Zero-Shot Contrastive Loss for Text-Guided Diffusion Image Style Transfer

Serin Yang Hyunmin Hwang Jong Chul Ye
{yangsr, hyunmin_hwang, jong.ye}@kaist.ac.kr
Kim Jaechul Graduate School of AI
Korea Advanced Institute of Science and Technology (KAIST)

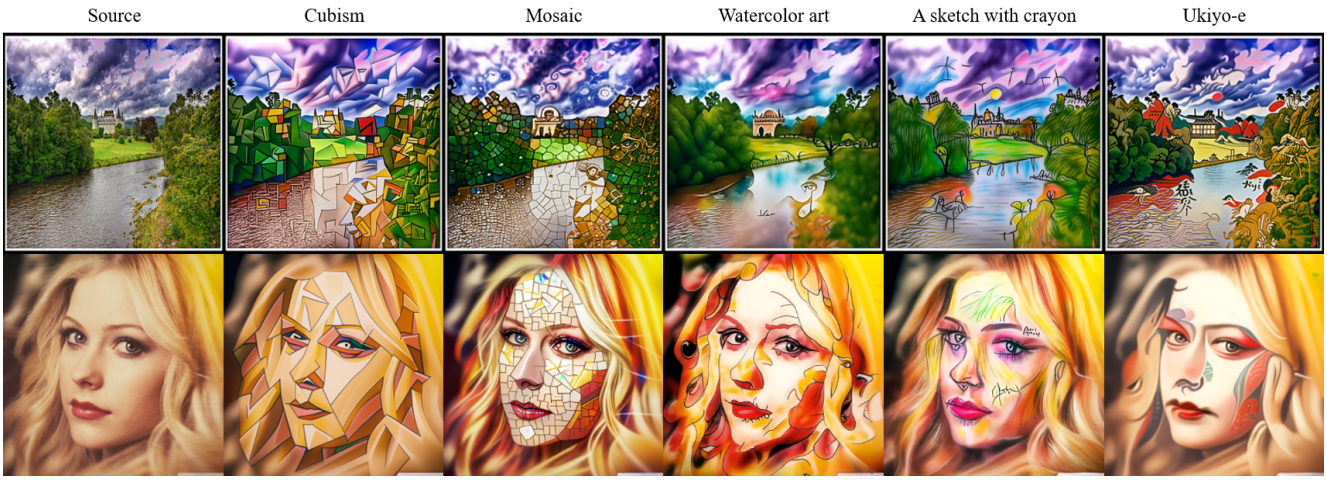

Figure 1. Our style transfer method produces impressive results when applied to a range of artistic styles. The method preserves the underlying structure of the source images while transforming them into the desired styles.

Abstract

Diffusion models have shown great promise in text-guided image style transfer, but there is a trade-off between style transformation and content preservation due to their stochastic nature. Existing methods require computationally expensive fine-tuning of diffusion models or additional neural network. To address this, here we propose a zero-shot contrastive loss for diffusion models that doesn't require additional fine-tuning or auxiliary networks. By leveraging patch-wise contrastive loss between generated samples and original image embeddings in the pre-trained diffusion model, our method can generate images with the same semantic content as the source image in a zero-shot manner. Our approach outperforms existing methods while preserving content and requiring no additional training, not only for image style transfer but also for image-to-image translation and manipulation. Our experimental results validate the effectiveness of our proposed method. Code is available at https://github.com/YSerin/ZeCon.

1. Introduction

Style transfer is the task that converts the style of a given image into another style while preserving its content. Over the past few years, GAN-based methods such as pix2pix [19], cycleGAN [42], and contrastive unpaired image-to-image translation (CUT) have been developed [28]. Recently, joint use of a pretrained image generator and image-text encoder enabled text-guided image editing which requires little or no training of the networks [31, 6, 30, 14, 25].

Inspired by the success of diffusion models for image generation [16, 35], image editing [27], in-painting [1], super-resolution [5], etc., many researchers have recently investigated the application of the diffusion models for image-to-image style transfer [33, 36]. For example, [33, 34] proposed conditional diffusion models that require paired dataset for image-to-image style transfer. One of the limitations of these approaches is that the diffusion models need to be trained with paired data set with matched source and target styles. As collecting matched source and target domain data is impractical, many recent researchers have focused on unconditional diffusion models. Uncondi-

tional diffusion models have limitations in maintaining content due to the stochastic nature of the reverse sampling procedure that doesn't explicitly impose content consistency. As a result, content and styles can change simultaneously, creating challenges for maintaining content.

To tackle this problem, the dual diffusion implicit bridge (DDIB) [36] exploits two score functions that have been independently trained on two different domains. Although DDIB can translate one image into another without any external condition, it also requires training of two diffusion models for each domain which involves additional training time and a large amount of dataset. On the other hand, DiffusionCLIP [24] leverages the pretrained diffusion models and CLIP encoder to enable text-driven image style transfer without additional large training data set. Unfortunately, DiffusionCLIP still requires additional fine-tuning of the model for the desired style. Furthermore, DiffuseIT [26] uses disentangled style and content representation inspired by the slicing Vision Transformer [37]. Although DiffuseIT has shown its superiority in preserving content, it still suffers from the trade-off between transforming the texture of images and maintaining the content. Also, an additional network is required for computing content losses in DiffuseIT.

To address this problem, here we propose a simple yet effective Zero-shot Contrastive (ZeCon) loss for diffusion models to transfer the style of a given image while preserving its semantic content in a zero-shot manner. Our approach is based on the observation that a pre-trained diffusion model already contains spatial information in its embedding that can be used to maintain content through patch-wise contrastive loss between the input image and generated images. Unlike DiffusionCLIP, our method doesn't require additional training. In other words, we could effectively preserve the content in a zero-shot manner by leveraging the patch-wise contrastive loss. Furthermore, unlike DiffuseIT, our method achieves more accurate texture modification while preserving the content.

To demonstrate the effectiveness of our proposed method, we show a text-driven style transfer using CLIP [31]. However, our method can be extended for general guidance beyond text inputs. Furthermore, we demonstrate that our method can be applied to text-driven image-to-image translation and image manipulation tasks, illustrating its wide applicability.

2. Related Works

Image style transfer Neural style transfer [15] iteratively optimizes the content image to match the style image, which is time-consuming. Alternatively, adaptive instance normalization (AdaIN) [18] transfers the style of a source image to a target image by matching their feature statistics.

In contrast, pix2pix [19], CycleGAN [42], and CUT [28] use different mechanisms for content preservation. CycleGAN's cycle consistency for content preservation is often too restrictive, while CUT maximizes mutual information between content and stylized images in a patch-based feature space. This maintains structural information while changing appearance.

CLIP model [31] has been shown to have semantic representative power resulting from a large-scale dataset of 400 million image and text pairs, which allows for text-driven image manipulation. StyleCLIP [30] uses CLIP and pretrained StyleGAN [22] to optimize the latent vector of the content input given a text prompt, but its image modification is limited to the domain of the pretrained generator. StyleGAN-NADA [14] proposes an out-of-domain image manipulation method that shifts the generative model to new domains. VQGAN-CLIP [6] demonstrates that VQGAN [11] can also be used as a pretrained generative model to generate or edit high-quality images without training. CLIPstyler [25] proposes a CNN encoder-decoder model that learns both content and style properties through patch-wise CLIP loss, allowing for image generation and manipulation beyond the domains of pretrained generators.

Diffusion models for image style transfer Diffusion models have become popular due to their impressive ability to generate high-quality images [16, 35]. Diffusion models have found application in various computer vision areas, including super-resolution [32], segmentation [2], image editing [1], medical image processing [23], and video generation [17].

This generative model works by progressively adding Gaussian noise through a Markov chain forward process. Then, a trained noise estimation model is used to generate clean samples from the latent noise through an iterative denoising process. Specifically, DDPM [16] directly samples x_t from x_0 by adding Gaussian noise with $\beta_t \in (0,1)$ at time $t \in [1,...,T]$,

$$x_t = \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon \tag{1}$$

where $\epsilon \sim \mathcal{N}(0,I)$, $\alpha_t = 1-\beta_t$, and $\overline{\alpha}_t = \prod_{i=0}^{t} \alpha_i$. The reverse sampling process to generate a clean image is then given by:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(x_t,t)\right) + \sigma_t\epsilon. \tag{2}$$

where the neural network $\epsilon_\theta(x_t,t)$ is used to estimate the noise component, which can be viewed as a score function up to a scaling factor.

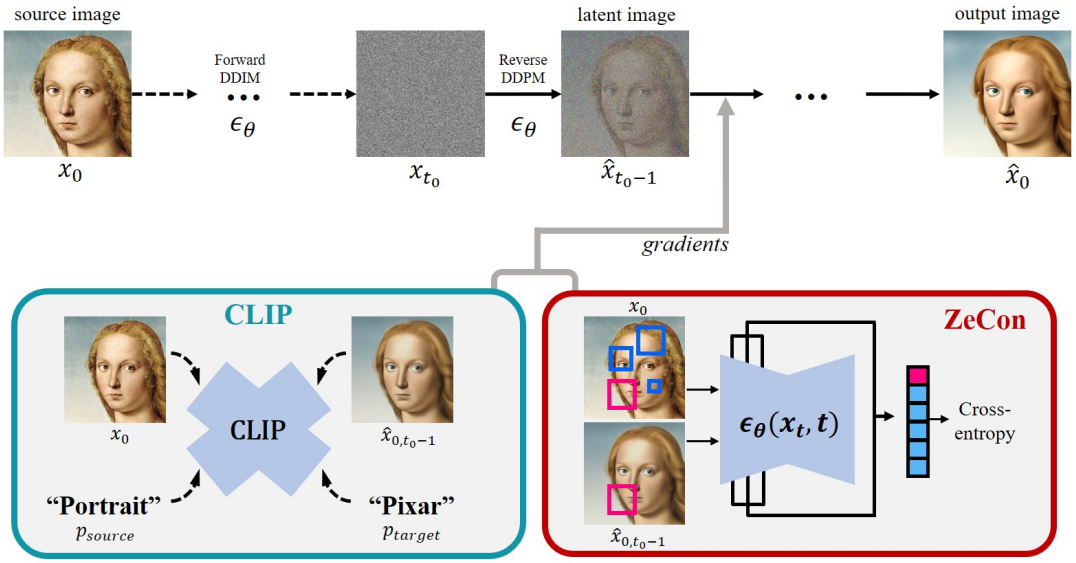

Figure 2. Our proposed method. To guide the diffusion model in our proposed method, we calculate the ZeCon loss using a noise estimator $\epsilon_\theta(\cdot)$ and the CLIP loss using the CLIP model. These losses allow us to add gradients to the denoised image at each time step.

While noise ϵ can help to achieve sample diversity in DDPM, it may also lead to a loss of content in the context of style transfer. The repeated application of stochastic operations can result in images with completely different content, even if the intermediate latent space is the same for each image. The content can be preserved with DDIM [35] whose sampling process is:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{x}_{0,t}(x_t) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_\theta(x_t,t) + \sigma_t^2\epsilon \tag{3}$$

where σ_t is the variance of noise which controls how stochastic the sampling process is, and $\hat{x}_{0,t}$ is a denoised image given by:

$$\hat{x}_{0,t}(x_t) := \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}. \tag{4}$$

If we set $\sigma_t = 0$ in (3), the noise term ϵ is eliminated, which allows us to preserve the content successfully. However, in this case, the sampling process becomes deterministic, which results in preserving the style as well. This is not desirable for style transfer, as illustrated in Figure 3.

To preserve semantics in style transfer, Palette [33] used conditional diffusion models that require paired datasets for training $\epsilon_\theta(x_t, t)$. On the other hand, ILVR [4] attempted to generate diverse samples conditioned on the input image using unconditional models, but the stochasticity introduced by ϵ still posed a challenge, as shown in Figure 3. Unconditional models were also employed in DDIB [36], which used two independently trained score functions for different domains, and in DiffusionCLIP [24], which fine-tuned a pre-trained diffusion model using identity and style losses. However, both methods require training or fine-tuning diffusion models for each style domain. Furthermore, DiffuseIT [26] necessitates an auxiliary network for computing a content loss.

3. Main Contributions

Sampling strategy Similar to (3), DDPM can be also represented as

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{x}_{0,t}(x_t) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_\theta(x_t,t) + \sigma_t^2\epsilon \tag{5}$$

if σ_t is given by

$$\sigma_t = \sqrt{(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}\sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t-1}} \tag{6}$$

We then define the loss function as follows

$$\ell_{total}(x) = \ell_{content}(x) + \ell_{CLIP}(x) \tag{7}$$

where $\ell_{content}$ and ℓ_{CLIP} denotes the content and style loss, respectively. Then, the denoised image estimate $\hat{x}_{0,t}(x_t)$ is supplemented with the gradient of the loss function:

$$\hat{x}_{0,t}(x_t) = \hat{x}_{0,t}(x_t) + \nabla_x \ell_{total}(x)|_{x=\hat{x}_{0,t}(x_t)} \tag{8}$$

Unlike DiffuseIT, which requires a substantial auxiliary network for computing content loss, our approach relies on a simpler but still effective content loss, as detailed below.

Content preservation loss In [28], it was demonstrated that the CUT loss effectively preserves structural information by maximizing the mutual information between input and output patches. Specifically, the original algorithm involves training an encoder to capture spatial information from the input. The resulting encoder features are then used to apply patch-wise contrastive loss, which utilizes the spatial information to preserve the contents.

Recent work by Baranchuk et al. [2] has shown that the U-Net noise predictor in the diffusion model contains spatial information. Therefore, a key contribution of this paper is to demonstrate that spatial features required for the CUT loss can be extracted from the diffusion model without additional training, as illustrated in Figure 2.

Specifically, at each reverse timestep $t \in [t_0, 0]$, both the original image x_0 and the reverse sampled denoised image $\hat{x}_{0,t}$ are forwarded to the noise estimator $\epsilon_\theta(x_t, t)$. The encoder part of the estimator is used to extract feature maps z_l and $\hat{z}_l$ for x_0 and $\hat{x}_{0,t}$, respectively. To apply patch-wise contrastive loss, the pixels of the feature maps are randomly selected and used to calculate cross-entropy loss. Pixels from the same location are considered "positive" and their mutual information is maximized. Pixels from different locations, considered "negative", have their mutual information minimized. This process can be expressed mathematically as:

$$\ell_{\text{ZeCon}}(\hat{x}_{0,\text{t}}, x_0) = \mathbb{E}_{x_0}\left[\sum_l \sum_s \ell(\hat{z}_l^s, z_l^s, z_l^{S\setminus s})\right] \tag{9}$$

Here, $\hat{z}_l$ and z_l denote the l-th layer features from $\hat{x}_{0,t}$ and x_0, respectively. s represents a spatial location in $1, \ldots, S_l$, where S_l is the number of spatial locations in feature z_l. The cross-entropy loss is denoted by $\ell(\cdot)$. By using the ZeCon loss in (9), we can maintain semantic consistency between the reverse sampled denoised image $\hat{x}_{0,t}$ and the original image x_0, preserving content information. More details can be found in the Supplementary material.

On top of the contrastive loss, we include the feature loss ℓ_{VGG}, which is the mean-squared error between the VGG feature maps of $\hat{x}_{0,\text{t}}$ and x_0, and the pixel loss ℓ_{MSE}, which is the ℓ_2 norm of the pixel difference between them.

$$\begin{aligned}\mathcal{L}_{\text{content}} = \ell_{\text{ZeCon}}(\hat{x}_{0,\text{t}}, x_0) \\ + \ell_{\text{VGG}}(\hat{x}_{0,\text{t}}, x_0) + \ell_{\text{MSE}}(\hat{x}_{0,\text{t}}, x_0)\end{aligned} \tag{10}$$

The weights for each loss function are hyperparameters which need to be chosen by users. The examples of these weights are given in the Supplementary material.

Style loss The CLIP model is trained on extensive language and image dataset which results in its great semantic power [31]. Thanks to this semantic capacity, we can generate images in diverse styles with only text prompts. The

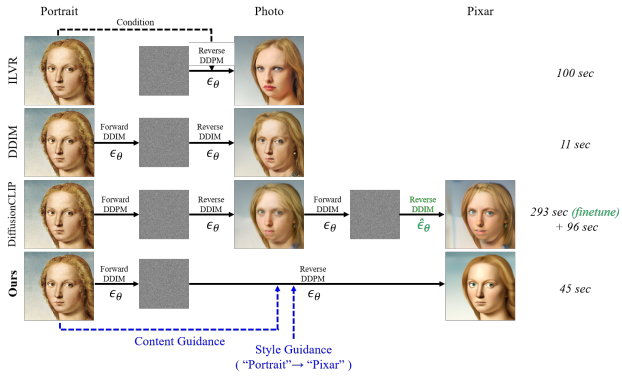

Figure 3. An illustration on sampling schemes of four diffusion models for style transfer.

CLIP loss for style guidance can be formulated as follows:

$$\begin{aligned}\ell_{\text{CLIP}} = \ell_{\text{global}}(\hat{x}_{0,\text{t}}, p_{\text{target}}) \\ + \ell_{\text{dir}}(\hat{x}_{0,\text{t}}, x_0, p_{\text{target}}, p_{\text{source}})\end{aligned} \tag{11}$$

Here, the global CLIP loss ℓ_{global} calculates the cosine distance in the CLIP embedding space between the generated image $\hat{x}_{0,t}$ and the style prompt p_{target} [30] by

$$\ell_{\text{global}}(\hat{x}_{0,\text{t}}, p_{\text{target}}) = D_{\text{CLIP}}(\hat{x}_{0,\text{t}}, p_{\text{target}}). \tag{12}$$

Since the global loss suffers from mode collapse and corrupted image quality, the directional CLIP loss ℓ_{dir} was proposed [14]. It aligns the direction in the CLIP embedding space between text and image pairs, which can be formulated as follows:

$$\ell_{\text{dir}}(\hat{x}_{0,\text{t}}, x_0, p_{\text{target}}, p_{\text{source}}) = 1 - \frac{\Delta I \cdot \Delta T}{\| \Delta I \| \| \Delta T \|}$$

where p_{source} denotes the source text prompt, and $\Delta I = E_{img}(x_0) - E_{img}(\hat{x}_{0,t})$, $\Delta T = E_{txt}(p_{\text{source}}) - E_{txt}(p_{\text{target}})$ for CLIP's image encoder E_{img} and text encoder E_{txt}. As the patch-based CLIP loss was proposed to enhance the generated images' quality [25], we adopt the patch-based scheme in both ℓ_{global} and ℓ_{dir}.

4. Experimental Results

4.1. Experimental setting

Dataset The images used as content reference are from FFHQ [21], CelebA-HQ [20], ImageNET [8], LSUN-Church [40], and CycleGAN dataset [42]. They contain images of human faces, objects, scenes, and churches. Furthermore, in order to evaluate the performance of our proposed model on the images from unseen domains, we utilize Wikiart dataset [7]. All the images are resized to 256 × 256 for the diffusion models. For patch-based guidance, we randomly crop 96 patches from a source image and then apply

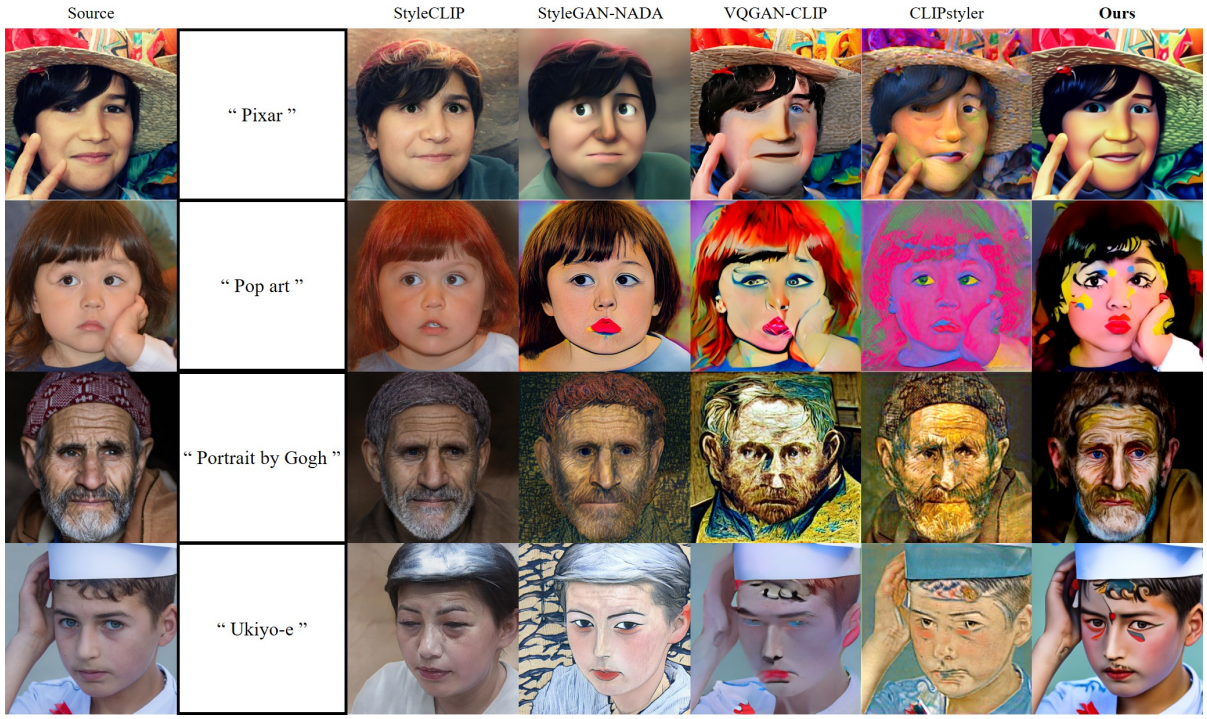

Figure 4. Comparison against GAN-based style transfer methods. When compared to four GAN-based methods, our approach achieves superior results in terms of style transformation and content preservation.

perspective augmentation and affine transformation. More details are illustrated in the Supplementary material.

Diffusion models We utilize the pre-trained unconditional diffusion model trained on ImageNET dataset with 256×256 image size [10] and the model trained on FFHQ dataset with 256×256 image size [4].

Either DDIM or DDPM method can be applied in our method during the forward and reverse diffusion steps. We basically adopt the DDIM strategy as the forward noising process and DDPM method as the reverse sampling. When T is the total time step, we respace the step size from T to T'. Then with the source image x_0, we obtain the latent x_{t_0} from the forward diffusion process, where $t_0 \in [0, T']$. We choose (T', t_0) as $(50, 25)$ as default when $T = 1000$. From this latent x_{t_0}, the stylized output image is sampled through diffusion processes. This approach not only preserves more latent information from the source image, but also enables the image to be effectively converted to a new style. Additionally, by reducing the number of iterations required, inference time can be significantly reduced.

The sampling scheme is illustrated in Figure 2 and 3, and the comparative studies on the choice of (T', t_0) are presented in the Supplementary material.

Methods	User study		CLIP score ↑	Face ID ↓
	Content ↑	Style ↑		
StyleCLIP	<u>4.10</u>	1.62	0.0925	**0.3750**
StyleGAN-NADA	3.42	2.94	0.1222	0.4948
VQGAN-CLIP	1.83	2.92	<u>0.1379</u>	0.7661
CLIPstyler	1.99	<u>2.96</u>	0.1347	0.6664
Ours	**4.61**	**4.23**	**0.1479**	<u>0.3881</u>

Table 1. User study and quantitative results for comparison with GAN-based methods for style transfer. The bold text and underline refer to the best and second best results, respectively.

4.2. Comparative studies

Figure 1 shows that our method achieves outstanding results across various artistic styles. In addition, we perform comparisons with GAN-based and diffusion-based style transfer methods, respectively.

Comparison with GAN-based models For GAN-based models, we compare four state-of-the-art methods - StyleCLIP [30], StyleGAN-NADA [14], VQGAN-CLIP [6], CLIPstyler [25]. The results of the comparison are illustrated in Figure 4.

Our proposed model clearly outperforms other methods in terms of retaining content. The outputs generated by StyleCLIP and StyleGAN-NADA exhibit distorted results where non-face objects, such as hands or hats, are removed from the output images. While results from VQGAN-CLIP

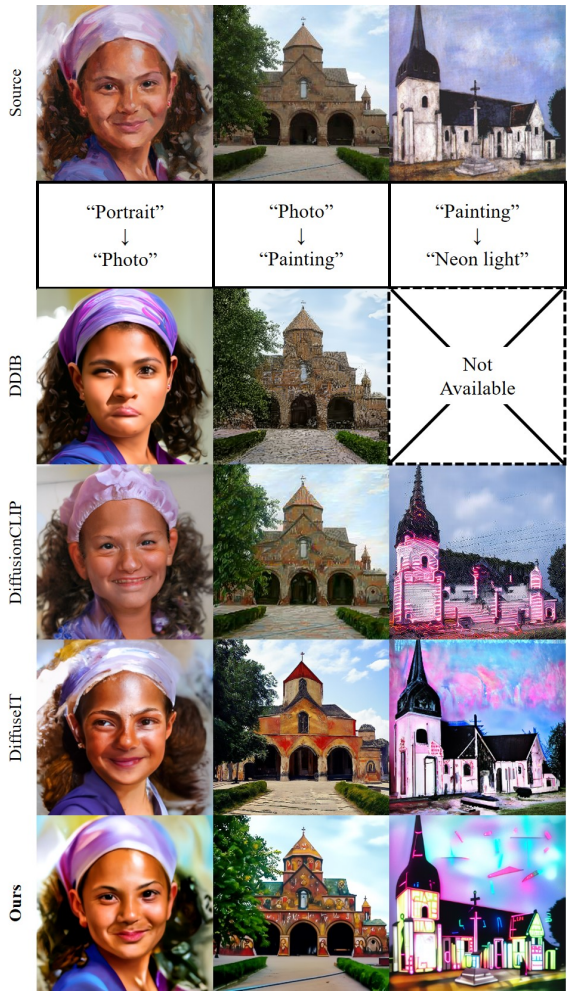

Figure 5. Comparing three diffusion-based style transfer methods, our proposed approach stands out by allowing for style modulation from unseen domain images, which the other diffusion models cannot achieve.

Methods	Photo domain		Unseen domain	
	Content ↑	Style ↑	Content ↑	Style ↑
DiffusionCLIP	3.71	2.95	3.29	3.05
Ours	**4.76**	**4.71**	**4.62**	**4.71**

Table 2. User study results on comparison with DiffusionCLIP.

Methods	DiffusionCLIP	DiffuseIT	Ours
CLIP score ↑	0.1220	0.1141	**0.1600**
Face ID ↓	0.8005	0.5228	**0.3240**

Table 3. Comparison with DiffusionCLIP and DiffuseIT.

and CLIPstyler show relatively better preservation of facial features, such as eyes and mouth, they still suffer from some loss of detail.

In contrast, our proposed method maintains structural information and retains hats and hands in the outputs, without

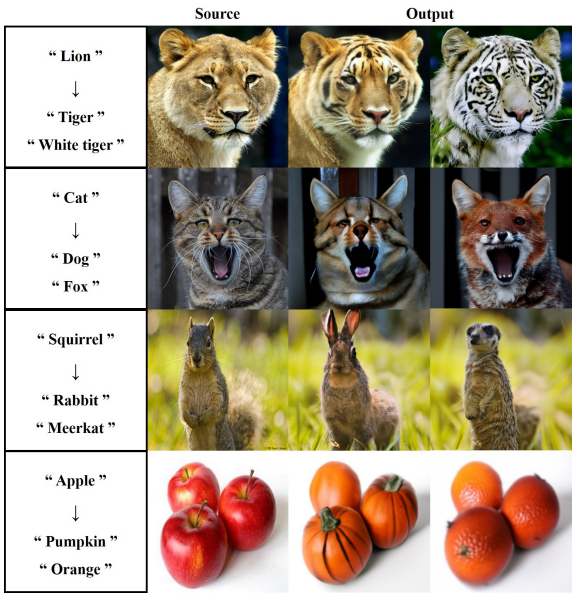

Figure 6. Image translation results.

crushing any details of the face, such as hairs or eyes. Additionally, our method generates outputs with feasible texture, unlike StyleCLIP which produces outputs that still look like photos, or StyleGAN-NADA which struggles to translate images into pop art style. Similarly, VQGAN-CLIP and CLIPstyler fail to generate Pixar and Uiyo-e style images, while our method provides high-fidelity samples transferred into the styles of the target prompt.

Our proposed method's superiority is further supported by quantitative evaluation. As shown in Table 1, our method achieves the highest scores in both user study and CLIP score. While StyleCLIP obtains the smallest face identity loss, this suggests that it preserves semantic information to such an extent that it fails to transform the style adequately. On the other hand, VQGAN-NADA and CLIPstyler overmodulate the images, resulting in significant content alteration. In contrast, our method achieves a balance between content preservation and style transfer. The CLIP score is calculated globally, as described in equation (12), and in a patch-based manner. Face identity loss is measured using ArcFace [9]. The same images used in the user study are used for the quantitative experiments.

Comparison with diffusion models We compared our proposed method with three diffusion-based models, DDIB [36], DiffusionCLIP [24], and DiffuseIT [26]. To evaluate the translation performance of DDIB between painting and photo domains, we trained a new diffusion model on the Wikiart dataset. For the photo domain, we used the pretrained diffusion models described above. The qualitative and quantitative results of the comparison are presented in Figure 5 and Table 2.

Figure 7. Image manipulation results with emotional prompts.

The third row of Figure 5 shows that DDIB suffers from identity loss, where the facial identity of the portrait is destroyed in the translation from portrait to photo domain. Moreover, the shape of the church is not well delineated in the output of DDIB. Additionally, diffusion models have to be trained for each new domain, which is a critical drawback of DDIB. Therefore, image translation from portrait to neon light style is not available with DDIB.

On the other hand, DiffusionCLIP shows relatively satisfying quality in translating photos into another style. However, when the input images are not well converted into the photo domain, the results from unseen domain images are unsatisfactory, as shown in the first and third columns of Figure 5. This is supported by user study results on DiffusionCLIP, as presented in Table 2, where the content score in unseen domains is 0.42 lower than the score in the photo domain. Furthermore, DiffuseIT shows the trade-off between style transfer and content preservation as shown in the fifth row of Figure 5. While changing the style of the source image, the facial identity is also modified. As demonstrated in the last column, the neon light is hardly seen when the shape of the church is well-preserved.

In contrast, our proposed method can stylize images not only from photo domains but also from unseen domains, such as portraits or paintings. The portrait is transformed into a photo while maintaining its facial identity, and the painting of a church is translated into neon light style while retaining small objects like a cross. These results are confirmed with user study results presented in Table 2, where the scores between the photo domain and unseen domains are highly similar. This means that our method can modulate images even from unseen domains. Regarding computational time, as shown in Table 4, our method is significantly faster than DiffusionCLIP. Additional examples are provided in the Supplementary material.

Figure 8. Image manipulation results with human faces.

Methods	Data preparation	# Train Param.	Training time	Inference time (sec)
ILVR	-	-	-	100
DDIM	-	-	-	11
DDIB	-	1104 M	$>$ 200 hrs	12
DiffusionCLIP	5.85 min	113 M	293 sec	96
DiffuseIT	-	-	-	40
Ours	-	-	-	38

Table 4. Comparison on computational complexity of various diffusion models. The symbol "-" indicates that data preparation or training is not required.

Figure 9. Comparative study results on image manipulation.

Image manipulation Our proposed method not only excels in image style transfer but also has potential for other tasks such as simple image translation and manipulation. The qualitative results for image translation are shown in Figure 6, where our method can translate different animal species while preserving the details and maintaining the overall coherence of the image. In addition, our method can also change from apples to pumpkins or oranges, as shown in the second row of Figure 6.

Moreover, our method can also be used for image manipulation tasks, as demonstrated in Figure 7 and Figure 8. Figure 7 shows an example of changing the expression of animals, while Figure 8 shows an example of appearance manipulation such as age, gender and make-up. These results demonstrate the potential of our method for various image manipulation tasks.

We performed comparative studies using four alternatives: Plug-and-Play[38], InstructPix2Pix[3], EGSDE[41], and Pix2Pix-zero[29]. As illustrated in Figure 9, all images were effectively translated into the dog class. However, the outcomes from the comparative methods demonstrated inferior performance in identity preservation. In contrast, our proposed method excelled in maintaining the cat's identity.

4.3. Ablation studies

Roles of content losses To investigate the effectiveness of content guidance losses, we performed ablation studies. Our proposed content loss for guidance in (10) consists of three different losses, namely ℓ_{ZeCon}, ℓ_{VGG}, and ℓ_{MSE}. To

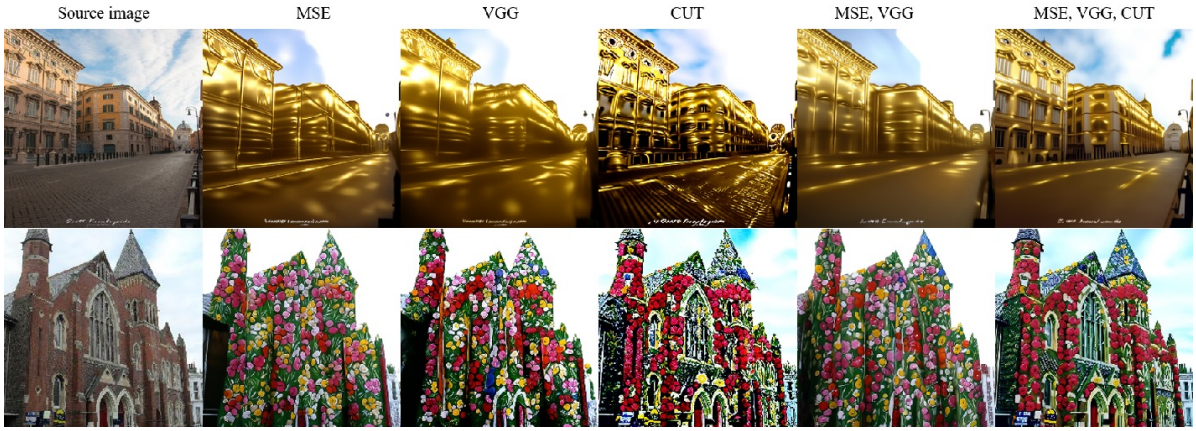

Figure 10. Ablation study focusing on three losses for content guidance - ℓ_{MSE}, ℓ_{VGG}, and ℓ_{ZeCon}. The results show that, in each row of translated images, which are transformed into the styles of "golden" and "oil painting of flowers", the proposed patch-wise content preservation loss is effective in preserving content information.

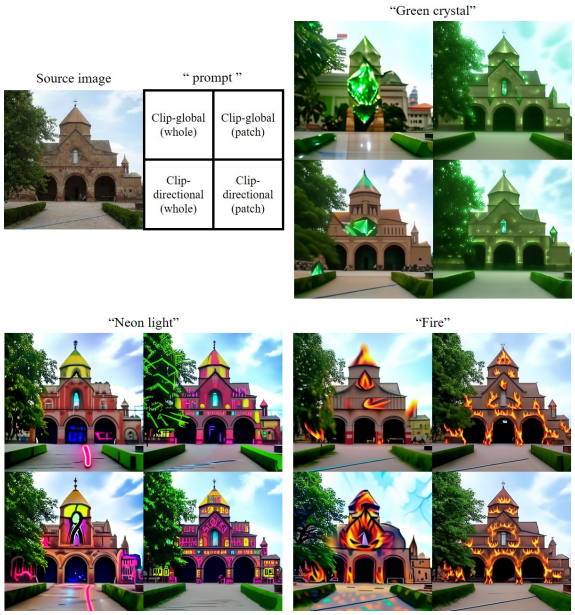

Figure 11. Ablation study on four losses for style guidance - CLIP global loss, patch-based CLIP global loss, CLIP directional loss, and patch-based CLIP directional loss.

examine the contribution of ℓ_{ZeCon}, we eliminated it from the total content loss and compared the results with the complete content loss. In Figure 10, we observed that excluding ℓ_{ZeCon} resulted in a loss of structural details such as windows in the building outlines, even though the overall shape was preserved. This suggests that ℓ_{VGG} and ℓ_{MSE} alone are insufficient to preserve the fine-grained details of the content.

On the other hand, employing all three losses yielded the best results in terms of content preservation. The user study results, presented in Table 5, support the superiority of ℓ_{ZeCon} compared to ℓ_{MSE}, ℓ_{VGG}, and even the combination of both. This implies that ℓ_{ZeCon} effectively preserves the structural details while avoiding over-fitting. In summary, the ablation studies demonstrate the crucial role of ℓ_{ZeCon} in our proposed method for preserving the structural properties of the input images.

	Content ↑	Style ↑
MSE	3.29	3.81
VGG	3.00	3.86
ZeCon	<u>4.29</u>	<u>4.57</u>
MSE, VGG	3.29	3.95
MSE, VGG, ZeCon	**4.81**	**4.81**

Table 5. User study results on ablation studies regarding content losses. Bold text and underline refer to the best and the second best scores, respectively.

Roles of style losses We conducted ablation studies to investigate the role of each loss function in our style loss. The loss function consists of two parts, ℓ_{global} and ℓ_{dir}, as shown in (11). To examine the contribution of each loss, we applied the loss functions individually and evaluated the results on three different styles - green crystal, neon light, and fire. The qualitative results are shown in Figure 11. We found that applying the directional CLIP loss in addition to the global CLIP loss led to more stylized images, as compared to applying the global CLIP loss alone. This implies that directional CLIP loss is more effective in modulating the style of images.

We also verified the role of patch-based guidance in our proposed method. We observed that using whole-image guidance tends to stylize the image in local parts, while patch-based guidance transforms the image into the given style by covering a large area. This is demonstrated in Figure 11, where the patch-based guidance is applied to stylize the entire background while preserving the foreground object.

Overall, the results of our ablation studies suggest that both ℓ_{global} and ℓ_{dir} are important for achieving high-quality style transfer results, and that patch-based guidance is effective in transforming images into a given style.

5. Conclusion

In this paper, we proposed a novel method for diffusion-based image style transfer without content changes using Zero-Shot Contrastive (ZeCon) loss. One of the major advantages of our method is that it is training-free and does not require additional training or data, which significantly reduces the computational time. Our experiments showed that contrastive loss with diffusion model leads to high capability in maintaining content while achieving effective stylization. Additionally, our method demonstrated potential for image translation and manipulation tasks. The limitations of our method are discussed in the Supplementary material.

6. Acknowledgement

This work was supported in part by the National Research Foundation of Korea under Grant NRF-2020R1A2B5B03001980; and in part by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 1711137899, KMDF_PR_20200901_0015);

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 1, 2

[2] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022. 2, 4

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 7, 14

[4] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14367–14376, October 2021. 3, 5

[5] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12413–12422, 2022. 1, 11

[6] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. *arXiv preprint arXiv:2204.08583*, 2022. 1, 2, 5

[7] Michael Danielczuk, Matthew Matl, Saurabh Gupta, Andrew Li, Andrew Lee, Jeffrey Mahler, and Ken Goldberg. Segmenting unknown 3d objects from real depth images using mask r-cnn trained on synthetic data. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2019. 4

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 5, 11

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2

[12] Heng Fan and Haibin Ling. Sanet: Structure-aware network for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 42–49, 2017. 12

[13] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven artistic style transfer. In *European Conference on Computer Vision*, pages 717–734. Springer, 2022. 12

[14] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 1, 2, 4, 5

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1, 2, 11

[17] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 2

[18] 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, 12

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 2

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 4

[21] 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. 4

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[23] Boah Kim, Inhwa Han, and Jong Chul Ye. Diffusemorph: Unsupervised deformable image registration along continuous trajectory using diffusion models. *arXiv preprint arXiv:2112.05149*, 2021. 2

[24] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 2, 3, 6, 11

[25] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 1, 2, 4, 5, 11

[26] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022. 2, 3, 6

[27] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. *arXiv preprint arXiv:2112.05744*, 2021. 1

[28] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for conditional image synthesis. In *ECCV*, 2020. 1, 2, 4

[29] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 7

[30] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021. 1, 2, 4, 5

[31] 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. 1, 2, 4

[32] 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

[33] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 1, 3

[34] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021. 1

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2, 3

[36] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022. 1, 2, 3, 6

[37] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 2

[38] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, June 2023. 7, 13

[39] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9045, 2019. 12

[40] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 4

[41] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *arXiv preprint arXiv:2207.06635*, 2022. 7

[42] 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. 1, 2, 4