relatively ideally because the weights of its decoder can be adjusted during the training process to generate output images that correspond better to the input. Second, although our multiple latent codes strategy reduces the loss of details in the images, we still cannot accurately reconstruct some fine details such as earrings, background, etc. The main reasons for this are the inability to control too fine-grained attributes by only modifying latent codes and StyleGAN's limited capacity to generate attributes other than facial. These cases can be seen by comparing the input and rightmost columns in Fig. 9 and the target and rightmost columns in Fig. 10.

Applications When paired with simple digital input devices, designers can easily use OME to complete preliminary design tasks with sketches or semantic segmentation maps as input. This generative power can also be used as an entertainment tool. Our OME can also be used as an image processing tool for tasks such as super-resolution. In conclusion, our study can be used in industries such as entertainment and design. Our work facilitates the work of professionals and enriches people's lives. And taking full advantage of pre-trained GANs can also save a lot of electricity and time consumed in designing and training GANs. From these points, our work can contribute to developing the design and entertainment industries.

6 Conclusion

In this work, we proposed OME, a general image-to-image translation framework based on a multi-scale codes GAN encoder using an iterative update image synthesis method. Compared with previous GAN-based methods, by taking full advantage of pre-trained GANs, our framework can skip the design and training of generators and enjoy the excellent generative power directly, which also saves a lot of electricity and time. Compared with other methods also based on GAN inversion, our framework can better utilize the gen-

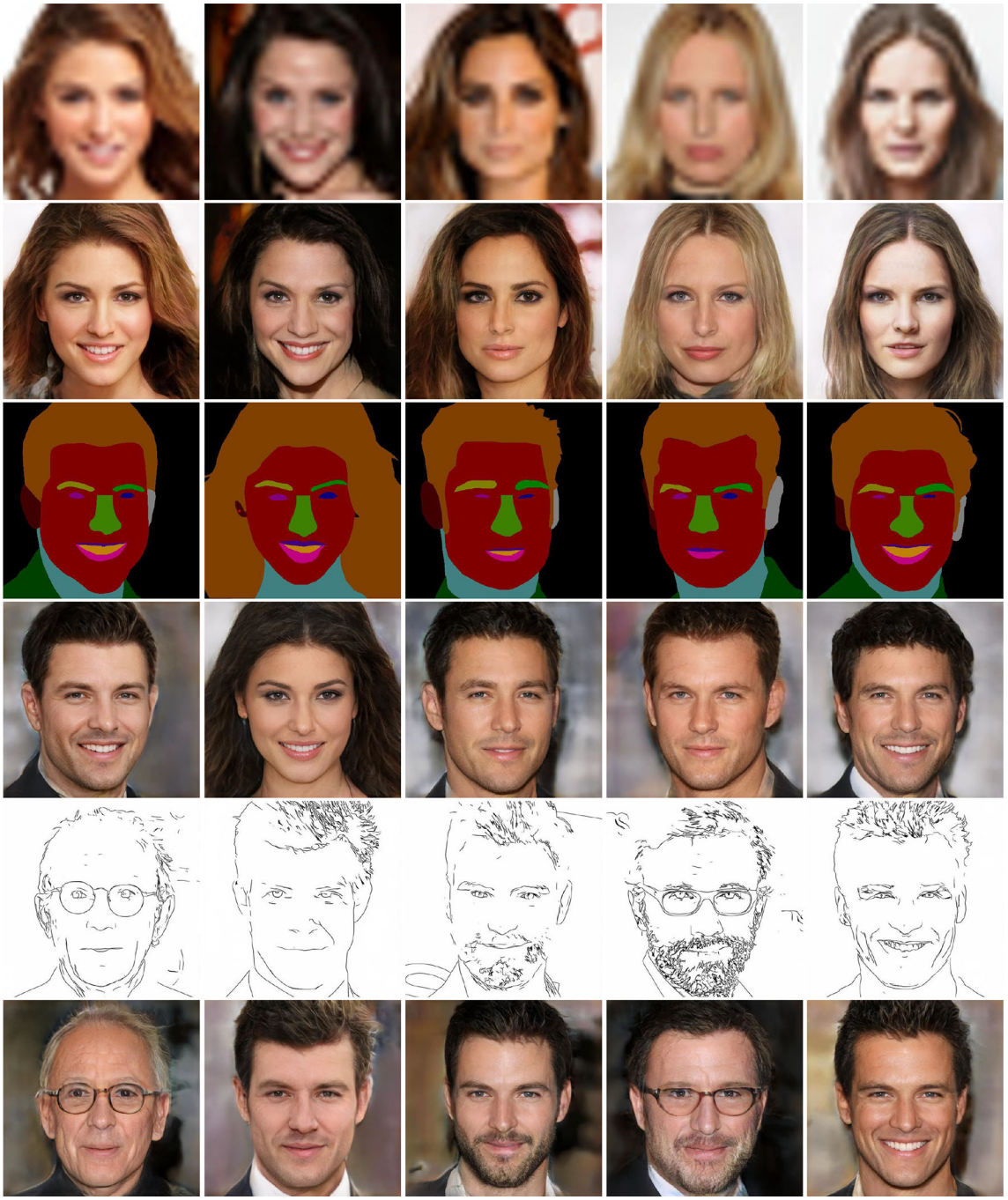

Fig. 13 Our OME can also combine with ProGAN for different image translation tasks. Rows 1, 3, and 5 are the inputs, and rows 2, 4, and 6 are the corresponding outputs. The downsampling factor of the first row is $\times 8$

erative power of pre-trained GANs. Specifically, instead of predicting latent codes in a single shot, our encoder predicts offset-codes based on the gap between the conditional inputs and synthesized images. Then the offset codes modulate the prediction results several times, which improves the quality of synthesized images and mitigates the semantic misalignment between conditional and synthesized images, with a negligible increase in inference. Moreover, unlike conventional single-code encoders, we use multi-scale latent codes to represent an image, which can contain more information and reduce the loss of details in synthesized images. And thanks to our multiple feature dynamic fusion module, our OME can combine with the common pre-trained GANs, showing a certain generalization to GAN models. Through extensive experiments, we prove that our OME is superior to state-of-the-art generic or specialized works. Also, since we

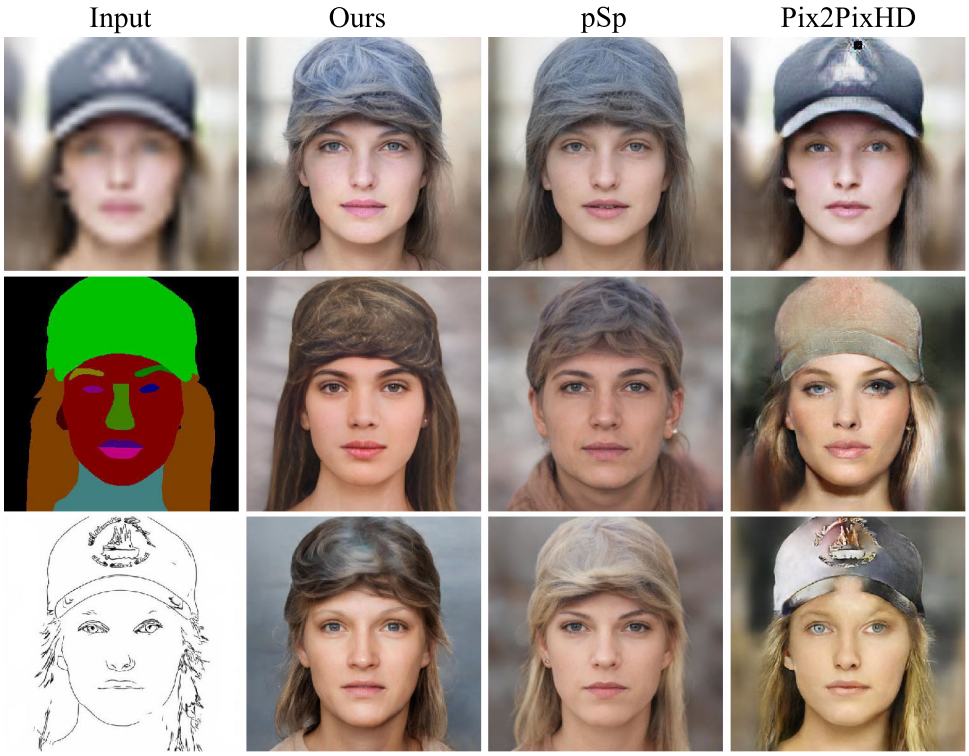

Fig. 14 Limitations of our work. Limited by the generative power of the pre-trained StyleGAN, neither our OME nor pSp can generate hats, while the scratch-trained pix2pixHD can generate hats better

can synthesize high-quality images from conditional images with severe information degradation, we believe our OME can be used in a wider range of image-to-image translation tasks.

Acknowledgements The authors are very indebted to the anonymous referees for their critical comments and suggestions for the improvement of this paper. This work was supported by National Key Research and development Program of China (2021YFA1000102), and in part by the grants from the National Natural Science Foundation of China (Nos. 61673396, 61976245), Natural Science Foundation of Shandong Province (No: ZR2022MF260).

Data Availability Statement The datasets generated or analyzed during this study are available in the CelebA-HQ repository, http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html, and the LSUN repository, https://www.yf.io/p/lsun.

Declarations

Conflict of interest The authors have no financial or proprietary interests in any material discussed in this article.

Ethical approval This work is original research that has not been published before and is not considered for publication elsewhere.

Humans or animal rights This article does not include any studies of humans or animals.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27**, 2672–2680 (2014)
2. Gui, J., Sun, Z., Wen, Y., Tao, D., Ye, J.: A review on generative adversarial networks: Algorithms, theory, and applications. IEEE Transactions on Knowledge and Data Engineering (2021). https://doi.org/10.1109/TKDE.2021.3130191
3. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
4. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
5. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119 (2020)
6. Song, H., Wang, M., Zhang, L., Li, Y., Jiang, Z., Yin, G.: S2rgan: sonar-image super-resolution based on generative adversarial network. The Visual Computer **37**(8), 2285–2299 (2021). https://doi.org/10.1007/s00371-020-01986-3
7. Menon, S., Damian, A., Hu, S., Ravi, N., Rudin, C.: Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In: Proceedings of the Ieee/cvf Conference on Computer Vision and Pattern Recognition, pp. 2437–2445 (2020)
8. Chan, K.C., Wang, X., Xu, X., Gu, J., Loy, C.C.: Glean: Generative latent bank for large-factor image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14245–14254 (2021)

9. Xiu, J., Qu, X., Yu, H.: Double discriminative face super-resolution network with facial landmark heatmaps. The Visual Computer (2022). https://doi.org/10.1007/s00371-022-02701-0
10. Bai, J., Chen, R., Liu, M.: Feature-attention module for context-aware image-to-image translation. The Visual Computer **36**(10), 2145–2159 (2020). https://doi.org/10.1007/s00371-020-01943-0
11. Li, L., Tang, J., Shao, Z., Tan, X., Ma, L.: Sketch-to-photo face generation based on semantic consistency preserving and similar connected component refinement. The Visual Computer, 1–18 (2021). https://doi.org/10.1007/s00371-021-02188-1
12. Reisfeld, E., Sharf, A.: Onesketch: learning high-level shape features from simple sketches. The Visual Computer (2022). https://doi.org/10.1007/s00371-022-02494-2
13. Kang, H.W., He, W., Chui, C.K., Chakraborty, U.K.: Interactive sketch generation. The Visual Computer **21**(8), 821–830 (2005). https://doi.org/10.1007/s00371-005-0328-9
14. Shao, M., Zhang, Y., Liu, H., Wang, C., Li, L., Shao, X.: Dmdit: Diverse multi-domain image-to-image translation. Knowledge-Based Systems **229**, 107311 (2021). https://doi.org/10.1016/j.knosys.2021.107311
15. Shao, M., Zhang, Y., Fan, Y., Zuo, W., Meng, D.: Iit-gat: Instance-level image transformation via unsupervised generative attention networks with disentangled representations. Knowledge-Based Systems **225**, 107122 (2021)
16. Song, X., Shao, M., Zuo, W., Li, C.: Face attribute editing based on generative adversarial networks. Signal, Image and Video Processing **14**(6), 1217–1225 (2020)
17. Xia, W., Zhang, Y., Yang, Y., Xue, J.-H., Zhou, B., Yang, M.-H.: Gan inversion: A survey. arXiv preprint arXiv:2101.05278 (2021)
18. Ma, F., Ayaz, U., Karaman, S.: Invertibility of convolutional generative networks from partial measurements. Advances in Neural Information Processing Systems **31**, 9651–9660 (2018)
19. Creswell, A., Bharath, A.A.: Inverting the generator of a generative adversarial network. IEEE transactions on neural networks and learning systems **30**(7), 1967–1974 (2018)
20. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4432–4441 (2019)
21. Zhu, J.-Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: European Conference on Computer Vision, pp. 597–613 (2016). Springer
22. Bau, D., Zhu, J.-Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., Torralba, A.: Seeing what a gan cannot generate. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4502–4511 (2019)
23. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6711–6720 (2021)
24. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
25. Gu, J., Shen, Y., Zhou, B.: Image processing using multi-code gan prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3012–3021 (2020)
26. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
27. Li, L., Tang, J., Ye, Z., Sheng, B., Mao, L., Ma, L.: Unsupervised face super-resolution via gradient enhancement and semantic guidance. The Visual Computer **37**(9), 2855–2867 (2021). https://doi.org/10.1007/s00371-021-02236-w
28. Fan, Y., Shao, M., Zuo, W., Li, Q.: Unsupervised image-to-image translation using intra-domain reconstruction loss. International Journal of Machine Learning and Cybernetics **11**(9), 2077–2088 (2020)
29. Lan, J., Ye, F., Ye, Z., Xu, P., Ling, W.-K., Huang, G.: Unsupervised style-guided cross-domain adaptation for few-shot stylized face translation. The Visual Computer (2022). https://doi.org/10.1007/s00371-022-02719-4
30. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2287–2296 (2021)
31. Jie, Shen: Samuel, Albanie, Gang, Sun, Enhua: Squeeze-and-excitation networks. IEEE transactions on pattern analysis and machine intelligence (2019). https://doi.org/10.1109/TPAMI.2019.2913372
32. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
33. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
34. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25**, 84–90 (2012)
35. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
36. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
37. Wright, L.: Ranger - a synergistic optimizer. GitHub (2019). https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer
38. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv arXiv preprint arXiv:1710.10196 (2017). http://arxiv.org/abs/1710.10196
39. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
40. Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8798–8807 (2018)
41. Ma, F., Ayaz, U., Karaman, S.: Invertibility of convolutional generative networks from partial measurements. Advances in Neural Information Processing Systems **31**, 9651–9660 (2018)
42. Zhu, J.-Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: European Conference on Computer Vision, pp. 597–613 (2016). Springer
43. Mohammadi, P., Ebrahimi-Moghadam, A., Shirani, S.: Subjective and objective quality assessment of image: A survey. Majlesi Journal of Electrical Engineering **9**, 55–83 (2014)
44. Venkatanath, N., Praneeth, D., Bh, M.C., Channappayya, S.S., Medasani, S.S.: Blind image quality evaluation using perception based features. In: 2015 Twenty First National Conference on Communications (NCC), pp. 1–6 (2015). https://doi.org/10.1109/NCC.2015.7084843
45. Shaham, T.R., Gharbi, M., Zhang, R., Shechtman, E., Michaeli, T.: Spatially-adaptive pixelwise networks for fast image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14882–14891 (2021)

46. Wang, X., Tang, X.: Face photo-sketch synthesis and recognition. IEEE transactions on pattern analysis and machine intelligence **31**(11), 1955–1967 (2008)
47. Simo-Serra, E., Iizuka, S., Sasaki, K., Ishikawa, H.: Learning to simplify: fully convolutional networks for rough sketch cleanup. ACM Transactions on Graphics (TOG) **35**(4), 1–11 (2016)
48. Chen, S.-Y., Su, W., Gao, L., Xia, S., Fu, H.: Deepfacedrawing: Deep generation of face images from sketches. ACM Transactions on Graphics (TOG) **39**(4), 72–1 (2020)

Zihao Guo received the B.Eng. degree from the College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China, in 2020, where he is currently pursuing the M.S. degree under the supervision of Prof. M. Shao. His current research interests include generative adversarial networks, computer vision, and deep learning.

Mingwen Shao received the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2005. He is currently a Professor and a Doctoral Supervisor with the China University of Petroleum (East China). He has published over 100 papers in national and international journals and conferences in research areas including deep learning, computer vision and data mining.

Shunhang Li received the B.Eng. degree from the College of Computer Science and Technology, China University of Petroleum (East China), Qingdao, China, in 2020, where he is currently pursuing the M.S. degree under the supervision of Prof. M. Shao. His current research interests include knowledge distillation, computer vision, and deep learning.