General Image-to-Image Translation with One-Shot Image Guidance

Bin Cheng*, Zuhao Liu*, Yunbo Peng, Yue Lin
NetEase Games AI Lab
{chengbin04, liuzuhao, gzpengyunbo, gzlinyue}@corp.netease.com

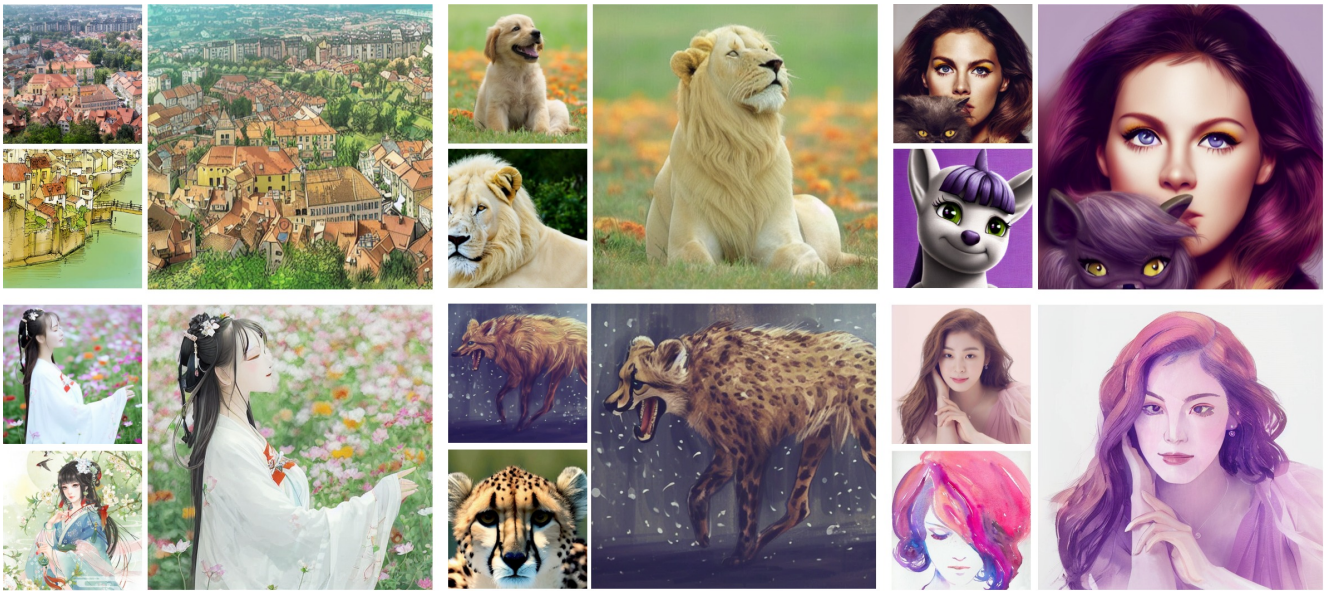

Figure 1: The image-to-image translation exhibition of the proposed visual concept translator (VCT). For each image group, the upper-left part is the source image, the lower-left part is the reference image, and the right part is the model output (target image).

Abstract

Large-scale text-to-image models pre-trained on massive text-image pairs show excellent performance in image synthesis recently. However, image can provide more intuitive visual concepts than plain text. People may ask: how can we integrate the desired visual concept into an existing image, such as our portrait? Current methods are inadequate in meeting this demand as they lack the ability to preserve content or translate visual concepts effectively. Inspired by this, we propose a novel framework named visual concept translator (VCT) with the ability to preserve content in the source image and translate the visual concepts guided by a single reference image. The proposed VCT contains a content-concept inversion (CCI) process to extract contents and concepts, and a content-concept fusion (CCF) process to gather the extracted information to obtain the target image. Given only one reference image, the proposed VCT can complete a wide range of general image-to-image translation tasks with excellent results. Extensive experiments are conducted to prove the superiority and effectiveness of the proposed methods. Codes are available at https://github.com/CrystalNeuro/visual-concept-translator.

1. Introduction

Image-to-image translation (I2I) task aims to learn a conditional generation function that translates images from source to target domain with source content preserved and target concept transferred[34, 46]. General I2I can complete a wide range of applications without dedicated model

*The first two authors contributed equally to this work.

design or training from scratch [45]. Traditionally, generative adversarial networks (GAN) or normalizing flow [11] are mainly applied to I2I tasks [19, 19, 34, 3]. However, these methods suffer from the problem of lacking adaptability [41]. The model trained in one source-target dataset cannot adapt to another one, so they fail to work in the scenario of general I2I.

Diffusion-based image synthesis has been developed rapidly in recent years due to the application of large-scale models [35, 37, 33]. Their strength is using a large number of image-text pairs for model training, so diverse images can be generated by sampling in the latent space guided by a specific text prompt. However, in our daily life, we accept massive visual signals containing abundant visual concepts. These visual concepts are difficult to describe in plain text just as the adage "A picture is worth a thousand words". In addition, I2I guided by reference images has wide applications including game production, artistic creation, and virtual reality. Therefore, research on image-guided I2I contains great potential in the computer vision community.

Several works try to extract visual information from images with the desired concepts. Specifically, [9] proposes a technique named textual inversion (TI) which freezes the model and learns a text embedding to represent the visual concepts. On the basis of TI, DreamBooth [36] and Imagic [20] are proposed to alleviate overfitting caused by model fine-tuning. The above methods are under the few-shot setting but sometimes collecting several related images containing the same concept is difficult. To address this problem, [7] proposes to use both positive and negative text embedding to fit the one-shot setting. However, these methods cannot be directly used in I2I tasks because they cannot preserve the content in the source image.

In order to preserve the source contents, the recently proposed DDIM inversion [6, 40] finds out the deterministic noise along the reverse direction of the diffusion backward process. Then, some studies[30, 12] further apply and improve the DDIM inversion to text-guided image editing. However, these methods are text-conditional so they fail to understand the visual concepts from reference images. Alternately, some works [49, 41] try to connect the source and target domain with image condition, but their models are task-specific so they cannot be used in general I2I.

In this paper, to complete the general I2I tasks guided by reference images, we propose a novel framework named visual concept translator (VCT) with the ability to preserve content in the source image and translate the visual concepts with a single reference image. The proposed VCT solves the image-guided I2I by two processes named content-concept inversion (CCI) and content-concept fusion (CCF). The CCI process extracts contents and concepts from source and reference images through pivot turning inversion and multi-concept inversion, The CCF process employs a dual-stream denoising architecture to gather the extracted information to obtain the target image. Given only one reference image, the proposed VCT can complete a wide range of general image-to-image translation tasks with excellent results. Extensive experiments including massive tasks of general I2I and style transfer are conducted for model evaluation.

In summary, our contributions are as follows

(1) We propose a novel framework named visual concept translator (VCT). Given only a single reference image, VCT can complete the general I2I tasks with the ability to preserve content in the source image and translate the visual concepts.

(2) We propose a content-concept inversion (CCI) to extract contents and concepts with pivot turning inversion and multi-concept inversion. We also propose a content-concept fusion (CCF) process to gather the extracted information with a dual-stream denoising architecture.

(3) Extensive experiments including massive tasks of general I2I and style transfer are conducted for model evaluation. The generation results show the high superiority and effectiveness of the proposed methods.

2. Related Works

2.1. Image-to-image Translation

The I2I aims to translate an image from the source domain to the target domain. The current I2I paradigms are mostly based on GANs [1, 29, 8, 53, 58, 50, 59]. However, these methods suffer from the problem of lacking adaptability [41]. The model trained in one source-target dataset cannot adapt to another one. In addition, large training images are always required for these methods.

The TuiGAN proposed by Lin et al. [27] can achieve translation with only one image pair, but their method necessitates retraining the whole network for each input pair, which is very time-consuming.

One specific type of I2I named image style transfer tries to transform the image style from source to target. The seminal work of Gatys et al. [10] shows that artistic images can be generated by separating content and style with the deep neural network. Then, to realize real-time style transfer, Johnson et al. [18] train a feed-forward network to handle the optimization problem mentioned by Gatys et al. Many works [47, 42, 43, 24, 17, 23] are categorized into per-style-per-model where the trained model can only fit one specific style. In order to increase the model flexibility, arbitrary style transfer is realized by many studies [15, 31, 16, 4, 28, 39, 48] where only single forward pass is needed for any input style image. However, these methods fail to generalize to general I2I tasks such as face swap because they lack the ability to process fine-grained information.

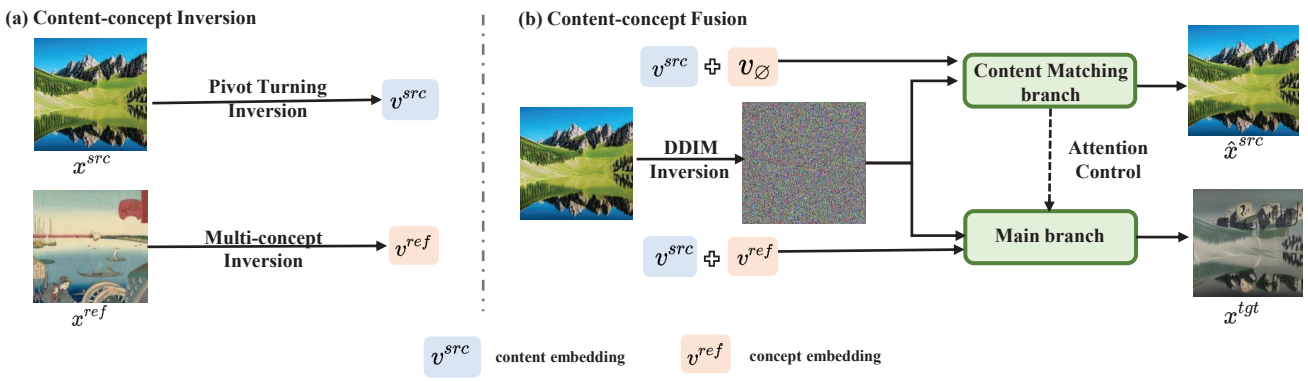

Figure 2: **The overall visual concept translator(VCT) framework**. Given a source image x^{src} and a reference image x^{ref}: (A) Content-concept inversion (CCI) process, we apply Pivot Turning Inversion with x^{src} to obtain the source text embedding v^{src}. Meanwhile, we apply Multi-concept Inversion with x^{ref} to learn the reference text embedding v^{ref}. (B) Content-concept fusion(CCF) process, we employ a dual-stream denoising architecture for image translation work, including a main branch $\mathcal{B}$ and a content matching branch $\mathcal{B}^*$. They start with the same initial noise inverted by x^{src} using DDIM inversion. The content matching branch reconstructs the source image and extracts the attention maps for the attention control mechanism. Finally, the main branch gathers all the information to obtain a target image x^{tgt}.

2.2. Diffusion-based Image Synthesis

Large-scale diffusion models conditioned on the plain text have shown good performance in high-resolution image syntheses recently, such as Stable Diffusion [35], Imagen [37] and DALL-E 2 [33]. The large text-image models [5, 32] are used by these methods to achieve text-guided synthesis. However, the text used to generate the target images is sometimes unavailable, so the inversion technique is used by some works [9, 36, 20] to learn a text embedding to guide the pre-trained large-scale diffusion models. To achieve translation of images from the source to the target domain, DDIM inversion [6, 40] finds out the deterministic noise vector with text condition along the reverse direction of the backward process, but this method is guided by text only. Our proposed method tries to handle the above drawbacks and fuses the abundant visual concepts from the image to complete the general I2I tasks.

3. Methods

3.1. Preliminaries

Latent Diffusion Models. Diffusion models are probabilistic generative models in which an image x_0 is generated by progressively removing noise from an initial Gaussian noise image $x_T \sim \mathcal{N}(0, \mathbf{I})$ in the sequence of $x_T, x_{T-1}, ..., x_1, x_0$.

With the remarkable capacity of image generation, the Latent Diffusion Model (LDM) [35] is utilized as our model backbone. Different from the conventional diffusion models that perform denoising operations directly in the image space, LDM conducts the process in the latent space with an autoencoder.

Specifically, an input image x is encoded into the latent space by the autoencoder $z = \mathcal{E}(x), \hat{x} = \mathcal{D}(z)$ (with an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$) pre-trained with a large number of images. Then, the denoising process is achieved by training a neural network $\epsilon_\theta(z_t, t, v)$ that predicts the added noise, following the objective:

$$\min_\theta E_{z_0, \epsilon \sim \mathcal{N}(0,I), t \sim \mathrm{U}(1,T)} \left\| \epsilon - \varepsilon_\theta(z_t, t, v) \right\|_2^2. \quad (1)$$

Note that v is the text embedding generated from the text condition and z_t is the noisy latent in timestamp t. z_t is generated by adding noise to the sampled data z_0 as

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t}\epsilon, \quad (2)$$

with $0 = \alpha_t < \alpha_{t-1} < ... < \alpha_0 = 1$, which are hyperparameters of the diffusion schedule, and $\epsilon \sim \mathcal{N}(0, I)$.

The text embedding v is obtained by $v = \tau(y)$ where τ is a BERT [5] tokenizer and y is a text prompt. The tokenizer τ converts each word or sub-word in an input string to a token, which is an index in a specific pre-defined dictionary. Each token is then linked to a unique embedding vector that can be retrieved through an index-based lookup.

Texture inversion. Textual Inversion (TI) [9] introduces a new concept in a pre-trained text conditional generative model by learning an embedding e^* as pseudo-words S^*. With a small collection of images X, TI do so by solving the following optimization problem:

$$\min_e E_{x \sim \mathcal{U}_X} E_{z_t \sim q(z_t|x)} \left\| \epsilon - \hat{\varepsilon}_\theta(z_t, t, \tau(y, S^*)) \right\|_2^2. \quad (3)$$

As such, it motivates the learned embedding e^* to capture fine visual details unique to the concept at a coarse level.

DDIM inversion. Inversion entails finding a noise map z_t that reconstructs the input latent code z_0 upon sampling. A simple inversion technique was suggested for the DDIM sampling [6, 40], based on the assumption that the ODE process can be reversed in the limit of small steps:

$$z_{t+1} = \sqrt{\bar{\alpha}_{t+1}} f_\theta\left(z_t, t, v\right) + \sqrt{1-\bar{\alpha}_{t+1}} \varepsilon_\theta\left(z_t, t, v\right). \quad (4)$$

where z_t is noised latent code at timestep t, $\bar{\alpha}_{t+1}$ is noise scaling factor as defined in DDIM[6], and $f_\theta\left(z_t, t, v\right)$ predicts the final denoised latent code z_0.

$$f_\theta\left(x_t, t, c\right) = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta\left(x_t, t, c\right)}{\sqrt{\bar{\alpha}_t}} \quad (5)$$

In other words, the diffusion process is performed in the reverse direction, that is $z_0 \rightarrow z_T$ instead of $z_T \rightarrow z_0$, where z_0 is set to be the encoding of the given real image.

Classifier-free guidance. The diffusion model may ignore the conditional input and produce results uncorrelated with this input. One way to address this is the classifier-free guidance[14]. During the denoising step, with a guidance scale $w \geq 1$, the classifier-free guidance prediction is defined by:

$$\tilde{\varepsilon}_\theta\left(z_t, t, v\right) = w \cdot \varepsilon_\theta\left(z_t, t, v\right) + (1-w) \cdot \varepsilon_\theta\left(z_t, t, v_\varnothing\right). \quad (6)$$

where $v_\varnothing$ represents the embedding of a null text.

3.2. Overall Framework

Given a source image x^{src} and a reference image x^{ref}, the goal of VCT is to generate a new image x^{tgt} that complies with x^{ref} while preserving the structure and semantic layout of x^{src}.

Fig. 2 shows the overall framework of the proposed VCT including a content-concept inversion (CCI) process and a content-concept fusion (CCF) process. As shown in Fig. 2 (a), the CCI process extracts contents and concepts from source image x^{src} and reference image x^{ref} into learnable embeddings. Then in Fig. 2 (b), the CCF process employs a dual-stream denoising architecture including a main branch $\mathcal{B}$ and a content matching branch $\mathcal{B}^*$, and both branches starts from the same initial noise inverted by x^{src}. The content matching branch reconstructs the source image and extracts the attention maps to guide the main process by the attention control mechanism. Then, the main branch gathers all information to obtain a target image x^{tgt}. For better understanding, we first explain the CCF process in Section 3.3, then we describe the CCI process in Section 3.4.

3.3. Content-concept Fusion

ϵ Space Fusion. Given two different text embedding v^{src} and v^{ref}, they can be guided separately and yield two different noise prediction ϵ^{src} and ϵ^{ref}:

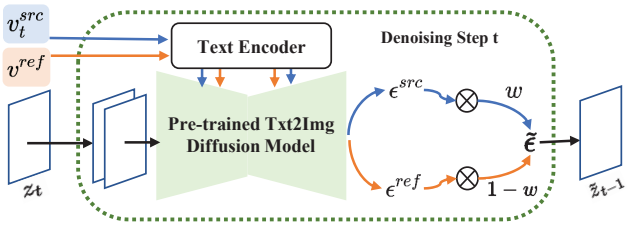

Figure 3: ϵ **fusion**. In each denoising step t, the text embeddings v_t^{src} and v^{ref} are extrapolated with guidance scale w in ϵ space.

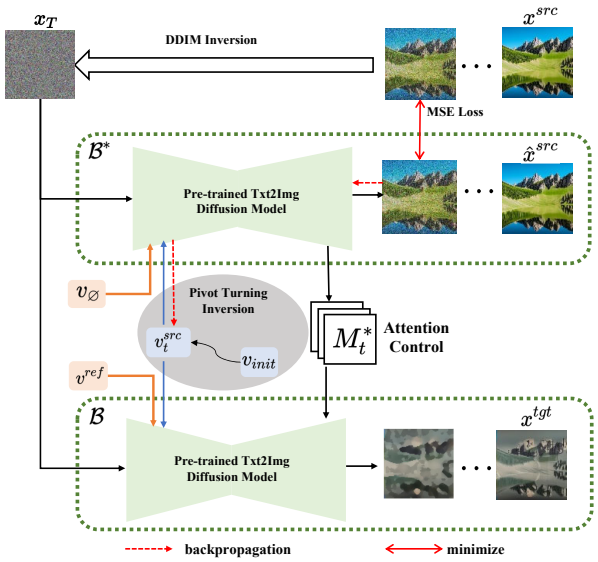

Figure 4: **Dual-stream denosing architectrure.**

$$\epsilon^{src} = \varepsilon_\theta\left(z_t, t, v^{src}\right), \epsilon^{ref} = \varepsilon_\theta\left(z_t, t, v^{ref}\right). \quad (7)$$

We call this space ϵ *space*, as shown in Fig. 3.

According to the conclusion stated by classifier guidance[6] and classifier-free guidance[14], the noise prediction in ϵ space in each diffusion step can be interpreted as score estimation function $\varepsilon_\theta\left(z_t, t, v\right) \approx -\sigma_t \nabla_{z_t} \log p\left(z_t \mid v\right)$, where $\nabla_{z_t} \log p\left(z_t \mid v\right)$ represents the gradient of log-likelihood of an implicit classifier $p\left(v \mid z_t\right) \propto p\left(z_t \mid v\right) / p\left(z_t\right)$.

Under the score estimation function view of ϵ space, the independent feature v^{src} and v^{ref} can be fused in the ϵ space to generate images containing certain attributes from both the source image and the reference image:

$$\tilde{\varepsilon}_\theta\left(z_t, t, v^{src}, v^{ref}\right) = w \cdot \epsilon^{src} + (1-w) \cdot \epsilon^{ref}. \quad (8)$$

w is the hyperparameter that balances the two terms. It's

noted that the classifier-free guidance is a special case of Eq. 8.

Dual stream denoising architecture. Based on the ϵ fusion mechanism, we now turn to the image translation task. As shown in Fig. 4, let x^T be the initial noise, obtained by inverting x^{src} using DDIM inversion with Eq. 4, where we set $v = v_\varnothing$. Starting from the same initial noise x^T, we employ a dual-stream denoising architecture for I2I, denoted as a main branch $\mathcal{B}$ and a content matching branch $\mathcal{B}^*$.

The content matching branch $\mathcal{B}^*$ is a denoising process that perfectly reconstructs the source image x^{src} (with z^{src} perfectly reconstructed in latent space for LDMs), and the main branch $\mathcal{B}$ is the denoising process that finally serves for the I2I tasks.

$$\begin{aligned} \mathcal{B}^* &: z_T \rightarrow z^*_{T-1} \rightarrow ... \rightarrow z^*_1 \rightarrow z^{src} \\ \mathcal{B} &: z_T \rightarrow z_{T-1} \rightarrow ... \rightarrow z_1 \rightarrow z^{tgt}. \end{aligned} \tag{9}$$

At each denoising step t, the content matching branch $\mathcal{B}^*$ aims to extract the text embedding v^{src}_t and the attention map M^*_t, which would serve for the parallel denoising step in the main branch. With $\mathcal{B}^*$, we obtain meaningful embedding and generated structure of the source image.

To better inject the information of the source image x^{src}, the dual stream diffusion processes have almost the same computation pipelines, except for the reference embeddings used in ϵ space fusion. We perform ϵ space fusion in the content matching branch as the main branch by:

$$\tilde{\varepsilon}_\theta\left(z_t, t, v^{src}, v_\varnothing\right) = w \cdot \epsilon^{src} + (1-w) \cdot \epsilon^{\varnothing}. \tag{10}$$

The above sampling procedure reduces to the classifier-free guidance. And we should ensure that Eq. 8 and Eq. 10 have the same w for the dual-stream diffusion architecture.

Attention control. Recent large-scale diffusion models [35, 37, 33] incorporate conditioning by augmenting the denoising network ε_θ with self-attention layer and cross-attention layer [2, 44]. Of particular interest are the *cross-attention map* and *self-attention map*, denoted as M in total, which is observed to have a tight relation with the structure of the image [12]. To this end, Amir at al.[12] pursue *prompt-to-prompt* editing framework for text-guided image translation task, which controls the attention maps of the edited image by injecting the attention maps of the original image along the diffusion process.

In our case, we employ soft attention control as described in *prompt-to-prompt* [12]. Let M^*_t be the attention map of a single step t of the content matching branch, and M_t be the attention map of the main branch. The soft attention control is defined as:

$$\widehat{M} = AC\left(M_t, M^*_t, t\right) = \begin{cases} M^*_t & \text{if } t < \tau \\ M_t & \text{otherwise} \end{cases} \tag{11}$$

where τ is a timestamp parameter that determines until which step the attention map replacement is applied. We define $\tilde{\varepsilon}_\theta\left(z_t, t, v^{src}, v^{ref}\right)\{M \leftarrow \widehat{M}\}$ to be the function that overrides the attention map M in $\tilde{\varepsilon}$ with additional given map $\widehat{M}$.

3.4. Content-concept Inversion

Pivotal turning inversion is proposed to generate the content embedding to guide the CCF process. We start by studying the DDIM inversion[6, 40]. In practice , a slight error is incorporated in every step. For unconditional diffusion models, the accumulated error is negligible and the DDIM inversion succeeds. However, recall that meaningful editing using the Stable Diffusion model [35] requires applying classifier-free guidance with a guidance scale w. Ron et al.[30] have presented that such a guidance scale amplifies the accumulated error.

To this end, Ron et al. [30] introduce null-text inversion technology to reconstruct the image and further for text-guided image translation tasks. Null-text inversion modifies the unconditional embedding in each timestamp t that is used for classifier-free guidance to match the initial conditional DDIM inversion trajectory.

In our image-guided case, we do not know the exact text prompt of the source image x^{src}. So, inspired by [30], we implement unconditional DDIM inversion, and optimize the source embedding v^{src}_t in each timestamp t for accurately matching the source image x^{src}, instead of the DDIM inversion trajectory.

In each timestamp t, we optimize the v^{src}_t by:

$$\min_{v^{src}_t} \left\| z_0 - \hat{z}_0\left(z_t, v^{src}_t\right) \right\|^2_2 \tag{12}$$

where $\hat{z}_0\left(z_t, v^{src}_t\right)$ refers to the estimated clean latent $\hat{z}_0$ given z_t and v^{src}_t, using the Tweedie's formula [21]. We rewrite it as:

$$\hat{z}_0\left(z_t, v^{src}_t\right) = \frac{z_t}{\sqrt{\bar{\alpha}_t}} - \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} \tilde{\varepsilon}_\theta\left(z_t, t, v^{src}, v_\varnothing\right) \tag{13}$$

where $\tilde{\varepsilon}_\theta\left(z_t, t, v^{src}, v_\varnothing\right)$ is defined in Eq. 10.

Note that for every $t < T$, the optimization should start from the endpoint of the previous step $t+1$ optimization, which computes a constant z_t using the optimized v^{src}_{t+1} and z_{t+1}. Otherwise, the learned embedding would not hold at inference.

Multi-concept inversion is proposed to represent complex visual concepts by generating the concept embedding. Lastly, we should learn a reference embedding v^{ref} from the reference image x^{ref}. The methodological approach is related to Textual Inversion [9] and DreamArtist [7].

To represent the concepts in the input images, Textual Inversion [9] learns an embedding as pseudo-words S_* from

few-shot images. DreamArtist [7] improves Textual Inversion, which learns a paired positive and negative multi-concept embeddings(S_*^p and S_*^n) from one-shot image and proposes reconstruction constraint for detail enhancement. In our case, we apply a similar strategy as DreamArtist, yet our method offers two improvements:

Firstly, we find that the multi-concept embeddings are useful for mining semantics information from the images, while the negative embeddings are optional. And in our pipeline, the negative embeddings are in conflict with the source embedding x^{src}. Thus, we use single positive multi-concept embeddings for learning the reference text embedding v^{ref}. We freeze the parameters of the generative diffusion model ε_θ, and optimize the v^{ref} using the denoising diffusion objective[13]:

$$\mathcal{L}_{ldm} = E_{\epsilon,t}\left[\left\|\epsilon - \varepsilon_\theta\left(z_t^{ref}, t, v^{ref}\right)\right\|_2^2\right]. \tag{14}$$

where v^{ref} is the multi-concept embeddings, z_t^{ref} is a noisy version of z_0^{ref} (the latent code of the reference image x^{ref}) obtained using Eq. 2, $\epsilon \sim \mathcal{N}(0, I)$ and $t \sim \mathrm{U}(1, T)$.

Secondly, we improve the reconstruction constraint for the mechanism of detail enhancement in DreamArtist. DreamArtist applys reconstruction constraint in x space, which can be denoted as $\mathcal{D}\left(\hat{z}_{t-1}\left(z_t, S^*\right)\right) \leftrightarrow x_0$. For one thing, optimization in x space suffers from huge resource consumption due to the gradient backpropagation inside decoder $\mathcal{D}$. For another thing, there is a gap between the estimated z_{t-1} and z_0, especially in the early stage of the denoising process.

Formally, we implement reconstruction constraint in z space. The reconstruction loss can be written as:

$$\mathcal{L}_{rec} = E_{\epsilon,t}\left[\left\|z_0^{ref} - \hat{z}_0\left(z_t^{ref}, v_t^{ref}\right)\right\|_2^2\right]. \tag{15}$$

where $\hat{z}_0\left(z_t^{ref}, v_t^{ref}\right)$ refers to the estimated clean latent $\hat{z}_0^{ref}$ given z_t^{ref} and v_t^{ref}, using Eq. 13.

4. Experiments

4.1. Implementation details

Putting all components together, our full algorithm is presented in our supplementary material. The core training process consists of two parts: pivotal tuning inversion with x^{src} and multi-concept inversion with x^{ref}, which can be implemented independently. For more details please refer to our supplementary material.

Our experiments were conducted using a single A100 GPU. We use Adam[22] optimizer for both training processes. We collect the evaluation images from the large-scale LAION 5B dataset [38] containing 5 billion images.

4.2. Comparison to Prior/Concurrent Work

General I2I Tasks. Here, we evaluate the performance of the proposed framework in general I2I tasks including leopard→dog, face swap, and mountain→snow mountain, as shown in Fig. 5. We compare the proposed method with TuiGAN [27], PhotoWCT [26], stable diffusion (SD) [35], textual inversion (TI) [9] and prompt-to-prompt (Ptp) [12]. For text-to-image models without learned embedding input including SD and Ptp, we use BLIP image caption model [25] to extract text description as input of diffusion model. From Fig. 5, the GAN-based translation methods TuiGAN and PhotoWCT cannot well translate the concept with only one image input with poor generation quality. For example, from columns 3-4 of Fig. 5, GAN-based methods only translate part of texture features from the reference image in leopard→dog and face swap task, and the image quality is poor in the mountain→snow mountain task. Therefore, the GAN-based methods cannot achieve satisfactory results in the one-shot setting. For diffusion-based methods SD and TI, the concepts of the reference image can be well preserved, but the information in the content image cannot be extracted. As shown in column 7 of Fig. 5, Ptp can well preserve content but the concepts in the reference images cannot be fused. By tackling all weaknesses of the above methods, the proposed VCT can generate the best results with concepts learned and content preserved.

Furthermore, to evaluate the strong concept translation ability of the proposed VCT, we keep the content image fixed and change different reference images in Fig. 6. The generation results of different reference images show satisfactory content preservation and concept translation ability. More results can be found in the supplementary material.

As shown in Fig. 7, we further make comparisons to concurrent one-shot baselines: Paint-by-example[49] and ControlNet[52]. These methods use additional conditions for controlling the generated image, while our method obtains better performance.

Image Style Transfer. In addition to general I2I, the proposed method also achieves excellent results in tasks of image style transfer. We compare our method with recent SOTAs in style transfer tasks with different art styles. As shown in Fig. 8, we totally compare with three GAN-based methods including TuiGAN [27], PhotoWCT [26] and ArtFlow [54], and three diffusion-based methods including SD [35], TI [9] and Ptp [12]. Following the setting of general I2I, we use the BLIP image caption model to extract text descriptions for text-to-image model SD and Ptp. From the results in Fig. 8, large defects exist for results generated by GAN-based methods, especially for TuiGAN and ArtFlow as columns 3 and 5 in Fig. 8. The same content preservation problem exists in diffusion-based methods SD and TI as general I2I. For Ptp, although the contents are preserved, the concept in the reference images cannot be well trans-

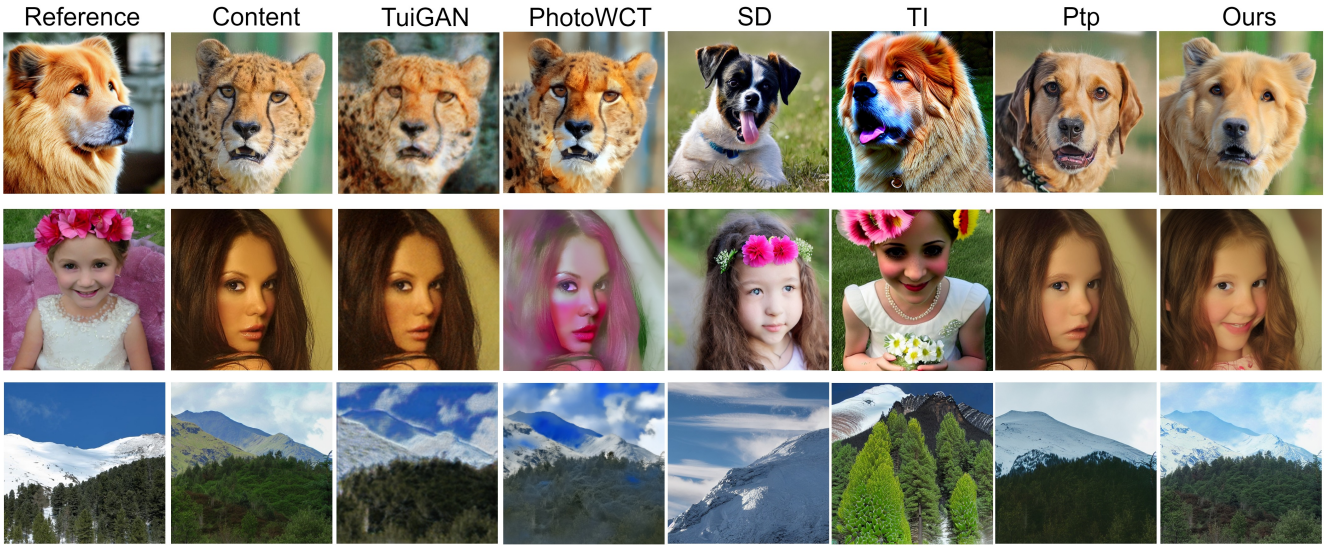

Figure 5: Model Performance in general image-to-image translation task including leopard→dog, face swap and mountain→snow mountain. The first and second columns show the reference and content images, respectively. The 3-7 columns show the translation results of different methods.

Figure 6: Our method can perform fine-grained image-to-image translation.

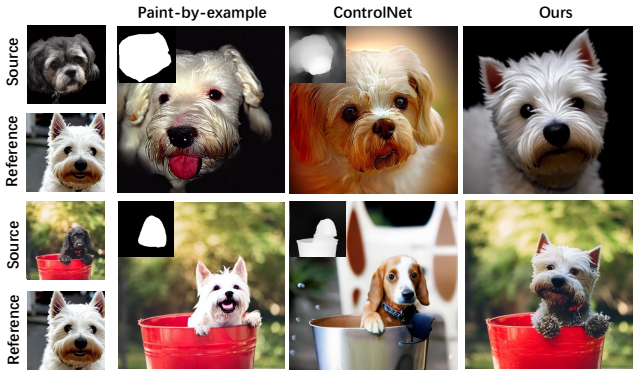

Figure 7: Comparisons to concurrent one-shot baselines: Paint-by-example[49] and ControlNet[52].

lated. The proposed method can also generate the most satisfactory images, as shown in column 9 of Fig. 8.

We also evaluate the model performance by keeping the reference image fixed and changing the content image, and vice versa. The results are shown in Fig. 9. The excellent translation results prove the generalization of the proposed method.

Quantitative Comparision. Due to the absence of ground truth for style transfer and the domain gap between the two domains, quantitative evaluation remains an open challenge. Following the same setting of StyTR2, we randomly choose 800 generated images from different translation tasks for quantitative comparison. We compare the proposed method with the state-of-the-art including Artflow [54], CAST [57], InST [51], StyleFormer [56] and StyTR2 [55], and the results are shown in Table 1. We use Learned Perceptual Image Patch Similarity (LPIPS) to evaluate the difference between output and source image, and

Table 1: Quantitative evaluation results.

	LPIPS↓	CLIPscore↑
Artflow[54]	0.42	0.51
CAST[57]	0.55	0.61
InST[51]	0.43	0.48
StyleFormer[56]	0.46	0.48
StyTR2 [55]	0.43	0.53
Ours	**0.35**	**0.66**

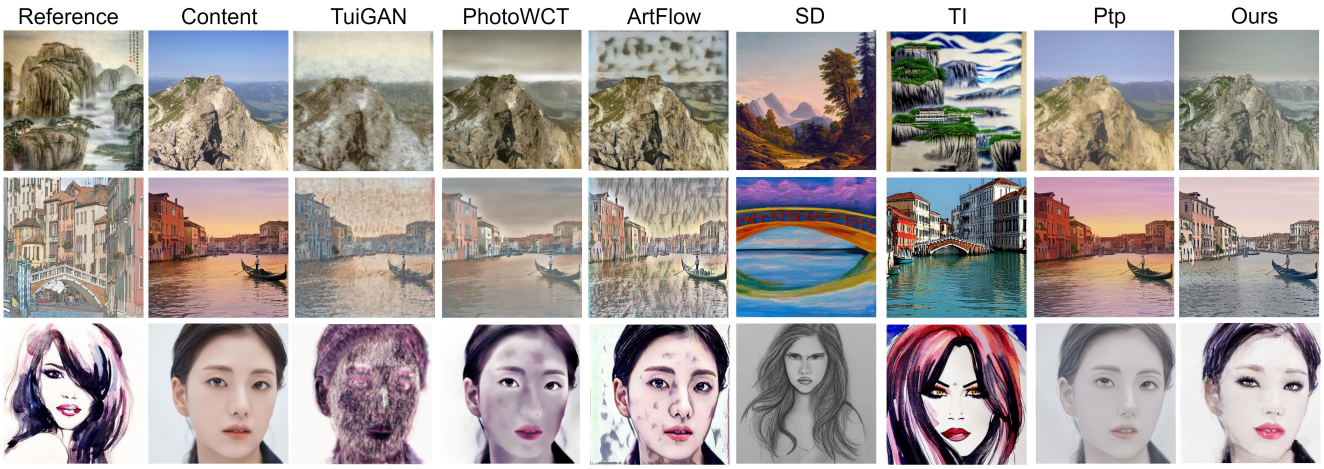

Figure 8: Model Performance in the style translation task. The first and second columns show the reference and content images, respectively. The 3-7 columns show the translation results of different methods.

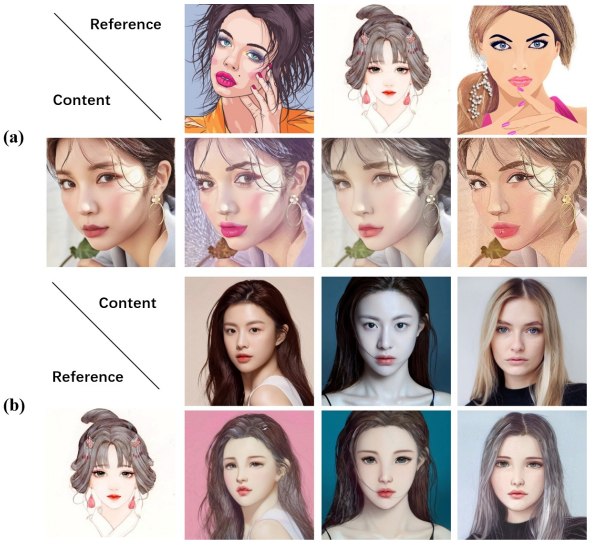

Figure 9: Content variance and style variance of the style transfer result.

CLIP score to evaluate the difference between output and reference image. The results show our proposed method can achieve the best performance with the lowest LPIPS and highest CLIPscore.

4.3. Ablation Study

Finally, we ablate each component of our method and show its effectiveness, including multi-concept inversion (MCI), pivotal turning inversion (PTI), and attention control (AC).

See visual ablation studies in Fig. 10: (a) By removing

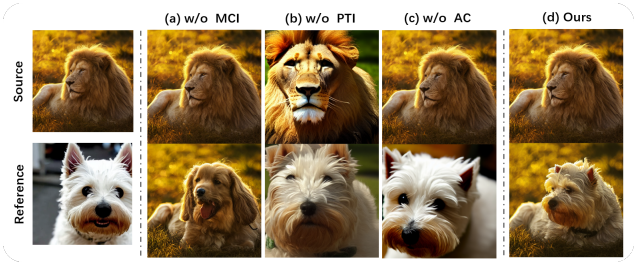

Figure 10: Visualization of the ablation study.

MCI, where we use the word 'dog' to generate the reference embedding v^{ref} in our pipeline, the generated result is not the specific dog in the reference image. (b) Without using PTI, the content matching branch cannot reconstruct the content image, due to the inconsistent DDIM sampling trajectory. (c) By removing AC, the result can not retain the structure of the content image.

Overall, we can obtain the best generation outputs by using all of our proposed components, which better preserves the structure and semantic layout of the content image, while complying with the reference image. Further ablations can be found in the supplementary material.

5. Conclusion

In this work, motivated by the importance of visual concepts in our daily life, we complete the general I2I with image guidance by proposing a novel framework named VCT. It can preserve the content in the source image and translate visual concepts guided by a single reference image. We evaluate the proposed model on a wide range of general image-to-image translation tasks with excellent results.

References

[1] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14154–14163, 2021. 2

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 5

[3] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in style: Uncovering the local semantics of gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5771–5780, 2020. 2

[4] Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2719–2727, 2020. 2

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2, 3, 4, 5

[7] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022. 2, 5, 6

[8] Aviv Gabbay and Yedid Hoshen. Scaling-up disentanglement for image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6783–6792, 2021. 2

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 3, 5, 6

[10] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 2

[11] Aditya Grover, Christopher Chute, Rui Shu, Zhangjie Cao, and Stefano Ermon. Alignflow: Cycle consistent learning from multiple domains via normalizing flows. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4028–4035, 2020. 2

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 5, 6

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 6

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*. 4

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2

[16] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. Dynamic instance normalization for arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4369–4376, 2020. 2

[17] Yongcheng Jing, Yang Liu, Yezhou Yang, Zunlei Feng, Yizhou Yu, Dacheng Tao, and Mingli Song. Stroke controllable fast style transfer with adaptive receptive fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 238–254, 2018. 2

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 2

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[20] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 2, 3

[21] Kwanyoung Kim and Jong Chul Ye. Noise2score: tweedie's approach to self-supervised image denoising without clean images. *Advances in Neural Information Processing Systems*, 34:864–874, 2021. 5

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 6

[23] Dmytro Kotovenko, Artsiom Sanakoyeu, Sabine Lang, and Bjorn Ommer. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4422–4431, 2019. 2

[24] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 702–716. Springer, 2016. 2

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 6

[26] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 453–468, 2018. 6

[27] Jianxin Lin, Yingxue Pang, Yingce Xia, Zhibo Chen, and Jiebo Luo. Tuigan: Learning versatile image-to-image translation with two unpaired images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 18–35. Springer, 2020. 2, 6

[28] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6649–6658, 2021. 2

[29] Yahui Liu, Enver Sangineto, Yajing Chen, Linchao Bao, Haoxian Zhang, Nicu Sebe, Bruno Lepri, Wei Wang, and Marco De Nadai. Smoothing the disentangled latent style space for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10785–10794, 2021. 2

[30] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 2, 5

[31] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5880–5888, 2019. 2

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3, 5

[34] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 1, 2

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 5, 6

[36] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 2, 3

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 3, 5

[38] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 6

[39] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8242–8250, 2018. 2

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3, 4, 5

[41] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *International Conference on Learning Representations*, 2022. 2

[42] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. *arXiv preprint arXiv:1603.03417*, 2016. 2

[43] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6924–6932, 2017. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[45] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. 2

[46] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 1

[47] Hao Wu, Zhengxing Sun, and Weihang Yuan. Direction-aware neural style transfer. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1163–1171, 2018. 2

[48] Zhijie Wu, Chunjin Song, Yang Zhou, Minglun Gong, and Hui Huang. Efanet: Exchangeable feature alignment network for arbitrary style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12305–12312, 2020. 2

[49] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2211.13227*, 2022. 2, 6, 7

[50] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Unsupervised image-to-image translation with generative prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18332–18341, 2022. 2

[51] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10156, 2023. 7

[52] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 6, 7

[53] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 5143–5153, 2020. 2

[54] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 862–871, 2021. 6, 7

[55] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336, 2022. 7

[56] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14618–14627, 2021. 7

[57] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–8, 2022. 7

[58] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. The spatially-correlative loss for various image translation tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16407–16417, 2021. 2

[59] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2