Computational Visual Media
https://doi.org/10.1007/s41095-023-0371-3 Vol. 10, No. 6, December 2024, 1031–1043

Research Article

Taming diffusion model for exemplar-based image translation

Hao Ma[1], **Jingyuan Yang**[1], **and Hui Huang**[1] (✉)

© The Author(s) 2024.

Abstract Exemplar-based image translation involves converting semantic masks into photorealistic images that adopt the style of a given exemplar. However, most existing GAN-based translation methods fail to produce photorealistic results. In this study, we propose a new diffusion model-based approach for generating high-quality images that are semantically aligned with the input mask and resemble an exemplar in style. The proposed method trains a conditional denoising diffusion probabilistic model (DDPM) with a SPADE module to integrate the semantic map. We then used a novel contextual loss and auxiliary color loss to guide the optimization process, resulting in images that were visually pleasing and semantically accurate. Experiments demonstrate that our method outperforms state-of-the-art approaches in terms of both visual quality and quantitative metrics.

Keywords exemplar; image translation; denoising diffusion probabilistic model (DDPM)

1 Introduction

Image-to-image translation aims to learn the mapping between two distinct domains, covering various tasks, such as image synthesis, image inpainting, and image manipulation. In this study, we focus on converting semantic masks into natural-looking images with a given exemplar image, which is called exemplar-based image translation [1].

Existing approaches often construct networks based on generative adversarial networks (GANs) [1–4]. Despite their great advantages, GANs often feature low-mode coverage, leading to a low diversity of the generated images. However, denoising diffusion probabilistic model (DDPM) has accurate mode coverage and can thus, produce high-quality images, which makes them superior for image generation and manipulation tasks [5–8].

In this study, we presented a DDPM-based approach for generating exemplar-based images. To generate semantically consistent images, we arranged a conditional DDPM with the SPADE module [4] and trained it in a classifier-free manner. In particular, we designed two critical losses to guide the entire network jointly, resulting in style-faithful and photorealistic outputs (see Fig. 1). Our main contributions are summarized as follows.



Fig. 1 Exemplar-based image synthesis. Our method translates the input masks to realistic images. Experiments are conducted on CelebAMask-HQ, ADE20K, and Cityscapes.

1 Visual Computing Research Center, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China. E-mail: H. Ma, mahao_fido@szu.edu.cn; J. Yang, jyyang@szu.edu.cn; H. Huang, huihuang@szu.edu.cn (✉).

Manuscript received: 2023-03-13; accepted: 2023-07-20

- To align the synthesized image with the input semantic map, we apply SPADE to the diffusion model, which enables the network to grasp the semantic layout.
- We propose a novel contextual loss and an auxiliary color loss to assist the translation process, aiming to guide the network to output images with style consistency.
- Experiments validate that our approach enables photo-realistic image generation and is competitive with leading methods qualitatively and quantitatively.

2 Related work

Image-to-image translation. Image reconstruction tasks emphasize the similarities between the input and output images, and the similarity metrics at the concrete appearance and abstract perceptual levels are carefully designed to guide the model in generating the desired images. By contrast, image-to-image translation aims to map images from different domains. From the perspective of network design, the existing prominent approaches can be conditional or unconditional, depending on how they leverage images or other information. pSp [9] is an end-to-end image-to-image translation framework that uses the representative power of a pretrained StyleGAN generator. Using a strong encoder, pSp can match each input image to corresponding code in the latent domain and supports multimodal synthesis via style resampling. Most image-to-image translation networks are conditional. Isola et al. [2] investigated conditional adversarial networks to solve image-to-image translation problems, and successfully applied their framework to various novel translation tasks. SEAN [10] can control the style of each semantic region individually owing to its normalization, which can extract style reference images per semantic region. Instead of directly feeding the semantic layout as input to the deep network, SPADE [4] modulates the activations in the normalization layers through a spatially adaptive learned transformation. Considering the remarkable performance of SPADE, researchers have applied it to networks to preserve semantics.

Models for unpaired training data were explored to relax the strict constraints of paired data. Kim et al. [11] proposed a method based on GANs to learn the relationships between different domains; thus, images in one domain can be translated to another. Yi et al. [12] developed a novel dual-GAN mechanism that enables image translators to train using unlabeled images from two domains. Cross-GAN assumes a shared latent space and proposes an unsupervised image-to-image translation framework based on Coupled GANs [13]. The limitations of Cross-GAN are the unimodal assumption of the Gaussian latent space and unstable training. Zhu et al. [3] integrated the cycle-consistent loss into GANs to achieve unpaired image-to-image translation. Royer et al. [14] introduced a dual adversarial autoencoder that captures a shared representation of the common domain semantic content and jointly learns domain-to-domain image translations.

Exemplar-based image synthesis. An exemplar offers contextual information for image synthesis [1, 15–18]. Therefore, leveraging a given exemplar is vital. Huang and Belongie [15] presented a novel adaptive instance normalization (AdaIN) layer that enables arbitrary style transfer by adjusting the mean and variance of the content input to match those of the style input. Similarly, Ma et al. [16] introduced AdaIN to transfer the style information of a target domain to the source domain and applied the concept of feature masks to avoid semantic inconsistencies. Qi et al. [17] designed a semiparametric approach that synthesizes images by compositing image fragments retrieved from a large database. A style consistency discriminator was proposed in Ref. [18] to determine whether the image pairs are consistent in style. However, their results could be affected by the performance of the state-of-the-art semantic labeling functions. SPADE learns an encoder to map the exemplar image into a low-dimensional vector and then delivers this vector to a generator accompanied by an input semantic image [4]. CoCosNet [1] aligns images from distinct domains to an intermediate domain, and then synthesizes images based on the appearance of semantically corresponding patches in the exemplar. However, this method may produce unsatisfactory results owing to one-to-many and many-to-one mappings, as mentioned in their paper. UNITE [19] introduces an optimal transport for feature matching, which treats the entire feature set as a whole and effectively overcomes many-to-one matching. However, this method cannot generate the correct semantics when there are no corresponding

semantics in the exemplar.

DDPM-based image synthesis. Compared with GANs, DDPM has achieved unprecedented success in various image-generation tasks [6, 20–26]. Xiao et al. [26] summarized the manner in which mainstream generative networks address the trilemma of sampling efficiency, quality, and diversity. GANs rapidly generate high-quality samples; however, they exhibit defects in the mode coverage. Although variational autoencoders (VAE) and normalizing flows fully and faithfully cover the data modes, they often produce low-quality samples. DDPM has emerged as a powerful generative model. These models demonstrated surprisingly good results in terms of sample quality, beating GANs in image generation. Enforced by high likelihood, they also have good mode coverage and sample diversity. DDIM [23] was proposed to reduce the sampling time with thousands of neural network inferences. In summary, DDPM has been increasingly developed and applied to tasks involving GAN, such as image manipulation, image inpainting, and text-to-image generation.

3 Method

3.1 DDPM

As shown in Fig. 2, DDPM [20] involves two processes: diffusion (forward) and denoising (reverse). Training data $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$, $\boldsymbol{x}_1, ..., \boldsymbol{x}_t$ are intermediate latents that share the same dimensions as data $\boldsymbol{x}_0$. The reverse process is defined as a Markov chain with a learned Gaussian transition starting at $p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_T; \boldsymbol{0}, \boldsymbol{I})$ and is formulated as the joint distribution $p_\theta(\boldsymbol{x}_{0:T})$:

$$\begin{cases} p_\theta(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \\ p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) := \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)) \end{cases} \tag{1}$$

The forward process gradually adds Gaussian noise to the data according to a precalculated schedule, which can also be learned through reparameterization $\beta_1, ..., \beta_T$:

$$\begin{cases} q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) := \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \\ q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}) \end{cases} \tag{2}$$

One property of the forward process is that $\boldsymbol{x}_t$ can be deduced in closed form:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_0, (1-\alpha_t)\boldsymbol{I}) \tag{3}$$

where $\alpha_t := \prod_{s=1}^{t}(1-\beta_s)$. Furthermore, the posterior given $\boldsymbol{x}_t$ and $\boldsymbol{x}_0$ is tractable:

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \widetilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0), \widetilde{\beta}_t \boldsymbol{I}) \tag{4}$$

where

$$\widetilde{\boldsymbol{\mu}}_t(\boldsymbol{x}_t, \boldsymbol{x}_0) := \frac{\sqrt{\alpha_{t-1}}\beta_t}{1-\alpha_t}\boldsymbol{x}_0 + \frac{\sqrt{1-\beta_t}(1-\alpha_{t-1})}{1-\alpha_t}\boldsymbol{x}_t \tag{5}$$

$$\widetilde{\beta}_t := \frac{(1-\alpha_{t-1})}{1-\alpha_t}\beta_t \tag{6}$$

To optimize DDPM, a surrogate objective that computes the standard mean-squared error loss between the estimated noise and normal noise is now widely adopted:

$$L_{\text{simple}} := E_{t\sim[1,T], \boldsymbol{x}_0\sim q(\boldsymbol{x}_0), \boldsymbol{\epsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}[\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{\epsilon}\|^2] \tag{7}$$

Here, $\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{\epsilon}\|$ is the mean-squared error loss between the true noise and the noise predicted from UNet [24]. UNet is the backbone of DDPM, whose input and output share the same dimensions. After estimating the true noise, we can predict the start variable $\hat{\boldsymbol{x}}_0$ as

$$\hat{\boldsymbol{x}}_0 = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x}_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)) \tag{8}$$

Additionally, DDIM makes deterministic diffusion and inference possible through Eqs. (9) and (10) [5, 23]

$$\text{Diffusion: } \boldsymbol{x}_{t+1} = \sqrt{\alpha_{t+1}}\hat{\boldsymbol{x}}_0 + \sqrt{1-\alpha_{t+1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \tag{9}$$

$$\text{Inference: } \boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\hat{\boldsymbol{x}}_0 + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \tag{10}$$

where $\hat{\boldsymbol{x}}_0$ is estimated from $\boldsymbol{x}_t$ using Eq. (8). Using deterministic diffusion and inference processes, face manipulation with identity preservation can be achieved through DDIM [6].

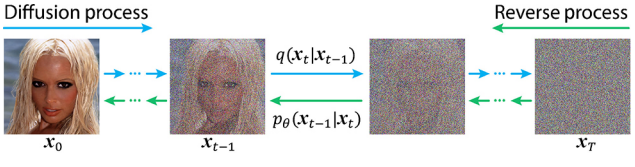


Fig. 2 Diffusion process and reverse process of DDPM.

3.2 Approach

In our study, DDPM is conditioned on a semantic map $\boldsymbol{y}$ which is integrated such that SPADE is developed. The proposed conditional DDPM follows the network

design of Ref. [27]. Because unconditional DDPM can generate higher-quality samples than conditional DDPM, classifier-free techniques have been further designed for DDPM [22]. Specifically, during training, a dropout probability is set to help determine whether condition $\boldsymbol{y}$ should be dropped. When being dropped, the value of $\boldsymbol{y}$ is set to zero and denoted as $\emptyset$. Thus, the variable tuple $(\boldsymbol{x}_t, t, \boldsymbol{y})$ or $(\boldsymbol{x}_t, t, \emptyset)$ is fed into the UNet of the conditional DDPM. During inference, the score is calculated using a linear combination of conditional and unconditional estimates as Eq. (11):

$$\widetilde{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t) = (1+w)\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{y}) - w\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \emptyset) \quad (11)$$

The exemplar-based image-translation method is illustrated in Fig. 3. After training the conditional DDPM, contextual and color diffusion guidance (CCDG) is introduced to further adjust the pretrained model for the desired outputs. Next, we introduce contextual loss and color loss and design a specific algorithm.

Contextual loss. In Algorithm 1, the contextual loss [28] is proposed to measure the similarity in appearance between two images. For clarity, we introduced notations in Ref. [28]. First, $X = \{x_i\}$ and $Y = \{y_j\}$ are the sets of feature maps of the two images x and y respectively. Both sets contain the same number of sample points. Let d_{ij} be the cosine distance between x_i and y_j, x_i and y_j are considered as similiar features when $d_{ij} \ll d_{ik}$ for $\forall k \neq j$. The original measurement is defined as

$$\mathcal{L}_{\mathrm{CX}}(x, y, l) := -\log \mathrm{CX}(\phi^l(x), \phi^l(y)) \quad (12)$$

where $\phi^l(x)$ and $\phi^l(y)$ are the feature maps from layer l of the pretrained perceptual network ϕ which is VGG19 [29]. To facilitate the introduction of the new measurement method, we define

$$\mathrm{CX}(x_i, Y) := \max_j \mathrm{CX}(\phi_i^l(x), \phi_j^l(y)) \quad (13)$$

further,

$$\begin{aligned}\mathrm{CX}(\phi^l(x), \phi^l(y)) &= \frac{1}{n_l}\sum_i \mathrm{CX}(x_i, Y) \\ &= \frac{1}{n_l}\sum_i \max_j \mathrm{CX}(\phi_i^l(x), \phi_j^l(y))\end{aligned} \quad (14)$$

where n_l is the number of feature maps in the VGG net (i and j are indices). The original contextual loss over all the layers between the predicted image $\hat{\boldsymbol{x}}_{B,0}$ and the given exemplar $\boldsymbol{z}_B$ in our case is

$$\mathcal{D}_{\mathrm{orig_context}} := \sum_l w_l \left[-\log\left(\frac{1}{n_l}\sum_i \max_j \mathrm{CX}(\phi_i^l(\hat{\boldsymbol{x}}_{B,0}), \phi_j^l(\boldsymbol{z}_B))\right)\right] \quad (15)$$

w_l controls the importance of layer l. Before introducing our novel contextual loss function, we borrow related original definitions (note that, following the original paper, the layer notation is ignored in the next few equations).

$$\widetilde{d}_{ij} = \frac{d_{ij}}{\min_k d_{ik} + \epsilon} \quad (16)$$

$$w_{ij} = \exp\left(\frac{1 - \widetilde{d}_{ij}}{h}\right) \quad (17)$$

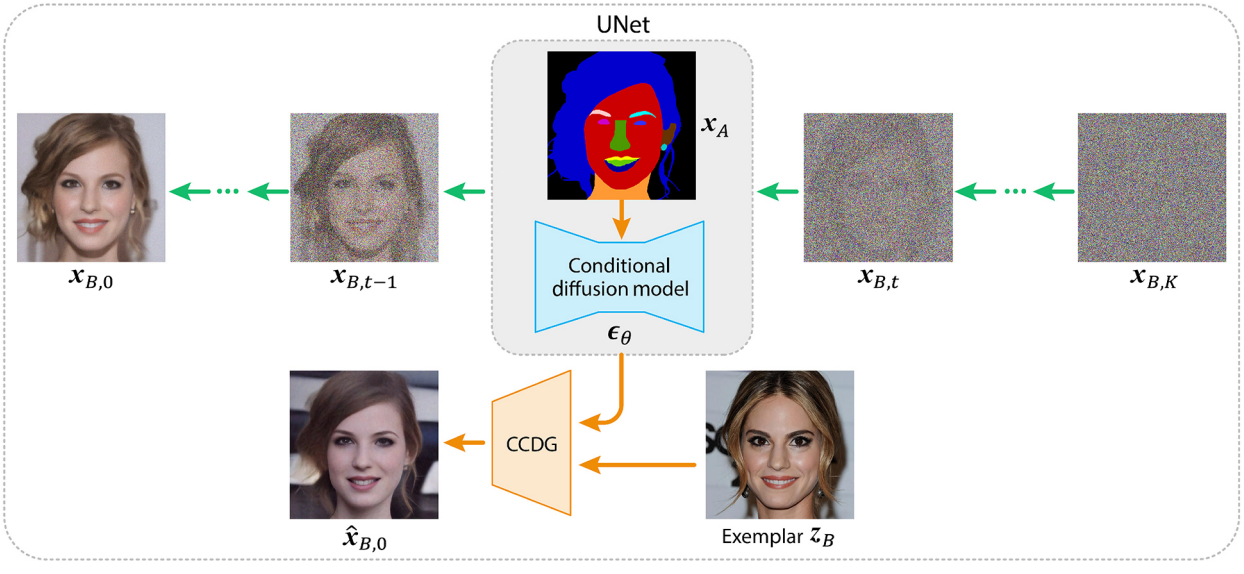


Fig. 3 Overview of our method. In each timestep, the pretrained conditional DDPM is guided by CCDG.

$$\mathrm{CX}_{ij} = w_{ij} / \sum_k w_{ik} \tag{18}$$

In our study, we redefine $\mathrm{CX}(x_i, Y)$ to constrain the similarity between the feature maps more effectively.

$$\mathrm{CX}(x_i, Y) = 1 - \frac{d_{ij}}{\min\limits_{k \neq j} d_{ik}} \tag{19}$$

According to this new equation, $\mathrm{CX}(x_i, Y)$ is large when the ratio of the second minimum distance and the minimum distance is relatively large (and vice versa). It is noteworthy that the original equation in Eq. (13) considers only one general distance that is processed uniformly inside. Our novel contextual loss function is defined as

$$\mathcal{D}_{\text{our_context}} := \sum_l w_l \left[-\log \left(\frac{1}{n_l} \sum_i \mathrm{CX}(\phi_i^l(\hat{\boldsymbol{x}}_{B,0}), \phi^l(\boldsymbol{z}_B)) \right) \right] \tag{20}$$

As detailed in Ref. [1], the low-level features in contextual loss contain richer style information, which is decisive for the transfer process of an exemplar appearance.

Color loss. We noticed that contextual loss alone is not powerful enough to maintain color consistency between the translated image and the exemplar. Thus, the color loss was designed to maintain consistency at the semantic region level. In CelebAMask-HQ dataset, the semantic map had 19 channels, each of which is set to 0 or 1, indicating invalid or valid, respectively. Given a semantic input $\boldsymbol{x}_A$, an exemplar $\boldsymbol{z}_B$, and its semantic map $\boldsymbol{z}_A$, we calculated the color distance for each semantic channel and set its weight using valid pixels. We determine that $w_c = \Sigma \boldsymbol{x}_A^c / (H \times W)$, where H and W denote the height and width of the input data respectively. Symbol $\cdot$ represents an element-wise multiplier. $\Sigma \boldsymbol{x}_A^c$ or $\Sigma \boldsymbol{z}_A^c$ measures the number of valid pixels in channel c. $\hat{\boldsymbol{x}}_{B,0}$ is the predicted image, and $\mathbb{1}$ is an indicator function whose value is zero if the inside value is $\mathbf{0}$:

$$\mathcal{D}_{\text{color}} := \sum_c \left[\left(w_c \left| \frac{\Sigma(\boldsymbol{x}_A^c \cdot \hat{\boldsymbol{x}}_{B,0})}{\Sigma \boldsymbol{x}_A^c} - \frac{\Sigma(\boldsymbol{z}_A^c \cdot \boldsymbol{z}_B)}{\Sigma \boldsymbol{z}_A^c} \right| \right) \mathbb{1}[\boldsymbol{x}_A^c \cdot \boldsymbol{z}_A^c] \right] \tag{21}$$

Algorithm. Our method is illustrated in Algorithm 1. The reverse process of the diffusion model can be implemented using DDIM with a smaller timestep K compared with the training process. To reduce the influence of random noise and synthesize images, DDIM was adopted during the translation process, which is further illustrated in Eqs. (9) and (10). Specifically, the exemplar image $\boldsymbol{z}_B$ is first encoded in the latent code space, leading to noisy data $\boldsymbol{x}_{B,K}$. Next, we iteratively refined the output of UNet through CCDG optimization. In our implementation, $\lambda_{\text{ctx_w}}$, $\lambda_{\text{color_w}}$, and w were set to 1, 10, and 0 on CelebAMask-HQ and Cityscapes and to 0.1, 10, and 1 on ADE20K by default, respectively.

Algorithm 1 Exemplar-based image synthesis via DDIM, given a diffusion mode $\boldsymbol{\epsilon}_\theta$ conditioned on semantic inputs

Input: source semantic image $\boldsymbol{x}_A$, exemplar $\boldsymbol{z}_B$, exemplar's semantic map $\boldsymbol{z}_A$, diffusion steps K, and $\boldsymbol{z}_{B,0} \leftarrow \boldsymbol{z}_B$
Output: synthesized image $\hat{\boldsymbol{x}}_B$
for t *from 0 to* $K-1$ **do**
  $\hat{\boldsymbol{z}}_{B,0} = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{z}_{B,t} - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_{B,t}, t, \boldsymbol{z}_A))$;
  $\boldsymbol{z}_{B,t+1} = \sqrt{\alpha_{t+1}}\hat{\boldsymbol{z}}_{B,0} + \sqrt{1-\alpha_{t+1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_{B,t}, t, \boldsymbol{z}_A)$;
end for
$\boldsymbol{x}_{B,K} \leftarrow \boldsymbol{z}_{B,K}$;
for t *from* K *to 1* **do**
  $\widetilde{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_{B,t}, t) = (1+w)\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_{B,t}, t, \boldsymbol{x}_A) - w\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_{B,t}, t, \emptyset)$;
  $\hat{\boldsymbol{x}}_{B,0} = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x}_{B,t} - \sqrt{1-\alpha_t}\widetilde{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_{B,t}, t))$;
  $\mathcal{L} \leftarrow \lambda_{\text{ctx_w}}\mathcal{D}_{\text{our_context}}(\hat{\boldsymbol{x}}_{B,0}, \boldsymbol{z}_B)$
    $+\lambda_{\text{color_w}}\mathcal{D}_{\text{color}}(\boldsymbol{x}_A, \hat{\boldsymbol{x}}_{B,0}, \boldsymbol{z}_A, \boldsymbol{z}_B)$;
  $\hat{\boldsymbol{\epsilon}} \leftarrow \widetilde{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_{B,t}, t) - \sqrt{1-\alpha_t}\boldsymbol{\nabla}_{\hat{\boldsymbol{x}}_{B,0}}\mathcal{L}$;
  $\boldsymbol{x}_{B,t-1} \sim \sqrt{\alpha_{t-1}}\hat{\boldsymbol{x}}_{B,0} + \sqrt{1-\alpha_{t-1}}\hat{\boldsymbol{\epsilon}}$;
end for
return $\hat{\boldsymbol{x}}_B \leftarrow \boldsymbol{x}_{B,0}$

4 Results and evaluation

Datasets. Our method is validated using three datasets to demonstrate its effectiveness.

- CelebAMask-HQ [30] consists of 30,000 high-resolution (1024×1024) face images selected from CelebA [31] and further post-processed by Ref. [32]. Each image has a corresponding segmentation mask of facial attributes corresponding to CelebA. The 512 × 512 masks were manually annotated with 19 classes including all facial components and accessories. CelebAMask-HQ can be used to train and evaluate algorithms for face parsing, face generation, and face editing.
- Cityscapes [33] collects 3000 training and 500 validation urban street scene images. The dataset has 35 classes and the image size is 2048×1024. During training, the aspect ratio of processed images is the same as the original.

- ADE20K [34] contains 20,210 training and 2000 validation images. Each image is associated with one segmentation mask and the dataset contains indoor and outdoor scenes with 150 semantic classes.

Ablation study. To evaluate the effectiveness of the proposed loss function, we constructed six experimental settings: (1) a semantic map (first column of Fig. 4) and an exemplar (second column of Fig. 4). We encode the exemplar into the latent code space and decode the code back into the image space without optimization. Contextual loss is introduced to help the model generate images that share a similar appearance with the exemplars. (2) The original contextual loss, (3) our novel contextual loss, (4) the proposed color loss alone, (5) the original contextual loss and color loss, and (6) our novel contextual loss and color loss optimize the model to reach the desired outputs. The results of these settings are presented in Fig. 4 from the third to the last columns. Directly feeding the exemplar into the pretrained model without optimization led to obvious style differences between the translated image and the exemplar in terms of hair color in the generated images. Contextual loss boosts the performance, and our novel contextual loss produces images that match exemplars better than the original contextual loss through close observation. The results in the last five columns are visually similar but have subtle differences. These differences are quantitatively presented in Table 1. In Fig. 4, the resolution of the synthesized images is 64×64, which is visually blurred. We also provide super-resolution results in other figures when appropriate (curated results are

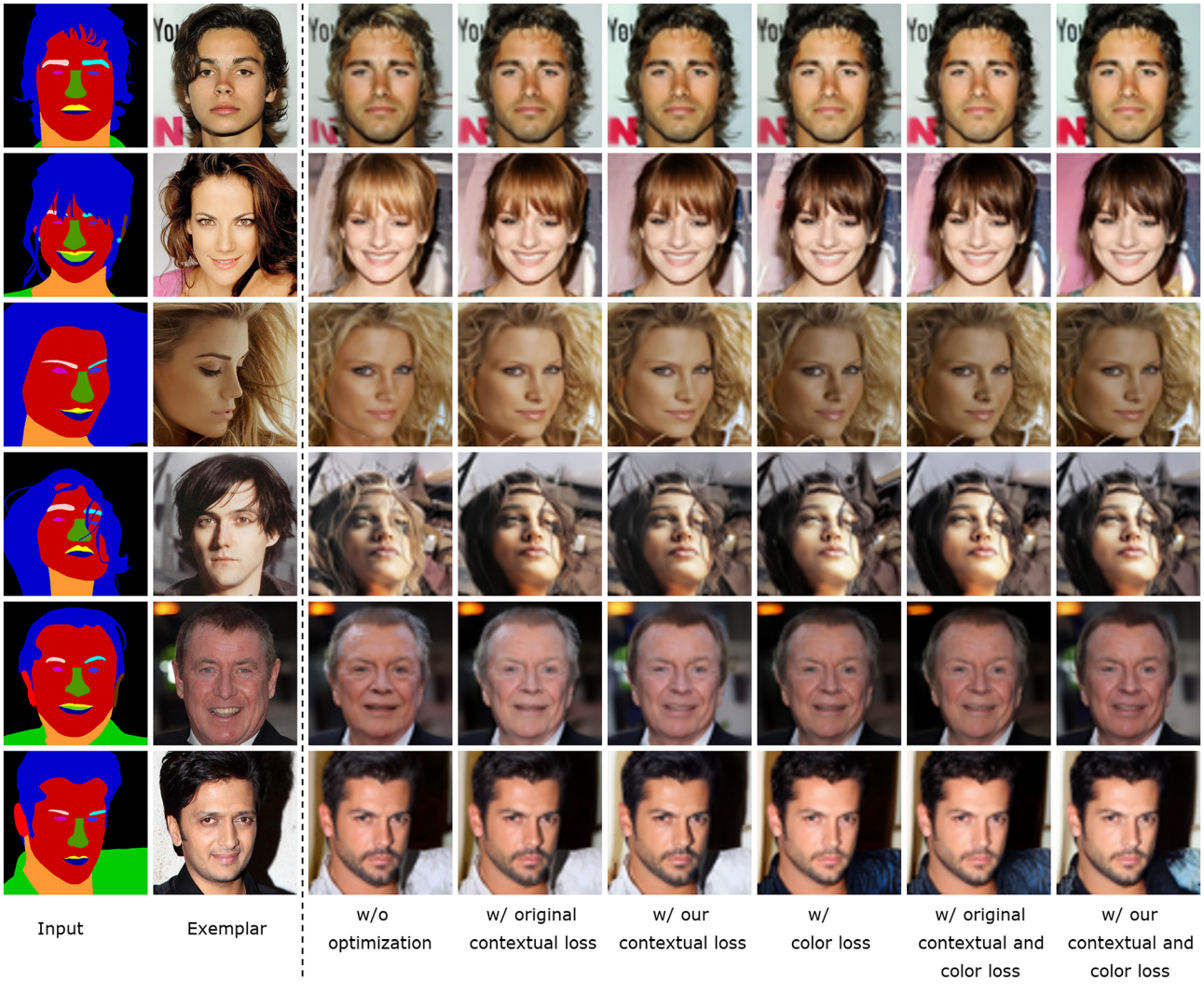


Fig. 4 Ablation study on loss terms. Six scenarios are considered.

Table 1 Quantitative results on ablation study and comparisons on CelebAMask-HQ

Method	Semantic consistency ↑	Style (color/texture) ↑	mIoU ↑	mAcc ↑	allAcc ↑	Realism (%) ↑	Artifacts (%) ↓
SPADE	0.912	0.978/0.946	69.8	78.2	95.4	4	36
SEAN	0.917	0.993/0.982	71.5	79.6	96.0	19	28
CoCosNet	0.926	0.987/0.967	70.1	79.1	95.7	44	15
w/ $D_{\text{orig_context}}$	0.907	0.962/0.934	—	—	—	—	—
w/ $D_{\text{our_context}}$	0.911	0.964/0.940	—	—	—	—	—
w/ D_{color}	0.903	0.971/0.937	—	—	—	—	—
w/ $(D_{\text{orig_context}} + D_{\text{color}})$	0.910	0.974/0.942	—	—	—	—	—
w/ $(D_{\text{our_context}} + D_{\text{color}})$	0.912	0.976/0.947	67.2	76.8	94.0	33	21

not suitable for ablation studies).

Reconstruction. First, we encoded the original RGB input image into the latent code space and then decoded it back into the image space. The entire process was conditioned on the RGB image's corresponding semantic map. Figure 5 presents the reconstruction results. The first column shows the semantic maps, whereas the second column shows the RGB images. The remaining columns present the results of our model. The third column presents the reconstruction results with 64×64, and the last column shows the super-resolution editions of the third column (we name the two types of results LR and SR, respectively). LR and SR are also shown in Figs. 6, 7, 8, and 12). Diffusion models are not efficient at producing high-resolution images because they must take in data at various noise levels during training and synthesize images iteratively through hundreds of inference processes. Meanwhile, general neural network models only need to understand the original clean data (if noise addition is not included in the data augmentation part) and can produce images efficiently. Additionally, SR3 [35] was utilized here to generate high-resolution results, which have been widely adopted to compensate for the shortage of low resolution in DDPM [25, 36].

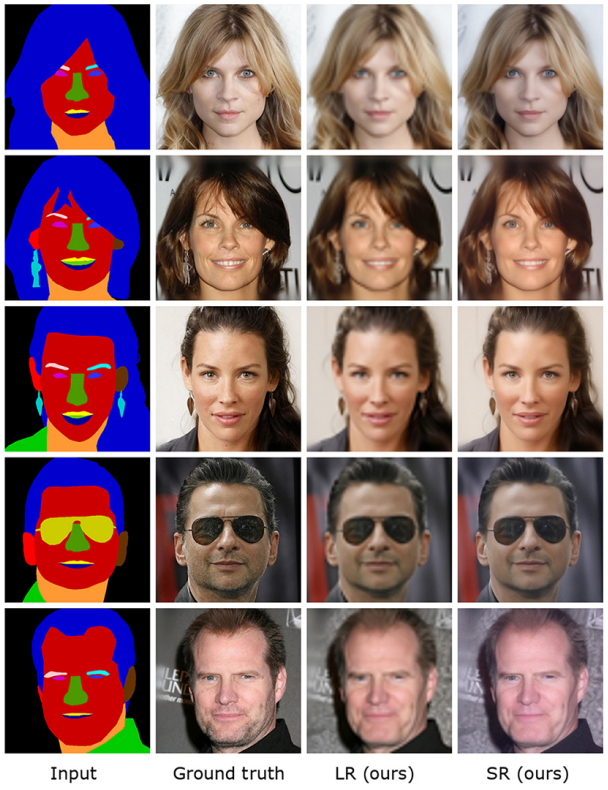


Fig. 5 Reconstruction results of the conditional DDPM.

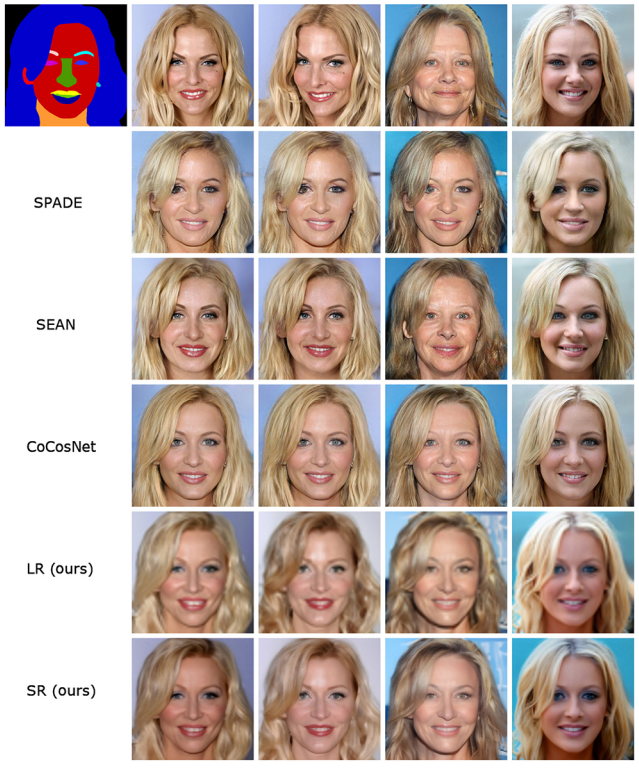


Fig. 6 Comparison with state-of-the-art methods. The semantic map and exemplars are in the first row.

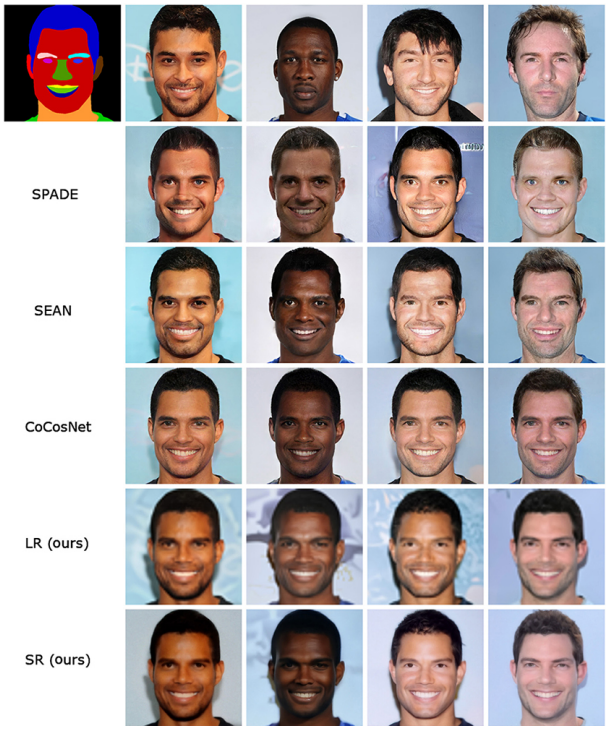


Fig. 7 Comparison with state-of-the-art methods. Four exemplars are provided for one semantic mask.



Fig. 8 Comparison with state-of-the-art methods on exemplars containing hands.

Baselines. We compared our method with several state-of-the-art image translation methods: (1) SPADE, a conditional generative model that can produce photorealistic images with a semantic layout. (2) SEAN, a network architecture that can individually control the style of each semantic region. (3) CoCosNet, a leading image translation method that learns cross-domain correspondences. (4) UNITE, a marginal contrastive learning network that learns domain-invariant features for realistic exemplar-based image translation.

Performance metrics. The synthesized images are expected to adhere to the semantic layout of the label map and resemble the appearance of an exemplar. To measure the segmentation accuracy, the mean intersection-over-union (mIoU), mean pixel accuracy (mAcc), and overall pixel accuracy (allAcc) were adopted in our experiments. For these metrics, the higher the value, the better the performance. We used the state-of-the-art segmentation networks BiSeNet [37] for CelebAMask-HQ, and PSANet [10] for ADE20K and Cityscapes. In addition, semantic consistency and style relevance described in Ref. [1] were employed.

Evaluation. For each semantic input from CelebAMask-HQ, we provided four reference images as exemplars for each semantic input from CelebAMask-HQ. As shown in Fig. 6, Fig. 7, and Fig. 8, SPADE and SEAN can capture styles with semantic consistency; however, their results are unnatural with obvious artifacts. CoCosNet is superior to SPADE and SEAN in generating photorealistic images, as shown in Fig. 6 and Fig. 7. However, in Fig. 8, CoCosNet's synthesized images are contaminated by hands from reference images. In this case, our method is much more similar to the exemplars. An extensive high-resolution correspondence is shown in the final row. In Table 1, all methods score more than 0.9 in terms of semantic consistency and style relevance. For mIoU, mAcc, and allAcc, SEAN performed the best, whereas all the methods behaved similarly without significant differences.

Figures 9 and 10 show the visual results for Cityscapes and ADE20K respectively. CoCosNet and UNITE aim to translate an exemplar's appearance to an input mask; however, they are unable to generate realistic styles for semantics that are not depicted in the exemplar, as shown in Fig. 10 the refrigerator (second row) and beach (third row) disappeared from the generated images. Although SPADE can produce reasonable objects that match semantics,

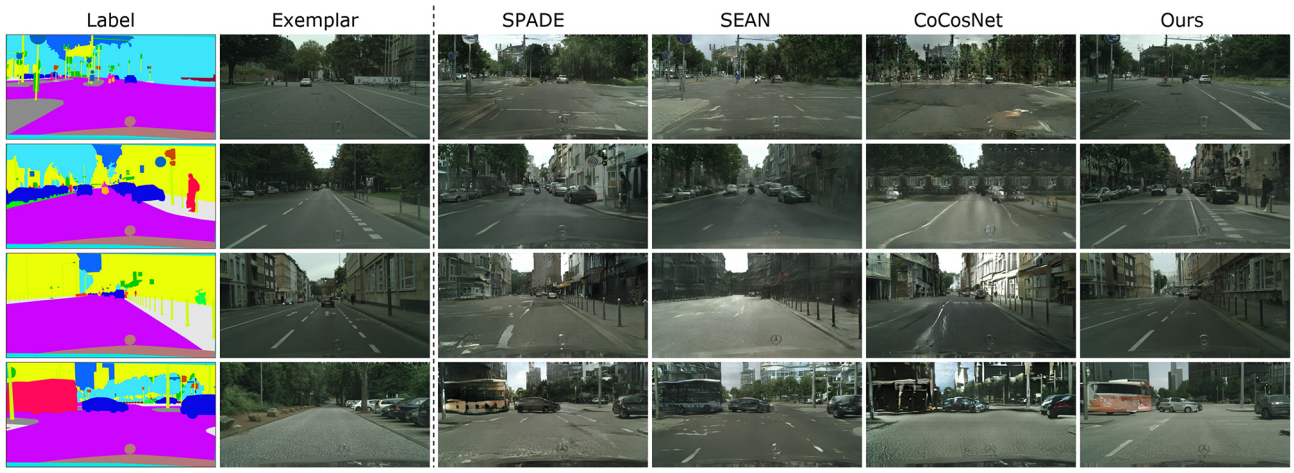


Fig. 9 Comparison with state-of-the-art methods on Cityscapes. SPADE, SEAN, and ours are able to generate realistic images on Cityscapes, except for CoCosNet.

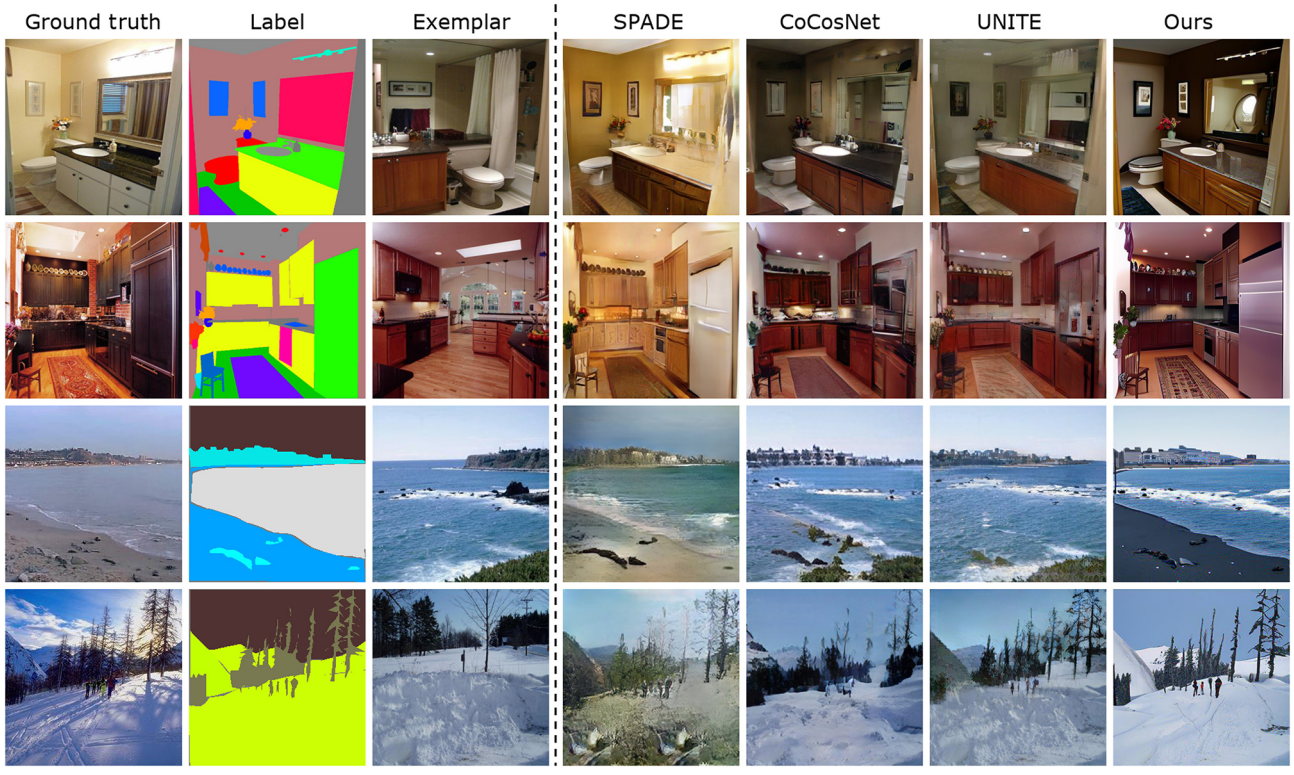


Fig. 10 Comparison with state-of-the-art methods on ADE20K. Corresponding RGB images of the semantic masks are also in the first column to help readers understand the specific semantics contained in this channeling dataset.

it fails to create consistent colors and textures from an exemplar. For Cityscapes and ADE20K, an obvious advantage of the proposed method is its ability to adhere to the semantic layout of the input mask. Specifically, in Fig. 9, clear demarcation lines can be observed between the carriageway and non-carriageway from the last column (ours) and the clear bus shape in the bottom-right synthesized image. In Fig. 10, the boundary between the sand and sea in the third row and the complicated tree boundaries in the last row validate this advantage, which is difficult for the compared methods to capture. The segmentation metrics in the last three columns of Tables 2 and 3 provide powerful evidence of shape maintenance capability. Our method has a low semantic consistency performance compared with existing methods but scores best in translating the color and texture on Cityscapes and maintaining clear

Table 2 Quantitative comparisons on Cityscapes

Method	Semantic consistency ↑	Style (color/texture) ↑	mIoU ↑	mAcc ↑	allAcc ↑
SPADE	0.917	0.972/0.931	59.1	66.4	93.1
SEAN	**0.930**	0.957/0.921	56.7	63.9	92.9
CoCosNet	0.891	0.974/0.942	29.5	34.9	82.2
Ours	0.899	**0.990/0.945**	**64.8**	**73.2**	**93.2**

Table 3 Quantitative comparisons on ADE20K

Method	Semantic consistency ↑	Style (color/texture) ↑	mIoU ↑	mAcc ↑	allAcc ↑
SPADE	0.861	0.873/0.890	46.0	53.4	83.9
CoCosNet	0.860	**0.978/0.956**	31.5	38.5	75.0
UNITE	**0.872**	0.968/0.954	34.2	41.6	75.6
Ours	0.832	0.955/0.890	**49.6**	**58.2**	**84.4**

boundaries of the segmentation mask. Overall, our method is comparable to state-of-the-art methods.

User study. Our user study adopted an image realism technique. For each algorithm, 20 synthesized images were randomly sampled, and each image was labeled with realism or artifacts by seven participants. We provided the SR images of our method to make it convenient for users. In Table 1 and Fig. 11, our method ranks the second, whereas CoCosNet generates more natural images through adversarial training. SPADE and SEAN are unsuitable for producing natural-looking images.

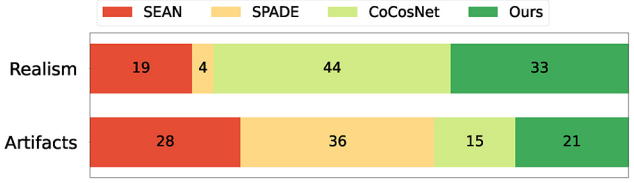


Fig. 11 User study of image realism and artifacts.

5 Conclusions and future work

Limitations. As shown in Fig. 12, our method can recover the beard in the first two columns, but fails in the last two. There are two possible explanations for this observation. Our method iteratively refines the output of the pretrained conditional diffusion model in the denoising process to synthesize images. However, the trained network may have a strong internal memory of some facial attributes. Another reason may be that our losses were not sufficiently powerful to accurately capture the textures of the given exemplar. This performance can be further improved by designing more powerful loss terms.



Fig. 12 Limitations of our method.

Future work. We preliminarily propose an exemplar-based image-translation method using the prevalent diffusion models. DDPM conditioned on semantic maps was first trained and subsequently guided via CCGD to generate the desired images. Our method is comparable to the state-of-the-art methods and can produce photorealistic images. In the future, we will explore the following two aspects. (1) More domains, including edge maps and pose keypoints, should be considered in the exemplar-based image translation. Most existing methods focus on synthesizing images from the semantic domain while neglecting other domains, which may also have potential research value. (2) The single-diffusion model is not applicable to high-resolution images. In contrast to concatenating several diffusion models, combining a diffusion model with GANs [26] can generate high-resolution and photorealistic images in shorter time. The latent diffusion model is trained on the latent space of pretrained autoencoders to reduce complexity and preserve detail, making high-resolution image synthesis possible [38]. As DDPM is now drawing increasing attention in image generation and editing tasks, there is an urgent need to improve the capability of DDPM to produce high-resolution images in shorter time and with lower complexity.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (U21B2023), DEGP Innovation Team (2022KCXTD025), Shenzhen Science and Technology Program (KQTD20210811090044003, RCJC20200714114435012), and Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ).

Declaration of competing interest

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

References

[1] Zhang, P.; Zhang, B.; Chen, D.; Yuan, L.; Wen, F. Cross-domain correspondence learning for exemplar-based image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5143–5153, 2020.

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

[3] Zhu, J. Y.; Park, T.; Isola, P.; Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, 2223–2232, 2017.

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

[5] Preechakul, K.; Chatthee, N.; Wizadwongsa, S.; Suwajanakorn, S. Diffusion autoencoders: Toward a meaningful and decodable representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10619–10629, 2022.

[6] Kim, G.; Kwon, T.; Ye, J. C. DiffusionCLIP: Text-guided diffusion models for robust image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2426–2435, 2022.

[7] Avrahami, O.; Lischinski, D.; Fried, O. Blended diffusion for text-driven editing of natural images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18208–18218, 2022.

[8] Liu, X.; Park, D. H.; Azadi, S.; Zhang, G.; Chopikyan, A.; Hu, Y.; Shi, H.; Rohrbach, A.; Darrell, T. More control for free image synthesis with semantic diffusion guidance. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 289–299, 2023.

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

[10] Zhu, P.; Abdal, R.; Qin, Y.; Wonka, P. SEAN: Image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5104–5113, 2020.

[11] Kim, T.; Cha, M.; Kim, M.; Lee, J. K.; Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, 1857–1865, 2017.

[12] Yi, Z.; Zhang, H.; Tan, P.; Gong, M. DualGAN: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision, 2849–2857, 2017.

[13] Liu, M.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, 700–708, 2017.

[14] Royer, A.; Bousmalis, K.; Gouws, S.; Bertsch, F.; Mosseri, I.; Cole, F.; Murphy, K. XGAN: Unsupervised image-to-image translation for many-to-many mappings. In: *Domain Adaptation for Visual Understanding*. Singh, R.; Vatsa, M.; Patel, V.; Ratha, N. Springer Cham, 33–49, 2020.

[15] Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, 1510–1519, 2017.

[16] Ma, L.; Jia, X.; Georgoulis, S.; Tuytelaars, T.; Gool, L. Exemplar guided unsupervised image-to-image translation. In: Proceedings of the International Conference on Learning Representations, 2019.

[17] Qi, X.; Chen, Q.; Jia, J.; Koltun, V. Semi-parametric image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8808–8816, 2018.

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

[19] Zhan, F.; Yu, Y.; Cui, K.; Zhang, G.; Lu, S.; Pan, J.; Zhang, C.; Ma, F.; Xie, X.; Miao, C. Unbalanced feature transport for exemplar-based image translation.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15028–15038, 2021.

[20] Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. In: Proceedings of the 34th Conference on Neural Information Processing Systems, 6840–6851, 2020.

[21] Dhariwal, P.; Nichol, A. Diffusion models beat GANs on image synthesis. In: Proceedings of the 35th Conference on Neural Information Processing Systems, 8780–8794, 2021.

[22] Ho, J.; Salimans, T. Classifier-free diffusion guidance. In: Proceedings of the Conference on Neural Information Processing Systems, 2021.

[23] Song, J.; Meng, C.; Ermon, S. Denoising diffusion implicit models. In: Proceedings of the International Conference on Learning Representations, 2021.

[24] Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Lecture Notes in Computer Science, Vol. 9351*. Navab, N.; Hornegger, J.; Wells, W. M.; Frangi, A. F. Eds, Springer Cham, 234–241, 2015.

[25] Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; Chen, M. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint* arXiv:2112.10741, 2021.

[26] Xiao, Z.; Kreis, K.; Vahdat, A. Tackling the generative learning trilemma with denoising diffusion GANs. In: Proceedings of the International Conference on Learning Representations, 2022.

[27] Wang, W.; Bao, J.; Zhou, W.; Chen, D.; Chen, D.; Yuan, L.; Li, H. Semantic image synthesis via diffusion models. *arXiv preprint* arXiv:2207.00050, 2022.

[28] Mechrez, R.; Talmi, I.; Zelnik-Manor, L. The contextual loss for image transformation with non-aligned data. In: Proceedings of the European Conference on Computer Vision, 768–783, 2018.

[29] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations, 2015.

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

[31] Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, 3730–3738, 2015.

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

[33] Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3213–3223, 2016.

[34] Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; Torralba, A. Scene parsing through ADE20K dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 633–641, 2017.

[35] Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; Norouzi, M. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 45, No. 4, 4713–4726, 2023.

[36] Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Lopes, R. G.; Ayan, B. K.; Salimans, T.; et al. Photorealistic text-to-image diffusion models with deep language understanding. In: Proceedings of the 36th Conference on Neural Information Processing Systems, 36479–36494, 2022.

[37] Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European Conference on Computer Vision, 325–341, 2018.

[38] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10684–10695, 2022.



Hao Ma obtained her Ph.D. degree in photogrammetry and remote sensing from Wuhan University in 2018. She is a postdoctoral researcher with the Visual Computing Research Center at the College of Computer Science and Software Engineering, Shenzhen University. Her main research interests are generative model based image synthesis, image and point cloud registration, and stereo vision.



Jingyuan Yang received her B.Eng. and Ph.D. degrees in electronic engineering, signal and information processing from Xidian University in 2017 and 2022, respectively. She is an assistant professor with the Visual Computing Research Center at the College of Computer Science and Software Engineering, Shenzhen University. Her current research interest is visual emotion analysis, image aesthetics, and generation tasks.



Hui Huang is a Distinguished TFA Professor of Shenzhen University, where she directs the Visual Computing Research Center and is also the Dean of the College of Computer Science and Software Engineering. Her research interests span on computer graphics, vision, and visualization. She is currently on the editorial board of *ACM Trans. on Graphics* and *IEEE Trans. on Visualization and Computer Graphics*.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made.

The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this licence, visit `http://creativecommons.org/licenses/by/4.0/`.

Other papers from this open access journal are available free of charge from http://www.springer.com/journal/41095. To submit a manuscript, please go to https://www.editorialmanager.com/cvmj.