# TEXT-GUIDED DIFFUSION IMAGE STYLE TRANSFER WITH CONTRASTIVE LOSS FINE-TUNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Recently, diffusion models have demonstrated superior performance in text-guided image style transfer. However, there exists fundamental trade-off between transforming styles and maintaining content in diffusion models. Although a simple remedy would be using deterministic sampling scheme such as denoising diffusion implicit model (DDIM) that guarantees the perfect reconstruction, it requires the computationally expensive fining-tuning of the diffusion models. To address this, here we present a text-guided sampling scheme using a patch-wise contrastive loss fine-tuning. By exploiting the contrastive loss between the samples and the original images, our diffusion model can generate an image with the same semantic content as the source image. Experimental results demonstrate that our approach outperforms the existing methods while maintaining content and requiring no additional training on the diffusion model.

## 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 (Isola et al., 2017), cycleGAN (Zhu et al., 2017), and contrastive unpaired image-to-image translation (CUT) have been developed (Park et al., 2020). 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 (Radford et al., 2021; Crowson et al., 2022; Patashnik et al., 2021; Gal et al., 2022; Kwon & Ye, 2022).

Inspired by the success of diffusion models for image generation (Ho et al., 2020; Song et al., 2020), image editing (Liu et al., 2021), in-painting Avrahami et al. (2022), super-resolution Chung et al. (2022), etc., many researchers have recently investigated the application of the diffusion models for image-to-image style transfer (Saharia et al., 2022; Su et al., 2022). For example, (Saharia et al., 2022; 2021) 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. To enable diffusion-based image style transfer without matched data set, many recent researchers have focused on unconditional diffusion models.

Figure 1: Results of our style transfer method on various artistic styles. The source images are translated into various styles while maintaining their structure.

For example, the dual diffusion implicit bridge (DDIB) (Su et al., 2022) 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 for each domain which involves additional training time and a large amount of dataset. On the other hand, DiffusionCLIP (Kim et al., 2022) 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 training to finetune the model for the desired style. Besides the additional complexity, unconditional diffusion models for image style transfer have further limitations in maintaining content. This is because the reverse sampling procedure of the diffusion models does not have an explicit constraint to impose the content consistency and the stochastic nature of diffusion models makes them easy to change the content and styles at the same time.

To address this, here we propose a diffusion model that transfers the style of a given image while preserving its semantic content by using contrastive loss similar to CUT (Park et al., 2020) that is fine-tuned during the run time by text guide. Since contrastive loss can exploit the spatial information in terms of positive and negative pairs, we found that it shows remarkable performance in maintaining the content. Furthermore, in contrast to DiffusionCLIP, it only requires the CUT loss fine-tuning via light weighted multi-layer perceptron (MLP) rather than the diffusion model, so the computational complexity can be significantly reduced.

To verify the effectiveness of this method, we present a text-driven style transfer using CLIP (Radford et al., 2021). In particular, we utilize CLIP in a patch-wise manner similar to (Kwon & Ye, 2022) thanks to its stable style translation. Our contribution can be summarized as following:

- Thanks to the content disentanglement using contrastive loss, to our best knowledge, our method is the first style transfer method with unconditional diffusion model that overcomes the trade-off between style and content.
- Our method only requires fine-tuning of contrastive loss in terms of MLP rather than the diffusion model for target domain, so the computational complexity is much low but still allows effective image transfer to any unseen domain.

## 2 Related Works

**Image style transfer** Neural style transfer (Gatys et al., 2016) is the first approach to change the style texture of the content image into a style image by iterative optimization process. However, these iterative process takes significant amount of time. Alternatively, the adaptive instance normalization (AdaIN) by (Huang & Belongie, 2017) converts the means and variances of the features of the source image to those of the target image, which enables arbitrary style transfer.

On the other hand, pix2pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017) and CUT (Park et al., 2020) rely on different mechanism for content preservation. Specifically, in CycleGAN (Zhu et al., 2017), the cycle consistency assumes bijective relationship between two domains for content preservation, whose constraint is often restrictive in some applications. In order to overcome this restriction, CUT (Park et al., 2020) was proposed to maximize the mutual information between the content input and stylized output images in a patch-based manner on the feature space. This leads to correspondence in the structure between the two images while changing appearance.

With the advent of CLIP model (Radford et al., 2021), it has been shown that text-guided image synthesis can be accomplished without collecting style images. CLIP has semantic representative power which results from large scale dataset consisting of 400 millions image and text pairs. This enables text-driven image manipulation. StyleCLIP (Patashnik et al., 2021) was proposed to optimize latent vector of the content input given text prompt by using CLIP and pretrained StyleGAN (Karras et al., 2020). However, image modification using StyleCLIP is limited to the domain of the pretrained generator. In order to solve this issue, StyleGAN-NADA (Gal et al., 2022) presented out-of-domain image manipulation method that shifts the generative model to new domains. VQGAN-CLIP (Crowson et al., 2022) has shown that VQGAN (Esser et al., 2021) can also be used as a pretrained generative model to generate or edit high quality images without training. In order not to be restricted to the domains of the pretrained generators, CLIPstyler (Kwon & Ye (2022)) proposed a CNN encoder-decoder model that learns both content and style properties through patch-wise CLIP loss.

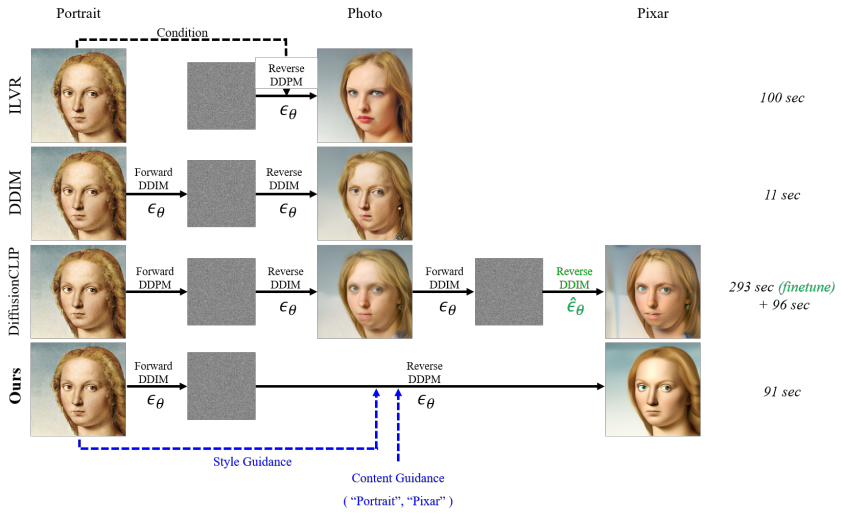


Figure 2: An illustration on sampling schemes of four diffusion models for style transfer.

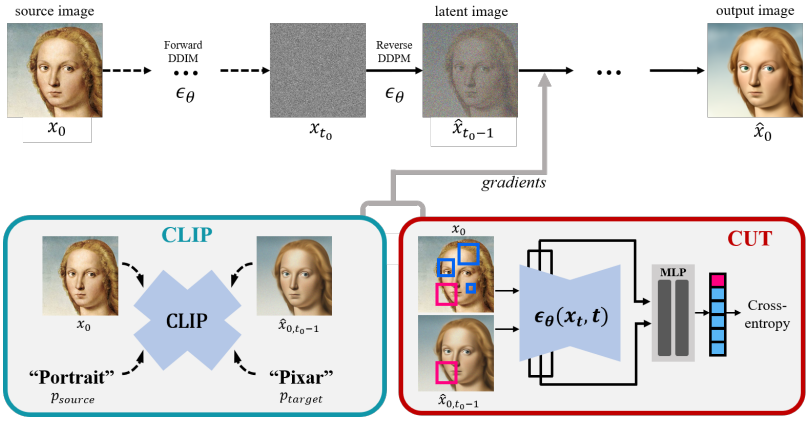


Figure 3: Our proposed method. In order to guide the diffusion model, CUT loss is calculated using the noise estimator $\epsilon_\theta(\cdot)$ and CLIP loss using the CLIP model. Through these losses, gradients are added to the predicted mean at each time step. More details regarding CUT implementation on the diffusion model are described in METHOD and APPENDIX sections.

**Diffusion models for style transfer** Diffusion probabilistic models have attracted great attention because of their superior performance in generating images despite their long training time (Ho et al., 2020; Song et al., 2020). The diffusion model is the generative model that involves Markov chain of forward process by gradually adding Gaussian noise (Ho et al., 2020). With noise estimation model, samples are generated from the latent noise by iterative denoising process. They have been applied to a wide variety of computer vision areas involving super resolution (Rombach et al., 2022), segmentation (Baranchuk et al., 2022), image editing (Avrahami et al., 2022), medical image processing (Kim et al., 2021), and video generation (Ho et al., 2022).

Recently, there have been several attempts to translate images from one domain to another using diffusion models. Palette (Saharia et al., 2022) demonstrated that various types of image-to-image translation can be performed by utilizing conditional diffusion models which require paired dataset for training. For unconditional diffusion models, DDIB (Su et al., 2022) proposed to exploit two score functions which have been independently trained on two different domains. However, it requires training diffusion models on each style domain, which takes a significant amount of time. Meanwhile, DiffusionClip (Kim et al., 2022) was proposed for image manipulation using a pre-trained diffusion model. Specifically, DiffusionCLIP fine-tunes the model with identity and style losses, which explicitly imposes constraints on transferring appearance while maintaining structure.

Though it could be trained easily with semantic knowledge of CLIP, the model need to be fine-tuned for each style domain. Additionally, diffusion models with guidance from regression models have been proposed for image-to-image translation tasks (Wolleb et al., 2022). Although it could transform the content images with the desired style without additional training on diffusion models, its generated image suffers from inconsistency with the content image.

# 3 METHODS

## 3.1 DILEMMA OF DIFFUSION MODELS

With respect to the style transfer, the generative process should convert the image into the given style while retaining the content. However, diffusion models have difficulties in maintaining the semantic information. Specifically, in diffusion models, given a source data distribution $q(x_0)$, latent variable $x_t$ is computed by forward diffusion process. DDPM (Ho et al., 2020) directly samples $x_t$ from $x_0$ by gradually adding Gaussian noise with $\beta_t \in (0, 1)$ at time $t \in [1, ..., T]$ which is fixed or learned variance schedule,

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

where $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$, $\alpha_t = 1 - \beta_t$, and $\overline{\alpha}_t = \prod_{i=0}^{t} \alpha_i$. The reverse sampling process is:

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

Although the noise $\epsilon$ contributes to achieve sample diversity, it in return leads to a loss of content in the context of style transfer. The stack of these stochastic steps results in the images with completely different content even if the images are sampled from the intermediate latent, e.g. $x_{200}$. In order to preserve the semantics, ILVR (Choi et al., 2021) in Fig. 2 tried to generate diverse samples with image condition. However, it still cannot generate images with same semantics and desired texture. After the conditioning ranges, the generative process suffers from the stochasticity by $\epsilon$. Meanwhile, the content can be preserved with DDIM (Song et al., 2020) whose sampling process is:

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

where $\sigma_t$ is the variance of noise which controls how stochastic the sampling process is. The clean image can be predicted from $x_t$ and $\epsilon_\theta(x_t, t)$ at each time step $t$ as follows:

$$f_\theta(x_t, t) := \frac{x_t - \sqrt{1 - \overline{\alpha}_t} \epsilon_\theta(x_t, t)}{\sqrt{\overline{\alpha}_t}}. \tag{4}$$

When $\sigma_t = 0$, the noise term $\epsilon$ is removed from (3) and then we can successfully preserve the content. Since the sampling process is deterministic, however, the style is also preserved, which is not desired for style transfer as shown in Fig. 2. DiffusionCLIP in Fig. 2 tried to overcome this problem by finetuning the score function, which takes much time and computation. In order to solve this dilemma, we propose diffusion-based style transfer method using guidance that requires no additional training on the diffusion model and is applicable even to unseen domains.

## 3.2 GUIDANCE FOR DIFFUSION MODELS

Guiding gradients in diffusion models is the method proposed in the context of class-conditional image generation (Dhariwal & Nichol, 2021). Even the unconditional diffusion model can generate conditional images using classifiers or CLIP. We propose to utilize this guidance method in style transfer task with the following losses. As proposed in the past work (Avrahami et al., 2022), gradients for the guidance is calculated as

$$\nabla_{\hat{x}_{0,t}} L_{total}(\hat{x}_{0,t}, x_0, p_{target}, p_{source})$$

where $p_{target}$ and $p_{source}$ are text prompts for target and source domain, respectively. From DDPM, the estimated clean image $\hat{x}_{0,t}$ from the latent variable $x_t$ (Avrahami et al., 2022) can be produced from the noise approximation model $\epsilon_\theta(x_t, t)$:

$$\hat{x}_{0,t} = \frac{x_t}{\sqrt{\overline{\alpha}_t}} - \frac{\sqrt{1 - \overline{\alpha}_t} \epsilon_\theta(x_t, t)}{\sqrt{\overline{\alpha}_t}}. \tag{5}$$

The gradients are then added to the predicted mean at each time step $t$. Thanks to the guidance method, diffusion models do not require additional training. In order to achieve superior performance, we incorporate two types of guidance - style and content guidance. The guiding method is illustrated in Figure 3.

**CLIP loss for style guidance** CLIP model is trained on extensive language and image dataset which results in its great semantic power (Radford et al., 2021). Thanks to this semantic capacity, we can generate images in diverse styles with only text prompts. The CLIP loss for style guidance can be formulated as:

$$\mathcal{L}_{CLIP} = \mathcal{L}_{global}(\hat{x}_{0,t}, p_{target}) + \mathcal{L}_{directional}(\hat{x}_{0,t}, x_0, p_{target}, p_{source}) \tag{6}$$

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 $p_{target}$ (Patashnik et al., 2021) as follows:

$$\mathcal{L}_{global}(\hat{x}_{0,t}, p_{target}) = D_{CLIP}(\hat{x}_{0,t}, p_{target}). \tag{7}$$

Since the global loss has shortcomings of mode collapse and corrupted image quality, the directional CLIP loss $L_{directional}$ was proposed (Gal et al., 2022). It aligns the direction in the CLIP embedding space between text and image pairs. In our case, it can be formulated as:

$$\mathcal{L}_{directional}(\hat{x}_{0,t}, x_0, p_{target}, p_{source}) = 1 - \frac{\Delta I \cdot \Delta T}{\parallel \Delta I \parallel \parallel \Delta T \parallel} \tag{8}$$

where $\Delta I = E_{img}(x_0) - E_{img}(\hat{x}_{0,t})$, $\Delta T = E_{txt}(p_{source}) - E_{txt}(p_{target})$ for CLIP's image encoder $E_{img}$ and text encoder $E_{txt}$. Meanwhile, the patch-based CLIP loss was proposed to enhance the generated images' quality (Kwon & Ye, 2022). So, we adopt the patch-based scheme in both $L_{global}$ and $L_{directional}$.

**CUT loss for content guidance** With respect to style transfer task, it is essential that diffusion models not only transform the style, but also preserve the content. CUT loss (Park et al., 2020) has been proven to effectively maintain structure information by maximizing mutual information between corresponding input and output patches. It requires training an encoder which can capture the spatial information from the input. Features $z$ extracted from the encoder are then used for contrastive learning. Meanwhile, the noise predictor, U-net in the diffusion model, has been shown to contain spatial information (Baranchuk et al., 2022). Thus, we could easily apply the patch-wise contrastive loss in order to preserve contents by utilizing the spatial information extracted from the diffusion model without its additional training as shown in Figure 3. It can be formulated as follows:

$$\mathcal{L}_{CUT}(\hat{x}_{0,t}, x_0) = \mathbb{E}_{\mathrm{x}_0}\left[\sum_l \sum_s \ell(\hat{z}_\ell^s, z_\ell^s, z_\ell^{S\setminus s})\right] \tag{9}$$

where $\hat{z}_l$ and $z_l$ are $\epsilon_\theta$'s $l$-th layer features from $\hat{x}_{0,t}$ and $x_0$, respectively. We denote $s$ as a spatial location in $\{1, \ldots, S_l\}$ where $S_l$ is the number of spatial locations in the feature $z_l$. Additionally, $\ell(\cdot)$ is cross-entropy loss and more details are described in the Appendix.

Meanwhile, it is important how contrastive loss for content guidance is incorporated into the diffusion model. As illustrated in the Figure 3, feature maps are extracted from the encoder part of the U-net $\epsilon_{\theta,enc}(x_t, t)$. Since these feature maps have different channel sizes, two linear layers are utilized. They are denoted as network $F_l$ for each feature map from $l$th layer. Their number of output channels are 256. ReLU activation is used between the two linear layers. $L2$ norm of the outputs become the final outputs, which are then utilized to calculate cross entropy loss. The outputs can be written as:

$$z_l = F_l(\epsilon_{\theta,enc}(x_0)), \quad \hat{z}_l = F_l(\epsilon_{\theta,enc}(\hat{x}_0)). \tag{10}$$

Though $\epsilon_\theta(x_t, t)$ is fixed, there exists MLP before the entropy loss that needs training. At the first sampling, MLP is trained while all the other weights are fixed. For each reverse diffusion time steps, MLP weights are updated and then gradients for guidance are calculated. Then, at the second stage, all the networks are fixed and the outputs are generated with guidance.

On top of the contrastive loss, we include the feature loss $L_{VGG}$, which is the mean-squred error between the VGG feature maps of $\hat{x}_{0,t}$ and $x_0$, and the pixel loss $L_{MSE}$, which is the $L_2$ norm of the pixel difference between them.

$$\mathcal{L}_{content} = \mathcal{L}_{CUT}(\hat{x}_{0,t}, x_0) + \mathcal{L}_{VGG}(\hat{x}_{0,t}, x_0) + \mathcal{L}_{MSE}(\hat{x}_{0,t}, x_0) \tag{11}$$

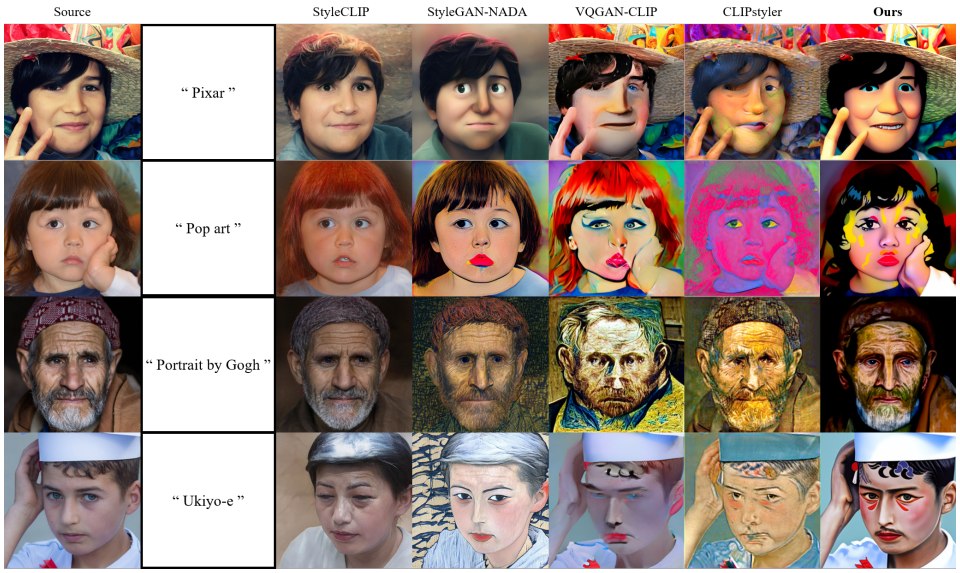


Figure 4: Comparison against GAN-based style transfer methods. Our method outperforms four GAN-based methods in perspective of both style transformation and content preservation.

| Methods | User study | | CLIP score ↑ | Face ID ↓ |
|---|---|---|---|---|
| | Content ↑ | Style ↑ | | |
| StyleCLIP | 3.84 | 1.46 | 0.0925 | **0.3750** |
| StyleGAN-NADA | 3.15 | 2.71 | 0.1222 | 0.4948 |
| VQGAN-CLIP | 1.83 | 2.79 | 0.1379 | 0.7661 |
| CLIPstyler | 2.05 | 2.93 | 0.1347 | 0.6664 |
| Ours | **4.33** | **4.39** | **0.1483** | <u>0.4219</u> |

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

Therefore, the total loss function for the guidance is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{CLIP}(\hat{x}_{0,t}, x_0, p_{target}, p_{source}) + \mathcal{L}_{content}(\hat{x}_{0,t}, x_0)$$

# 4 Experiments

## 4.1 Experimental setting

**Dataset** The images used as content reference are from FFHQ (Karras et al., 2019), CelebA-HQ (Karras et al., 2017), ImageNET (Deng et al., 2009), LSUN-Church (Yu et al., 2015), and CycleGAN dataset (Zhu et al., 2017). 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 (Danielczuk et al., 2019) consisting of more than 50 thousands painting images. All the images are resized to 256 $\times$ 256 for the diffusion models. For patch-based guidance, we randomly crop 96 patches from a source image and then apply perspective augmentation and affine transformation. More details are illustrated in the Appendix.

**Pretrained diffusion models** We utilize the unconditional diffusion model trained on ImageNET dataset with 256 $\times$ 256 image size (Dhariwal & Nichol, 2021) and the model trained on FFHQ dataset with 256 $\times$ 256 image size (Choi et al., 2021). The weights of the pretrained diffusion models are fixed during all sampling processes.

**Sampling scheme** 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 the total time step $T$ is 1000, we respace the step size from 1000 to 50. Then with $T' = 50$, we forward the source image $x_0$ to latent image $x_{t_0}$ where

Figure 5: Comparison between three diffusion-based style transfer methods. Our proposed method can modulate styles for unseen domain images that other diffusion models cannot.

| Methods | Photo domain | | Unseen domain | |
|---|---|---|---|---|
| | Content ↑ | Style ↑ | Content ↑ | Style ↑ |
| DiffusionCLIP | 4.05 | 2.90 | 3.20 | 2.75 |
| Ours | **4.70** | **4.70** | **4.55** | **4.55** |

Table 2: User study results on comparison against DiffusionCLIP.

$t_0 = 25$. From this latent image, the stylized output image is sampled through diffusion processes with style and content guidance. In this way, not only can more latent information be preserved from the source image, but the image can be sufficiently converted to a new style at the same time. By greatly reducing the number of iterations, inference time could be significantly reduced, which is the critical drawback of the diffusion models. The sampling scheme is illustrated in the Figure 2, 3 and the Appendix.

### 4.2 Comparative Studies

Figure 1 shows that our method achieves outstanding results on various artistic styles. In addition, we demonstrate remarkable performance of our proposed method through comparisons between GAN-based and diffusion-based style transfer methods, respectively.

#### 4.2.1 GAN-based models

For GAN-based models, we compare four state-of-the-art methods - StyleCLIP (Patashnik et al., 2021), StyleGAN-NADA (Gal et al., 2022), VQGAN-CLIP (Crowson et al., 2022), CLIPstyler (Kwon & Ye, 2022). The results of the comparison are illustrated in Figure 4. We could clearly see that our proposed model outperforms in aspect of retaining the content. The generated outputs from both StyleCLIP and StyleGAN-NADA show distorted results that non-face objects, such as hands or hats, are removed from the output images. Although results from VQGAN-CLIP and CLIPstyler show relatively better result in perspective of content, we observe that some details of the face, such as eyes or mouth, are crumbled. This leads to larger face identity loss (Deng et al., 2019) in VQGAN-CLIP and CLIPstyler as shown in the Table 1. In contrast, our proposed method does not compromise the structural information in that hats and hands are retained in the outputs, and neither hairs nor eyes are crushed.

Also, the results show that our proposed method generates outputs with feasible texture. StyleCLIP gives unsatisfactory results that still look like photos. StyleGAN-NADA has difficulty in translating images into pop art style. VQGAN-CLIP and CLIPstyler also fails at generating Pixar and Uiyo-e

style images. On the contrary, we could see that our method provide high fidelity samples which are transferred into the styles of the target prompt.

In addition to the qualitative results, the superiority of our method is proved by quantitative results. In the Table 1, we could see that our proposed method achieves the best scores in both user study and CLIP score. Although StyleCLIP attains the smallest loss in face identity, its style transformation shows inferior performance in CLIP score. This shows that StyleCLIP maintains semantic information too much that it cannot transform the style enough. In contrast, VQGAN-NADA and CLIPstyler modulates the images too much that they even transforms the content. However, our method could achieve both of them. Meanwhile, CLIP score is calculated in a global manner as explained in the equation (7) and in a patch-based manner. Face identity loss is calculated using ArcFace (Deng et al., 2019). The images used for quantitative experiments are the same as those used for the user study.

Figure 6: 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.

#### 4.2.2 Diffusion models

For diffusion-based models, we adopt two methods for comparison - DDIB (Su et al., 2022) and DiffusionCLIP (Kim et al., 2022). Since DDIB requires two different score functions, we trained a new diffusion model on Wikiart dataset in order to evaluate its translation performance between painting and photo domains. For the photo domain, we utilized the pretrained diffusion models described above. The qualitative and quantitative results of the comparison are demonstrated in the Figure 5 and the Table 2.

The third column of the Figure 5 shows that DDIB suffers from identity loss. Though the portrait is translated into photo, its facial identity is destroyed. Also, it is hard to delineate the shape of the church with the output of DDIB. Furthermore, the most critical drawback of DDIB is that diffusion models have to be trained for each new domain. In this regard, image translation from portrait to neon light style is not available. Meanwhile, DiffusionCLIP shows relatively satisfying quality in translating photos into another style. Once the images are not well converted into the photo domain, however, it is inevitable to get unsatisfactory results from unseen domain images. As shown in the first and third rows of the Figure 5, DiffusionCLIP has difficulty translating images from an unseen domain, painting and portrait respectively, resulting in image quality degradation. This is supported by user study results on DiffusionCLIP. As can be seen in the Table 2, its content score in unseen domains is 0.85 lower than the score in photo domain. More examples are provided in the Figure 10 In contrast, we can observe that our proposed method can stylize images not only from photo domain but also unseen domains such as portrait or painting. The portrait is transformed into photo while maintaining its facial identity. Also, the painting of church is translated into neon light style while retaining even small objects like a cross. It is confirmed with user study results. The scores between photo domain and unseen domains are highly similar, which means that our method can modulate images even from unseen domain.

Figure 7: Ablation study on three losses for content guidance - $L_{MSE}$, $L_{VGG}$, $L_{CUT}$. In each row, source images are translated into the style of "golden" and "oil painting of flowers", respectively. Patch-wise contrastive loss helps to preserve content information.

| | MSE | VGG | CUT | MSE, VGG | MSE, VGG, CUT |
|---|---|---|---|---|---|
| **Content** ↑ | 2.80 | 2.05 | <u>3.90</u> | 2.90 | **4.75** |
| **Style** ↑ | 2.05 | 3.10 | <u>3.85</u> | 3.70 | **4.20** |

Table 3: User study results on ablation studies regarding content losses. Bold text and underline refer to the best and the second best scores, respectively.

In terms of computational time, as shown in Figure 3, our method is significantly faster than DiffusionCLIP.

### 4.3 Ablation Studies

In order to check the roles of each guidance loss, we conducted ablation studies. The results are illustrated in the Figure 6 and 7.

**Roles of style losses** Our proposed style loss consists of $L_{global}$, and $L_{directional}$ as shown in the equation (6). We examine the role of each loss by applying the loss functions one by one. In the Figure 6, we evaluate on three styles - green crystal, neon light, and fire. The images in the lower row tend to be more stylized than images in the upper row, which means that directional CLIP loss leads the diffusion models to higher performance in style modulation. In addition, we verified the role of patch-based guidance. The results show that CLIP guidance with whole image tend to stylize the image in local parts. In contrast, the patch-based guidance transforms the image into the given style that covers large area.

**Roles of content losses** As proposed in the equation (11), the content loss for guidance is comprised of three different losses, $L_{CUT}$, $L_{VGG}$, and $L_{MSE}$. Among the three losses, we investigate the power of $L_{CUT}$ by eliminating it from the total content loss. In Figure 7, we can see that $L_{VGG}$ and $L_{MSE}$ are not enough to sufficiently preserve the structural properties. The outlines of the buildings remain, but not their details such as windows. When we employ all the three losses, we could get the best results regarding content preservation. In the upper row of the Figure 7, even letters in the source image are preserved in the output with contrastive content loss. The superiority of $L_{CUT}$ is also justified with user study results. As shown in the table 3, $L_{CUT}$ scores higher than $L_{MSE}$, $L_{VGG}$, and even both of them.

## 5 Conclusion

In this paper, we proposed a strategy to guide diffusion models with content losses. In terms of the guidance strategy, our method possesses several strengths. First, our method does not require additional training on the diffusion models. Moreover, we have shown that combination of content and style losses provide highly exceptional results. The source image can be transformed into another style while preserving its content information. Incorporating patch-wise contrastive loss with diffusion model results in the high capability to maintain content. In addition, our proposed method could modulate even unseen domain images, which is distinct from other text-driven image manipulation methods. The discussions on limitations are given in the Appendix.

## REFERENCES


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*, pp. 18208–18218, 2022.

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. URL `https://openreview.net/forum?id=SlxSY2UZQT`.

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)*, pp. 14367–14376, October 2021.

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*, pp. 12413–12422, 2022.

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.

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.

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*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

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*, pp. 4690–4699, 2019.

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

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*, pp. 12873–12883, 2021.

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.

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*, pp. 2414–2423, 2016.

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

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

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*, pp. 1501–1510, 2017.

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*, pp. 1125–1134, 2017.

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.

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*, pp. 4401–4410, 2019.

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*, pp. 8110–8119, 2020.

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.

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*, pp. 2426–2435, 2022.

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*, pp. 18062–18071, 2022.

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.

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

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)*, pp. 2085–2094, October 2021.

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*, pp. 8748–8763. PMLR, 2021.

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*, pp. 10684–10695, 2022.

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.

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*, pp. 1–10, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022.

Julia Wolleb, Robin Sandkühler, Florentin Bieder, and Philippe C Cattin. The swiss army knife for image-to-image translation: Multi-task diffusion models. *arXiv preprint arXiv:2204.02641*, 2022.

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.

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*, pp. 2223–2232, 2017.

# A DETAILS OF IMPLEMENTATION

## A.1 DATA MANIPULATION AND HYPERPARAMETERS

In order to utilize directional CLIP loss in a patch-based manner (Kwon & Ye, 2022), we randomly crop the source image. The patch size can have various ranges, but we mainly utilize $(0.01, 0.05)$ for texture styles, such as golden or green crystal, and $(0.01, 0.3)$ for artistic styles, such as painting by Gogh or pop art. The cropped images are then augmented with perspective function and random affine transformation.

For style and content guidance, we utilized various weights for each loss according to different styles. Even though the values can be varying for each style, the weights for $L_{global}$ and $L_{directional}$ usually range from 5000 to 30000. In addition, the weights for $L_{CUT}$, $L_{VGG}$, and $L_{MSE}$ are generally $100$, $100$, and $10000$, respectively. However, one can change these values to improve image quality.

## A.2 PATCH-WISE CROSS ENTROPY LOSS FOR CUT GUIDANCE

We explain cross-entropy loss in the equation (9) in more detail (Park et al., 2020). The inputs for the loss are a query $\boldsymbol{v}$, and its positive $\boldsymbol{v}^+$ and negatives $\boldsymbol{v}_i^-$ where $i \in [1, \ldots, N]$ . The query is a patch from the generated image and the positive is the corresponding patch of the source image. The negatives are the other non-corresponding patches of the source image. Then the cross-entropy loss helps a patch share embedding space with the corresponding patch of the input and not with the other patches. It can be written as,

$$\ell(\boldsymbol{v}, \boldsymbol{v}^+, \boldsymbol{v}^-) = -log\left[\frac{e^{\boldsymbol{v}\cdot\boldsymbol{v}^+/\tau}}{e^{\boldsymbol{v}\cdot\boldsymbol{v}^+/\tau} + \Sigma_{i=1}^{N} e^{\boldsymbol{v}\cdot\boldsymbol{v}_i^-/\tau}}\right] \tag{12}$$

where $\tau$ is a temperature

## A.3 TRAINING DDIB

For comparison with DDIB, we trained a diffusion model with 13 thousands of images from Wikiart dataset. They were $256 \times 256$ in size. The architecture of the model is based on the guided diffusion (Dhariwal & Nichol, 2021). The model use $128$ base channels and the attention at $16 \times 16$ and $8 \times 8$ resolutions. Residual blocks for upsampling and downsampling are not used. We fixed the variance as a constant (Ho et al., 2020). The model was trained during 50,000 iterations with batch size $8$ on a NVIDIA RTX 3090.

# B ADDITIONAL EXPERIMENTAL RESULTS

## B.1 EFFECT OF THE NUMBER OF TIMESTEPS

Since the diffusion process usually takes lots of time, two techniques are widely used - respacing and skipping time steps (Chung et al., 2022; Kim et al., 2022). The last time step $T$ is respaced into $T'$. Then we forward the diffusion model to time $t_0 < T'$ and reverse the diffusion process from $x_{t_0}$. $T'$ and $t_0$ have various effects on both image quality and time consumption. As shown in the Figure 8 (a) and (b), image quality with respect to style transformation enhances as respacing time step $T'$ increases. However, its growth rate decreases and its difference is imperceptible even though sampling time still increases. In the mean time, CLIP score decreases as skip time step $t_0$ increases as illustrated in the Figure 8 (c). On the other hand, the content information is not fully preserved in the early time steps $t_0 = 15$ or 20 as shown in the Figure 8 (a). Thus, we set ($T'$, $t_0$) as (50, 25) for our baseline.

## B.2 DIFFUSION MODELS' TRADE-OFF BETWEEN STYLE AND CONTENT

With respect to style transfer, one of the challenges posed by unconditional diffusion models is to maintain content of the given image. When transforming styles of the given image, its content

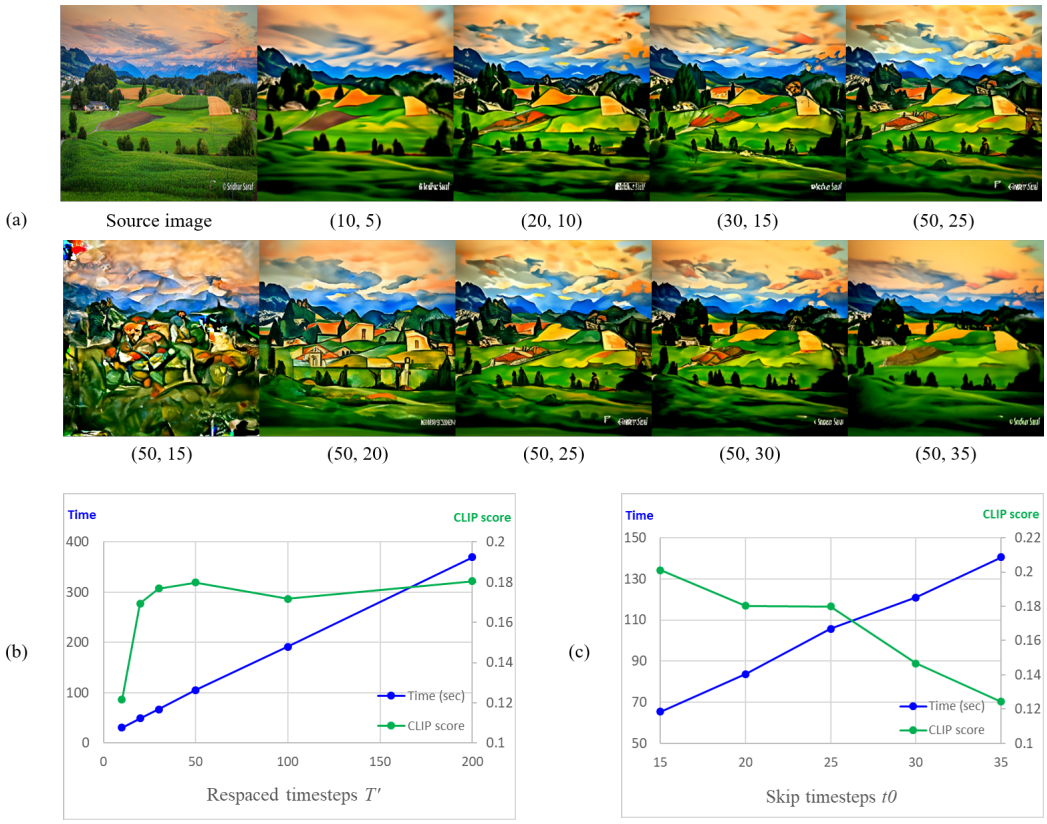


Figure 8: The effect of the respaced time $T'$ and skipped time $t_0$. (a) demonstrates the images sampled with ($T'$,$t_0$). The first row shows the difference between various $T'$ when $t_0$ is its half and the second row shows the difference between various $t_0$ when $T' = 50$. (b) and (c) illustrates the relationship between sampling time and CLIP score as graphs for the first and second rows of (a), respectively.

changes simultaneously. GAN-based methods explicitly impose content losses, such as a reconstruction loss. This results in good performance in content preservation. In contrast, diffusion models have no constraint during training phase. They generate high quality images in correspondence with the training data domain. The semantic constraints are not considered which finally results in the degradation in the quality of the generated images.

Here, we compare four diffusion models - ILVR, DDIM, DiffusionCLIP, and our proposed method - with respect to style and content in the Figure 2. ILVR utilizes down-sampled reference image as condition in each reverse denoising steps. The condition helps the generated image share its content information with the reference image. However, it cannot have same identity because reverse DDPM steps without condition should be given sufficiently in order to generate images in photo style. This accordingly results in a loss of content. DDIM can reconstruct the source image when the variance of noise $\sigma_t$ is set as 0. However, the style is also preserved with zero variance. When we control $\sigma_t$ as larger than zero, we can get photo style images but their content is altered with stochastic noise. DiffusionCLIP tried to solve the trade-off between content and style by fine-tuning the score function $\epsilon_\theta$. However, it requires much more time due to the model training for each style and data preparation. In addition, the content cannot be maintained when the source images are from unseen domains. This is more illustrated below. In contrast, our proposed method does not require additional training on the diffusion model. This leads to shorter time than DiffusionCLIP. With the help of CUT guidance, we could retain the content of source image from any domain and translate it into different styles.

### B.3 UNSEEN DOMAINS

DiffusionCLIP tried to solve the trade-off between content and style by fine-tuning the diffusion model with identity loss. Because of the constraints imposed on the finetuned model, the transformed image shows high performance in identity preservation. However, the fine-tuned model $\hat{\epsilon}_\theta$ converts only the photo domain images. When it comes to unseen domains, such as portraits or paintings, they should be converted to photo images through $\epsilon_\theta$. Since it has not been fine-tuned with identity loss, the semantic information is lost during the reverse sampling process due to the stochastic property of the diffusion model. Thus, the final output from images of unseen domains is degraded in its quality. In contrast, our proposed method can transform even the unseen domain images with only one step. Since our method can preserve the identity with content guidance, the final outputs do not suffer from quality degradation. Also, our method takes about 90 seconds whereas DiffusionCLIP requires about 400 seconds for model fine-tuning and sampling. As described in the Figure 2, DiffusionCLIP requires two steps from portrait to photo to Pixar domains. In this process, the face identity is destroyed. However, the proposed method could preserve the identity while transforming into the style of Pixar.

### B.4 USER STUDY

For quantitative analysis, we conducted user study. For comparison with GAN-based methods, 60 images with four styles have been used in total. The styles involved are "golden", "clay", "3d render in the style of Pixar", and "pop art." We utilized human face images because StyleCLIP and StyleGAN-NADA are based on face dataset. In addition, we totally generated 24 images with six styles for comparison with DiffusionCLIP. We chose three styles ("neon light", "green crystal", and "Ukiyo-e") for the photo domain and the other three styles ("3d render in the style of Pixar", "pop art", and "golden") for unseen domains. We used portraits and paintings from Wikiart dataset for unseen domain images. Besides, for ablation study on content losses, we used 15 images for three styles ("golden", "oil painting of flowers", "leather") in total. The number of questions were 14. 20 users participated in the user study, and their ages range from 20 to 60 years old. They were randomly recruited online.

## C LIMITATIONS

Although our proposed method has various strengths and shows great performance, there remain some limitations. As described in the Appendix A.1, one should find weights for each loss though their relevant ranges are given in this paper. In addition, although diffusion model does not require addition training, a light weight MLP network requires training which doubles time for image generation overall. As a future work, we will try to solve these issues.

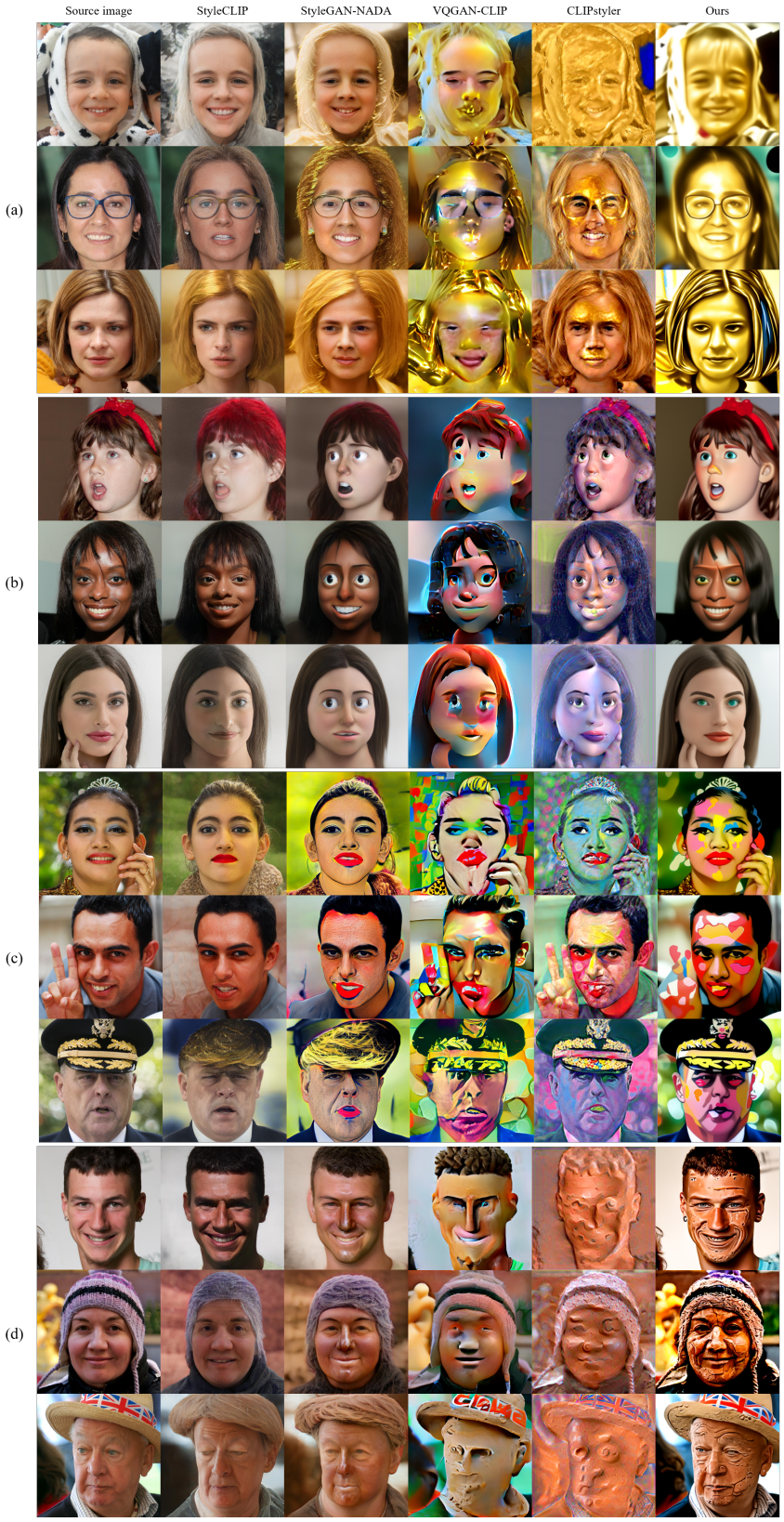


Figure 9: Additional results on the comparative studies. (a), (b), (c), and (d) are the results on styles of “golden”, “3d render in the style of Pixar”, “pop art”, and “clay”, respectively.

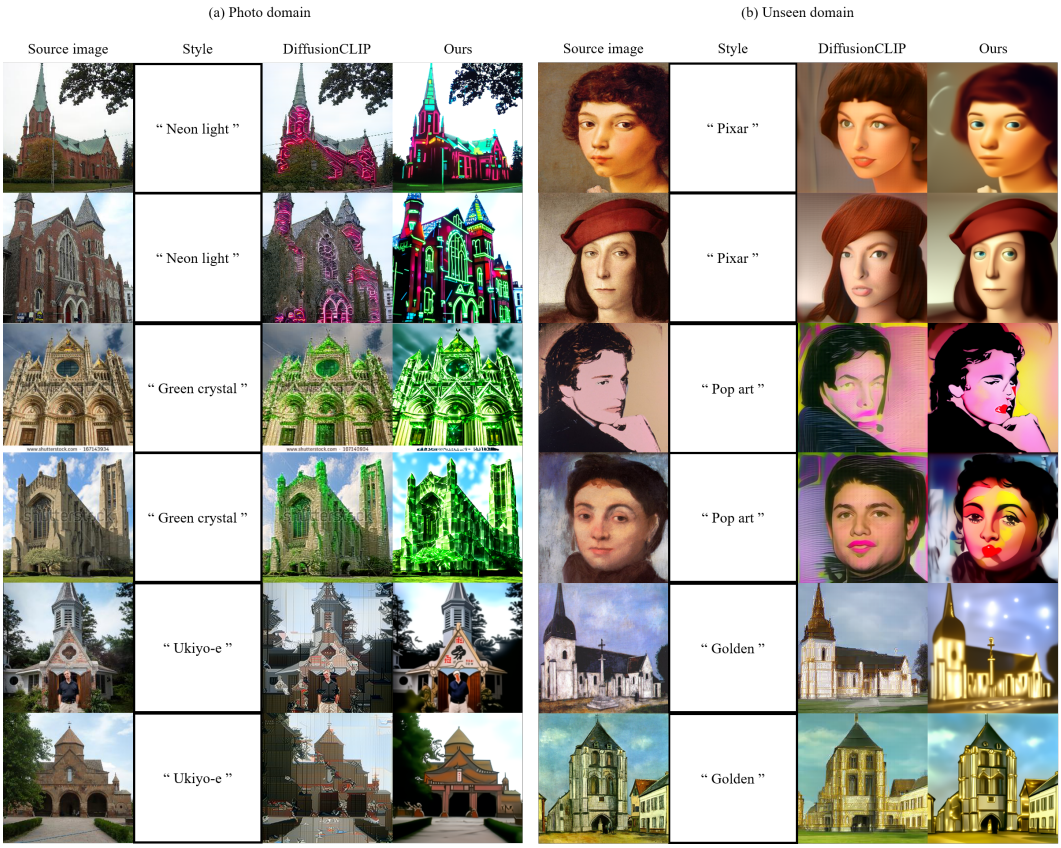


Figure 10: Additional results on the comparative studies. Source images in (a) are from photo domain and ones in (b) are from unseen domains such as portrait or painting.