DUAL-DOMAIN DIFFUSION BASED PROGRESSIVE STYLE RENDERING TOWARDS SEMANTIC STRUCTURE PRESERVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we propose a Dual-Domain Diffusion based Progressive Style Rendering (D3PSR) method to achieve style rendering from the semantic Domain A to the style Domain B. Although some existing models obtain style transfer in Image-to-Image (I2I) translation, little is known about how computers draw images due to the black box problem of classic generative model. Leveraging diffusion models for I2I style transfer, for the first time, we have implemented a progressive method to visualize the intermediate steps in image generation, which provides interpretability for style transfer. As far as we know, we are the first to innovatively use the diffusion model for dual-domain image style transfer. Numerous experimental results and comparisons with state-of-the-art methods in the same field show that our approach has extremely superior performance in stylization and extraordinary preservation of semantic structure.

1 INTRODUCTION

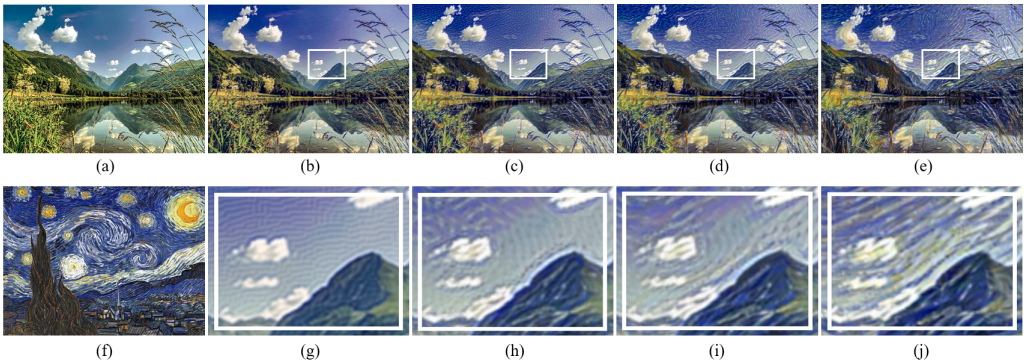

Figure 1: The step-by-step process of our diffusion based progressive style rendering method. (a) and (f) indicates the content image and target style image from Domain A and Domain B, respectively. (b) $\sim$ (e) stand for the output images generated by our progressive rendering model. (g) $\sim$ (j) shows the detail of output images selected by white box.

Since the remarkable work on StyleGAN (Karras et al., 2019), image-to-image (I2I) style transfer has been a matter of huge interest, where a content image can often be rendered into a new artistic style using a referenced image (Deng et al., 2022). Currently, one of the critical challenges in the field of I2I style transfer lies in how to accurately extract essential feature information of images from two different domains (Zhou et al., 2021), and combine them together creatively.

Early work (Efros & Freeman, 2001; Bruckner & Gröller, 2007) can generate stylized images based on texture synthesis. However, the process of texture rendering requires complex computations. The emerged popular solutions in the past few years include convolution neural networks (CNNs) based generative models (Gatys et al., 2016; Gatys et al., 2017) that extract the style and content targets

by the pre-trained network to optimize the joint content and style loss of pending images. Later, various CNN-based style transfer methods were proposed (Kolkin et al., 2019; Wang et al., 2020; Kalischek et al., 2021; Chen et al., 2021; Hong et al., 2021) that are highly efficient for feature extraction in source image. Similarly, Zhu et al. (2017) proposed cycle-GAN to realize the unpaired translation based on generative adversarial nets (Goodfellow et al., 2020). However, none of these works demonstrate a step-by-step progressive process in their image rendering. In this paper, by leveraging diffusion based models, we aim to develop a progressive I2I rendering tool capable of step-by-step rendering while preserving semantic structures of the source image.

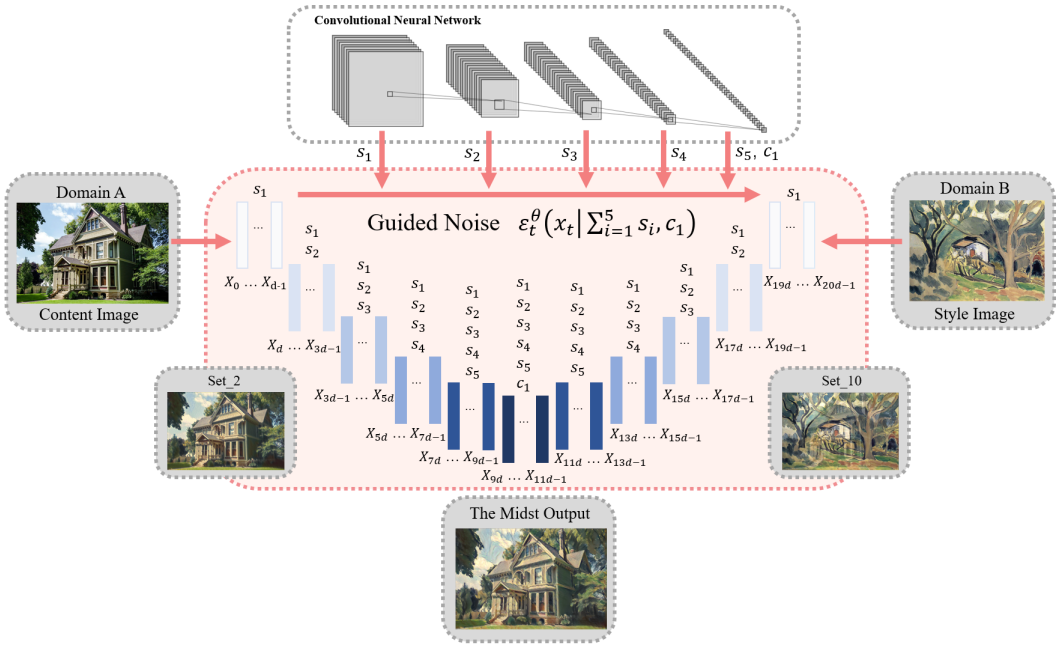

Figure 2: The 'U' shape architecture of our D3PSR method. Content Image denotes the input data from Domain A that contains the semantic structure we need to extract and Style Image denotes the input data from Domain B that contains the styles such as texture of painting strokes we try to imitate. s_1, s_2, s_3, s_4, s_5 denote 5 style feature targets extracted from Domain B image by utilizing the CNN as an encoder. c_1 denotes content feature target extracted from Domain A image. The deeper the set in the 'U' shaped structure the more feature targets will be imposed. The guided noises modify the image generation, which is guided with different number of s_i and c_1 depending on certain set. d controls the size of each set which depends on the total step T of the diffusion model. The final output is obtained at the midst layer on the bottom of the 'U' shape architecture.

Traditional convolutional neural network (CNN)-based generative models (Goodfellow et al., 2020; Kingma & Welling, 2013; Makhzani et al., 2015) achieve high performance at the expense of interpretability, and black boxes have become a fatal flaw in CNNs (Zhang & Zhu, 2018). The diffusion model differs from these algorithms in that it gives us a way to visualise intermediate steps while obtaining the final generated image.

Diffusion model is based on the mathematical framework described in the work (Sohl-Dickstein et al., 2015). Since Ho et al. (2020) provided the foundational work for subsequent diffusion models, numerous works have emerging on this base, such as Denoising Diffusion Implicit Models (DDIMs) (Song et al., 2020), Classifier-Free Guidance Diffusion (CFGD) (Ho & Salimans, 2022), GLIDE (Nichol et al., 2021), OpenAI's DALL E 2 (Ramesh et al., 2022) and Google's Imagen (Saharia et al., 2022). In the past year, many remarkable works have been done. Choi et al. (2021) proposed Iterative Latent Variable Refinement (ILVR) method, by adding reference to the DDPM generation process, they can control the generation and obtain high-quality images at the same time. Kawar et al. (2022) proposed an efficient and unsupervised posteriori sampling method for image restoration, coloring and deblurring. Wolleb et al. (2022) proposed a method to guide the denoising process by an external gradient, which achieves to guide the image generation to the desired out-

put. Su et al. (2022) proposed Dual Diffusion Implicit Bridges (DDIBs), which train two diffusion models independently on each domain and leverage two ordinary differential equations (ODEs) for image translation. However, there is so far no diffusion model based work that can achieve stunning I2I style transfer via the precise rendering of textures and style with semantic structure preservation.

In this work, to address the black box problem of traditional image generation models, we creatively leverage diffusion models to provide interpretability to the image generation process. As shown in Fig. 2, for the first time, we propose a novel Dual-Domain Diffusion based Progressive Style Rendering (D3PSR) method that progressively deepens the style rendering while extraordinary preserving the semantic structure of the source image. Unlike existing diffusion-based methods that use the U-net (Ronneberger et al., 2015) to obtain a noise predictor ε_t^θ, we use a CNN as an encoder to extract style and content feature targets from both domains , and then impose these feature targets to guide the noise $\varepsilon_t^\theta(x_t|\sum_{i=1}^5 s_i, c_1)$. Consequently, we can make controlled stylistic modifications to the generated image, allowing for a style transfer on the original image while retaining the desired semantic structure.

In summary, our substantial contributions includes:

- Creatively proposing the novel D3PSR method which provides the interpretability for I2I translation task by visualizing the intermediate steps.
- For the first time implementing the two-domain I2I style transfer with diffusion model.
- Comprehensive experiment results and detailed comparisons with state-of-the-art methods demonstrating that D3PSR outperforms baseline methods and obtains extraordinary stylization while preserving remarkable semantic structure.

2 Preliminary on Diffusion Models

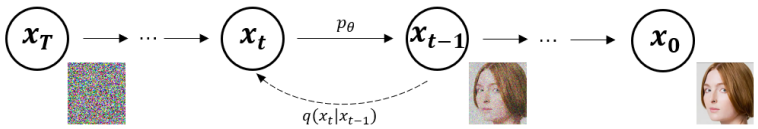

Figure 3: The basic structure of diffusion model

Denoising diffusion probabilistic models (DDPM) (Ho et al., 2020) is a parameterized Markov chain that uses variational inference to generate samples that match the data after a finite time T. Given a data point sampled from the true data distribution $x_0 \sim q(x)$, the forward process incrementally adds noise $\varepsilon_t \sim \mathcal{N}(0, \mathbf{I})$ over $0 \sim$ T time steps, producing a series of noisy samples: $x_1, ..., x_T$. The approximate posterior $q(x_{1:T}|x_0)$ of the diffusion model is a fixed Markov chain, which can be written as follows:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

Adding Gaussian noise to the data is controlled by *variance schedule* $\beta_1, ..., \beta_T$. When β_t is small enough, the $q(x_{t-1}|x_t)$ of the reverse process also abides Gaussian distribution. Therefore, $q(x_t|x_{t-1})$ can be written as:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{2}$$

The derivation process uses the reparameterization trick proposed in VAE works (Kingma & Welling, 2013; Rezende et al., 2014), given $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{i=1}^{T} \alpha_i$, x_t can be expressed as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\bar{\varepsilon}_t\text{: where } \bar{\varepsilon}_t \sim \mathcal{N}(0, \mathbf{I}) \tag{3}$$

Then, sampling of x_t can be written as:

$$q(x_t|x_0) := \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{4}$$

Because the distribution of the whole dataset $q(x_{t-1}|x_t)$ is intractable, the original work (Ho et al., 2020) approximates this conditional probability with p_θ for the reverse diffusion process: $p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$, where θ denotes the learning parameters. The loss function can be expressed as follows:

$$\mathcal{L}_t^{simple} = \mathbb{E}_{x_0,\varepsilon}[\|\varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, t)\|^2] \tag{5}$$

where ε_θ is a set of T function with trainable parameters $\theta(t)$. Figure 3 shows the structure of classic diffusion model, where an image can be generated from noises or vice versa.

3 The Proposed Diffusion based Progressive Style Rendering

In this section, we present our Dual-Domain Diffusion based Progressive Style Rendering (D3PSR) method. The architecture of our model is shown in Figure 2. The traditional diffusion model is inspired by the non-equilibrium thermodynamic entropy increase, where noise is continuously added to a given input data in the forward process to finally obtain an isotropic Gaussian noise. In principle, such kind of model can only handle single-domain image problems, since in the forward process, sampling of x_t can be expressed as:

$$q(x_t|x_0) := \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{6}$$

which means the generated data is only sampled from one domain. Even if labels or conditions y_{target} are imposed during the training process, the required crucial information between two domains cannot be efficiently and flexibly combined such as Choi et al. (2021). In order to efficiently integrate the feature between two domains, we innovatively propose a double-end input 'U' shape structure.

Existing diffusion-based methods leverage the U-net (Ronneberger et al., 2015) to obtain a noise predictor ε_t^θ, we creatively utilize a CNN as an encoder to extract style and content feature targets $(s_1, s_2, s_3, s_4, s_5, c_1)$ from two domains, and then impose these feature targets to guide the noise $\varepsilon_t^\theta(x_t|\sum_{i=1}^{5} s_i, c_1)$ (the number i of the imposed feature targets depends on the certain set). Since we can only extract 6 feature targets ($6 \ll T$) on different convolution layers of CNN and aim to visualize the intermediate steps, we divide the time step T into 11 sets and impose different number of feature targets on each set to obtain a progressive visualization of the output changing. In the U-shaped structure, the closer of the set to the bottom, the more feature targets will be imposed. Thus, we need to have 11 sets to ensure the gradually style rendering effect, so that we can impose $(1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1)$ feature targets on each set by order.

Given data x_A sampled from the true data distribution $p_d(X_A)$ from Domain A ($\mathcal{X}_A$), we continuously add noise $\varepsilon_t \sim \mathcal{N}(0, \mathbf{I})$ over $0 \sim$ T time steps. $x_1, ..., x_t$ denote the samples in each step. On the basis of diffusion model as illustrated in section 2 and Eq. (6), we can obtain the sampling of x_{last} as $q(x_{last}|x_A)$, where x_{last} denotes the data sampled from the last step of the our model with input data x_A. Given x_B sampled from the true data distribution $p_d(X_B)$ from Domain B ($\mathcal{X}_B$), we can obtain the Kullback-Leibler Divergence (Kullback, 1997) between the true target distribution and the data distribution generate by our model:

$$\mathcal{L}_B = \mathcal{KL}[p_d(x_B)||q(x_{last}|x_A)] \tag{7}$$

which guides the diffusion model to rendering the data from $\mathcal{X}_A \to \mathcal{X}_B$.

Let v be the convolutional layers codes vector extracted by utilizing a convolution neural network (CNN) as encoder and $p_\theta(v)$ be the prior distribution we want to impose on the codes with CNN parameters θ, conditional style and content targets codes distribution can be obtained as $p_\theta(v_c|x_A)$ and $p_\theta(v_s|x_B)$. Then the aggregated distribution of $p_\theta(v_c)$ and $p_\theta(v_s)$ can be obtained as $p_\theta(v_c) = \int p_\theta(v_c|x_A)p_\theta(x_A)dx_A$ and $p_\theta(v_s) = \int p_\theta(v_s|x_B)p_\theta(x_B)dx_B$ that represent the content and style targets codes distribution, respectively.

Given x_t sampled from the intermediate steps $x_t \sim q(x_t|x_A)$ in diffusion model and $p_\theta(v_c|x_t)$, $p_\theta(v_s|x_t)$ be the encoding distributions. Then the aggregated distribution of content and style codes of sampled data can be written as:

$$q_\theta(v_c) = \int\int p_\theta(v_c|x_t)q(x_t|x_A)p_\theta(x_A)dx_t dx_A \tag{8}$$

$$q_\theta(v_s) = \int\int p_\theta(v_s|x_t)q(x_t|x_B)p_\theta(x_B)dx_t dx_B \tag{9}$$

To combine the dual-domain feature, we joint the content and style targets codes losses then obtain optimal dual-domain targets codes loss function with Kullback-Leibler divergence (Kullback, 1997) as follows:

$$\begin{aligned}\mathcal{L}_{dual} &= \mathcal{KL}[p_\theta(v_c)||q_\theta(v_c)] + \mathcal{KL}[p_\theta(v_s)||q_\theta(v_s)] \\ &= \mu\mathcal{L}_{content} + \lambda\mathcal{L}_{style}\end{aligned} \tag{10}$$

where μ, λ denotes the weight coefficients for style and content codes vector loss respectively, with which the progressive stylized images can be more rationally and optimally controlled. The first term represents the texture and style loss compared with targets from $\mathcal{X}_A$ and the second term represents the content loss compare with the target from $\mathcal{X}_B$.

As shown in Figure 2, we separate the whole steps into 11 data sets ($set\_1 - set\_11$). Since we expect to obtain the progressive style transfer of different step and observe the style rendering pattern of neural network, we impose different numbers of style and content targets at different step.

As we have the diffusion sampling $q(x_t|x_A)$, we sample the data x_{mid} from the midst layer of our model, then the midst data distribution can be obtained from $q(x_{mid}|x_A)$. In order to make the content preservation of the midst output results adjustable, we introduce the parameter ϕ defined as the Control Factor (CF) of semantic structure from $\mathcal{X}_A$. Then we can obtain the optimal function for the midst data distribution loss based on the true data distribution $p_d(x_A)$ from $\mathcal{X}_A$ as follows:

$$\mathcal{L}_A = \mathcal{KL}[p_d(x_A)||q(x_{mid}|x_A)] \tag{11}$$

Ultimately, with Eq. (7), Eq. (10) and Eq. (11) we can obtain the final object function for our D3PSR models, as shown below:

$$\mathcal{L}_{total} = \mathcal{L}_{dual} + \phi\mathcal{L}_A + \mathcal{L}_B \tag{12}$$

where the fist term represents the convolutional layers feature loss between [$v_c \sim p_\theta(v_c|x_t)$, $v_s \sim p_\theta(v_s|x_t)$] (with $x_t \sim q(x_t|x_A)$) and [$v_c \sim p_\theta(v_c|x_A)$, $v_s \sim p_\theta(v_s|x_B)$] (with $x_A \sim p_d(X_A)$, $x_B \sim p_d(X_B)$) from dual-domain extracted by our model, the second and third term denotes the Kullback-Leibler divergence (Kullback, 1997) between the true data sampled from true distribution representing $\mathcal{X}_A$ & $\mathcal{X}_B$ and the data sampled from midst step of diffusion model. In experiments, we can optimize the midst output by adjusting the hyperparameters μ, λ, ϕ to be as close to the target as possible on an artistic level ($\mathcal{X}_B$) while preserving the semantic structure ($\mathcal{X}_A$). Meanwhile, we can also modify these three hyperparameters to control each intermediate step, so as to make the whole style rendering process in a visibly progressive way.

4 EXPERIMENTS

In this section, we will give details of our experimental setup designs as well as our experimental results. We selected a collection of classic artworks, set them as Domain B, used them to train our D3PSR model to extract their style targets s_1, s_2, s_3, s_4, s_5. A collection of nature landscape pictures is set as Domain A, processed by our model to obtain content target c_1. As shown in Figure 2, we use d to control the size of each set, in this experiment, we set $d = 50$. To better demonstrate the performance of our model, MS-COCO (Lin et al., 2014) was used as the content dataset and WikiArt (Phillips & Mackintosh, 2011) is used as the style dataset in section 4.1.

4.1 COMPARISON AGAINST OTHER STATE-OF-THE-ART METHODS

We compared our approach with StyTr2 (Deng et al., 2022), StyleFormer (Wu et al., 2021), AdaAttN (Liu et al., 2021), SANet (Park & Lee, 2019) and AdaIN (Huang & Belongie, 2017), as shown in Figure 4. We use the Unifying Structure and Texture Similarity proposed by Ding et al. as the basis for our analysis, which is a method for analysing the structural and textural similarity between images. Compared to this method, Mean Absolute Error (MAE), Multi-Scale Structural Similarity (MS-SSIM) (Wang et al., 2003) and some other metrics such as SSIM (Wang et al., 2004) are relatively inaccurate and perform weakly in analysing texture similarity between images, since they

Table 1: Quantitative comparison of content and style DISTS value with other approaches

	Ours	**StyTr^2**	**StyleFormer**	**AdaAttN**	**SANet**	**AdaIN**
sample_1						
D_S	**0.2643**^{1}	**0.2685**^{2}	0.3574	0.3244	0.2709	0.2871
D_C	**0.4251**^{1}	0.4384	0.4745	0.4423	0.4569	**0.4308**^{2}
sum	**0.6894**^{1}	**0.7069**^{2}	0.8319	0.7667	0.7278	0.7179
sample_2						
D_S	**0.2482**^{1}	**0.2803**^{2}	0.3492	0.3832	0.2879	0.3327
D_C	0.4522	0.4531	0.5285	**0.3614**^{1}	0.4864	**0.4442**^{2}
sum	**0.7004**^{1}	**0.7334**^{2}	0.8777	0.7446	0.7743	0.7769
sample_3						
D_S	**0.3013**^{2}	0.3121	0.3160	0.3016	**0.2938**^{2}	0.3394
D_C	**0.2780**^{1}	0.3311	**0.3310**^{2}	0.3293	0.3510	0.3686
sum	**0.5793**^{1}	0.6432	0.6470	**0.6309**^{2}	0.6448	0.7080
sample_4						
D_S	0.4009	**0.2969**^{2}	0.3222	0.3415	**0.2857**^{1}	0.3389
D_C	**0.3902**^{1}	0.4940	0.5674	**0.4672**^{2}	0.5214	0.4848
sum	**0.7911**^{2}	**0.7909**^{1}	0.8896	0.8087	0.8071	0.8237

rely on a simple introjection mapping and tends to make more conservative estimates which produces a superposition of all possible results.

The Table 1 shows the DISTS value (Ding et al., 2020) of the output images to target images after style transfer by approaching different methods, where D_S, D_C indicates the distance to content and style images from two domains. In the table, the first and second ranked scores are **bolded** and superscripted with ^{1} and ^{2}, respectively. From the quantitative result in the table 2, we can see that our model is reliably in first or second place. Notably, S_C shows that our model performs remarkably effectively in terms of preserving semantic content and is basically in the first place, which means that our model can obtain efficient stylistic rendering of the images while preserving valid content features from the images in Domain A, which means that our model is particularly distinguished in terms of its semantic recognizability performance.

Figure 4: The comparison of our model against several famous I2I style rendering approaches.

Further, the distinctive feature of our model compared to other methods is the conditioning mechanism which enable the liberty to modify the degree of texture rendering over semantic structure

Table 2: Quantitative DISTS value with different set

	set_1	set_2	set_3	set_4	set_5	set_6
sample_1						
D_S	0.3892	0.3671	0.3438	0.3444	0.3323	0.4309
D_C	0.3673	0.3968	0.4043	0.3928	0.3939	0.3325
sample_2						
D_S	0.3788	0.3312	0.3059	0.2654	0.2637	0.2817
D_C	0.2016	0.2437	0.2674	0.3108	0.3109	0.2810
sample_3						
D_S	0.4041	0.3571	0.3206	0.3101	0.3086	0.3089
D_C	0.2989	0.3573	0.3952	0.4123	0.4039	0.3889
sample_4						
D_S	0.4242	0.3661	0.3716	0.3607	0.3597	0.3663
D_C	0.3433	0.4333	0.4748	0.4662	0.4716	0.4492
sample_5						
D_S	0.3637	0.3165	0.2914	0.2851	0.2892	0.3561
D_C	0.3673	0.3968	0.4043	0.3928	0.3939	0.3325
Average						
D_S	0.3920	0.3476	0.3267	0.3131	0.3107	0.3488
D_C	0.3279	0.3739	0.3977	0.4046	0.4065	0.3601

preservation. Our model is allowed to freely obtain various extent of stylized results by modifying the Control Factor (CF) ϕ to control the trade off between structure and texture. We illustrate the effect of ϕ on style transfer in detail in Appendix B.

4.2 Progressive Style Transfer Evaluation

Here, we used style images and content images from two domains to evaluate the progressive style rendering performance of our model. We selected the samples from $set\_1 - 6$ as examples shown in Figure 5. These progressive style rendering samples demonstrate how our D3PSR model paints step by step.

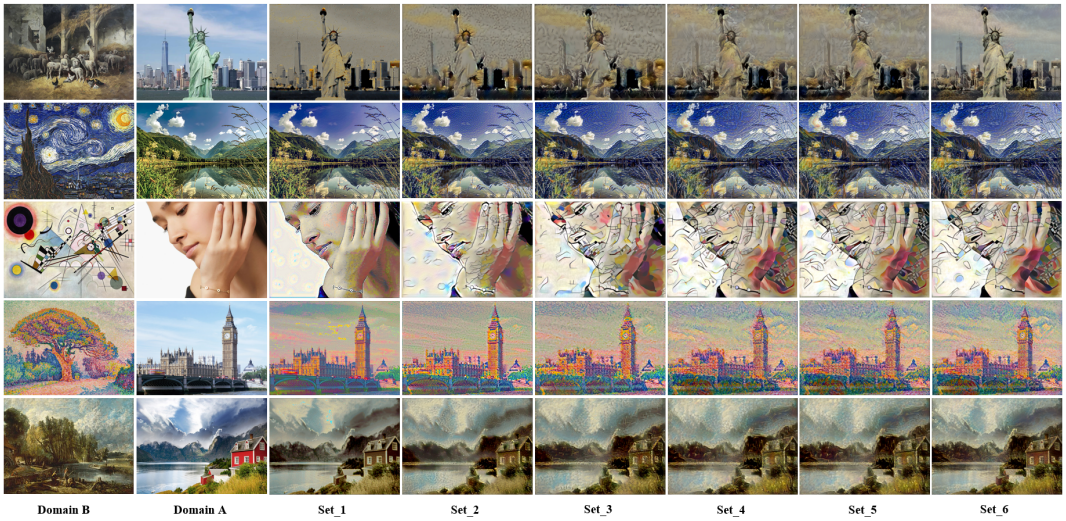

Figure 5: The progressive changes at each step in the style rendering process of our D3PSR model.

Qualitative evaluation. As we can observe from Figure 5, from set_1 to set_5 as more and more style feature targets were imposed, the stylization in the images became more apparent (e.g. brush strokes and textures), but also the semantic structure from the original content images became more

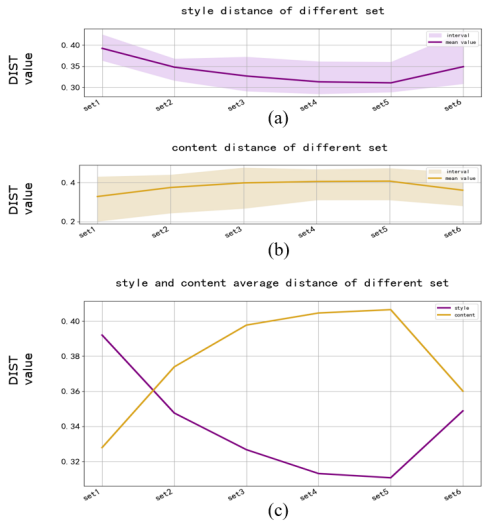

Figure 6: The DISTS value varies over different sets.

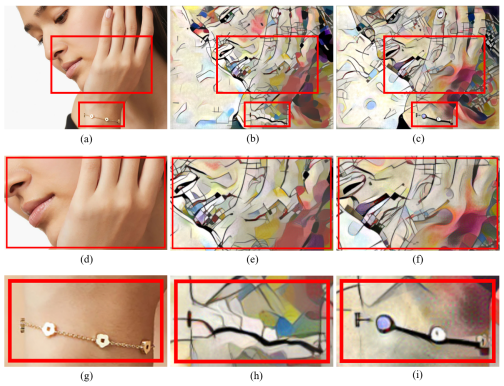

Figure 7: The effect of low/high level feature targets on structure during progressive rendering.

Table 3: Quantitative comparison of content and style DISTS value with other approaches

	Ours	**StyTr**^{2}	**AdaAttN**	**SANet**	**AdaIN**
General	**4.16**^{1}	**4.08**^{2}	3.69	3.03	3.13
Texture	**4.01**^{1}	**3.98**^{2}	3.16	3.56	3.28
Structure	**4.34**^{1}	4.09	**4.16**^{2}	3.03	3.41

blurred. However, with the imposition of ϕ and content target features, the semantic structure in set_6 becomes much sharper than in the previous output.

Quantitative evaluation. From the Table 2, it can be observed that as more style feature targets $(s_1, s_2, s_3, s_4, s_5)$ are progressively imposed on each set, the style DISTS value (Ding et al., 2020) is getting smaller which indicates that the generated images increasingly closely resemble the target style images in terms of texture and style. In contrast, the content DISTS value is increasing which indicates that the generated images lose more and more information about their semantic structure as they are stylised, which also results in a gradual blurring of contours from the content image. However, with the guidance of the content feature target (c_1) and Control Factor ϕ in set_6, the content DISTS value plummets, meaning that the image becomes more similar to the content image from Domain A in terms of semantic structure (the generated images in Figure 5 indicate that the semantic structure in set_6 becomes sharper). At the same time, the style DISTS value becomes larger, demonstrates that there is a drop in performance at the stylised level compared to the previous set.

From the above analysis and Figure 6 we can reveal that in the process of style transfer there is a trade-off between preserving the semantic structure and stylization, which means that it is difficult to preserve the structure and maximise the style transfer at the same time. This is where the superiority of our model emerges, as we can have the freedom to control the extent to which the semantic structure is preserved through Control Fctor ϕ. For example, we can select results from set_6 when we require the output image with clearer content and less style; we can select results from set_5 when we need strong stylization and do not require much sharp semantic structure.

4.3 User Study

To further evaluate the performance of our method, we conduct a user study with StyTr2 (Deng et al., 2022), AdaAttN (Liu et al., 2021), SANet (Park & Lee, 2019) and AdaIN (Huang & Belongie, 2017) as baselines. We set up our questionnaire with reference to the work (Li & Chen, 2009), the

questionnaire format with specific options definition is described in Appendix D. The options in the questionnaire have: A. General: how well the image is stylized; B. Texture: how well the texture strokes are imitated; C. Structure: whether the image preserves its content structure after the stylization. There are total 117 participants involved in our user study, including 56 males and 61 females, with random ethnic sampling. The professional background of the participants covers art workers, computer science researchers and the general public. We utilized the images for comparison in the section 4.1 and 20 images were randomly selected for each participant to rate. The minimum usage time for each marking is 20 seconds. Finally we counted the results obtained and recorded them in Table 3. In the table, the first and second ranked scores are **bolded** and superscripted with ^{1} and ^{2}, respectively. From the table 3 we can clearly observe that our method outperforms other methods in terms of overall style rendering, learning of stylised brushstroke textures, and semantic structure preservation. In particular, the preservation of the semantic structure is extremely impressive.

5 DISCUSSION

A major benefit of our model is its nature to preserve semantic structures in its style rendering process. So far, all existing diffusion model utilized the U-Net to obtain the noise predictor ε_t^θ. Inspired by the work (Gatys et al., 2016), we utilized CNN as an encoder to extract the style and content features $(s_1, s_2, s_3, s_4, s_5, c_1)$ from target style image and semantic structure image to guidance the noise $\varepsilon_t^\theta(x_t|\sum_{i=1}^5 s_i, c_1)$ added to each step, and the number of features imposed depends on the set index. With this method, our model performs well in simulating the style texture of target style images while preserving semantic structure of original content image.

As shown in Figure 7, we can clearly notice that the outline of the hand and human face is becoming sharper with the imposing of more feature targets extracted by the CNN. In the original image, the woman wears a bracelet on her wrist. Since during learning, set_3 does not joint the target c_1 or control factor (CF) ϕ which adjusts the sharpness of the original semantic structure in the output image and is more specifically defined in Methods section, the bracelet does not appear in the rendered result. Also, because set_3 is only imposed with the lower-level vector of CNN, it is more biased to simulate the local small structure texture from Domain B, which results in a cluster of small structures near the nose and mouse. In contrast, set_6 adds more high-level features (s_4, s_5) and imposed with the content target (c_1); moreover, set_6 is imposed with CF ϕ, which controls the extent of the semantic structure. Consequently, we can see that the bracelet and the hand curve, two semantic structures in the source image, are nicely preserved in their structures in the final rendering results in set_6, which demonstrates the superior performance of our method.

6 CONCLUSION

In this work, we proposed a novel Dual-Domain Diffusion based Progressive Style Rendering (D3PSR) method for I2I style transfer. Utilizing the feature of diffusion model that can visualize each step and possess remarkable performance in image generation, our D3PSR method provides interpretability for the neural network in image generation, allowing the model to imitate human artists in drawing step by step to a certain extent, providing substantial contribution to I2I style transfer that was treated mostly as a black box in previous methods. To the best of our knowledge, our work is the first one to implement two-domain style transfer with diffusion model, achieving remarkable stylization while extraordinarily preserving the semantic structure of the source image. Numerous experimental results and comparisons with state-of-the-art method, further demonstrate the superiority of our method in the field of I2I style transfer.

REFERENCES

Stefan Bruckner and M Eduard Gröller. Style transfer functions for illustrative volume rendering. In *Computer Graphics Forum*, volume 26, pp. 715–724. Wiley Online Library, 2007.

Haibo Chen, Lei Zhao, Huiming Zhang, Zhizhong Wang, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Diverse image style transfer via invertible cross-space mapping. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14860–14869. IEEE Computer Society, 2021.

Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

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*, pp. 11326–11336, 2022.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 341–346, 2001.

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.

Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3985–3993, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Kibeom Hong, Seogkyu Jeon, Huan Yang, Jianlong Fu, and Hyeran Byun. Domain-aware universal style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14609–14617, 2021.

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.

Nikolai Kalischek, Jan D Wegner, and Konrad Schindler. In the light of feature distributions: moment matching for neural style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9382–9391, 2021.

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.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10051–10060, 2019.

Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

Congcong Li and Tsuhan Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):236–252, 2009. doi: 10.1109/JSTSP.2009.2015077.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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*, pp. 6649–6658, 2021.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

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*, pp. 5880–5888, 2019.

Fred Phillips and Brandy Mackintosh. Wiki art gallery, inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3):593–608, 2011.

Yuchen Pu, Weiyao Wang, Ricardo Henao, Liqun Chen, Zhe Gan, Chunyuan Li, and Lawrence Carin. Adversarial symmetric variational autoencoder. *Advances in neural information processing systems*, 30, 2017.

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.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

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.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

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.

Wenjing Wang, Shuai Yang, Jizheng Xu, and Jiaying Liu. Consistent video style transfer via relaxation and regularization. *IEEE Transactions on Image Processing*, 29:9125–9139, 2020.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

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.

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*, pp. 14618–14627, 2021.

Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018.

Wujie Zhou, Qinling Guo, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Irfr-net: Interactive recursive feature-reshaping network for detecting salient objects in rgb-d images. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2021. doi: 10.1109/TNNLS.2021.3105484.

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 NON-MARKOVIAN FORWARD PROCESS

In this section, we give a more detailed mathematical basis for our model. Different from the classic DDPM Ho et al. (2020) which is a parameterized Markov Chain (as shown in Figure 3), our Dual-Domain Diffusion based Progressive Style Rendering (D3PSR) method leverages input data from dual-domain (as shown in Figure 2), which means any step in a finite time (T) sequence is no longer just related to the previous step, but also to the future step.

Given data x_A, x_B sampled from the true data distribution $p_d(X_A)$, $p_d(X_B)$ from dual-domain $\mathcal{X}_A$, $\mathcal{X}_B$ respectively, we continuously add noise $\varepsilon_t \sim \mathcal{N}(0, \mathbf{I})$ starting from x_A over $0 \sim$ T time steps. $x_1, ..., x_t$ denote the samples in each step. Unlike DDPM Ho et al. (2020) which ends up with an isotropic Gaussian distribution, our goal is to generate data that approximate the target data distribution sampled from $\mathcal{X}_B$ in the last step.

Based on the discussion in section 2 and section 3, we can obtain the sampling of x_{last} as $q(x_{last}|x_A)$.Then the Kullback-Leibler Divergence (Kullback, 1997) between the true target distribution and the data distribution generated by our model can be written as:

$$\begin{aligned}\mathcal{L}_B &= \mathcal{KL}[p_d(x_B)||q(x_{last}|x_A)] \\ &= \sum_{x\in X_A, x\in X_B}\left[p_d(x_B)log\frac{p_d(x_B)}{q(x_{last}|x_A)}\right] \\ &= \mathbb{E}_{x\sim p_d(x_B), x\sim p_d(x_A)}\left[log\frac{p_d(x_B)}{q(x_{last}|x_A)}\right]\end{aligned} \tag{13}$$

which guides the model to reconstruct the data from $\mathcal{X}_A \to \mathcal{X}_B$.

In order to impose condition on the intermediate steps, we need to extract the target code vector v. As discussed in section 3, the aggregated distribution of $p_\theta(v_c)$ and $p_\theta(v_s)$ can be obtained as follows:

$$p_\theta(v_c) = \int p(v_c|x_A)p_\theta(x_A)dx_A \tag{14}$$

$$p_\theta(v_s) = \int p(v_s|x_B)p_\theta(x_B)dx_B \tag{15}$$

that denotes the content and style target codes vector distribution, respectively.

Here, we discuss the relation of our target codes vector v with the latent code z in VAE (Kingma & Welling, 2013; Rezende et al., 2014). In original work of VAE, consider an observed data sample x, modeled as being drawn from $p_{vae}(x|z)$ with latent code z. $p_{vae}(z)$ denotes the prior distribution of latent space, which follows isotropic distribution in general. $p_{vae}(z|x)$ denotes the posterior distribution on the latent code, which is intractable. In original work, Kingma & Welling (2013) let this variational approximate posterior be a multivariate Gaussian:

$$logq(z|x^{(i)}) = log\mathcal{N}(z; \mu^{(i)}, \sigma^{2(i)}\mathbb{I}) \tag{16}$$

Then VAE utilize the optimal function as follows:

$$\begin{aligned}\mathcal{L}_{vae} &= \mathcal{KL}[q_{vae}(x,z)||p_{vae}(x,z)] \\ &= \mathbb{E}_{x\sim q_{vea}(x)}\Big[-\int q_{vae}(z|x)logp(x|z)dz + \mathcal{KL}[q_{vae}(z|x)||p_{vae}(z)]\Big] \\ &= \mathbb{E}_{x\sim q_{vea}(x)}\Big[-logp_{vae}(x|z) + \mathcal{KL}[q_{vae}(z|x)||p_{vae}(z)]\Big]\end{aligned} \tag{17}$$

However, Pu et al. (2017) proposed that the above optimization approach has certain errors due to the cumulative posterior on latent codes: $\int p_{vae}(z|x)q_{vae}(x)dx \approx \int q_{vae}(z|x)q_{vae}(x)dx = q_{vae}(z)$ is

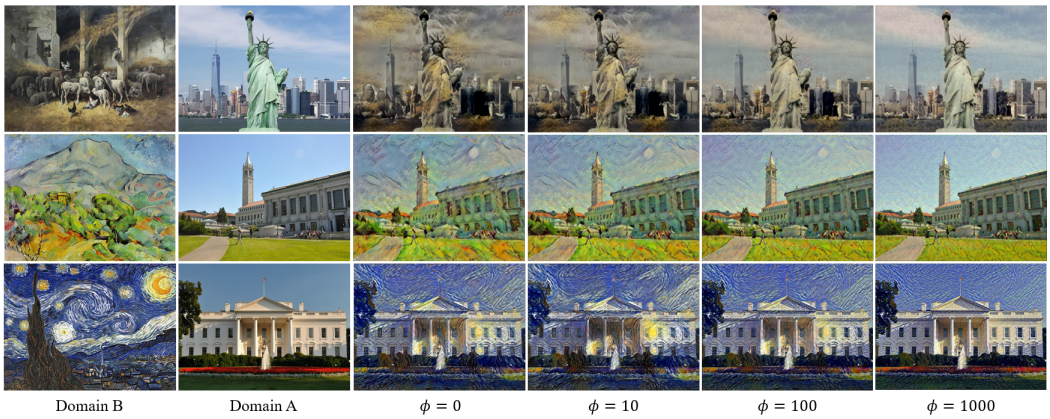

Figure 8: The outputs with different control factor ϕ.

usually different from $p_{vae}(z)$. Since $q_{vae}(z|x)$ is a multivariate Gaussian, there will be a certain error when learning complex data samples, and as the data set increases, the error will become cumulatively larger.

In our model, convolution neural networks (CNNs) utilized as an encoder like VAE to extract the target code vector. Different from the generation method of VAE, the data generated by diffusion model is incrementally adding noise. With more learning time and longer steps in the diffusion model, the performance of diffusion model to fit the observed distribution by superimposing noise can be better than multivariate Gaussian method of VAE. Which means that the diffusion sampling process (shown as follows) could be more accurate.

$$q(x_t|x_0) := \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{18}$$

In our model, to realize the progressive style rendering with structure preservation, we impose the target codes vector v that obtained with Eq. (14) and Eq. (15) on the intermediate steps. Given $x_t \sim q(x_t|x_A)$ sampled from the intermediate steps in diffusion model and $p_\theta(v_c|x_t)$, $p_\theta(v_s|x_t)$ be the encoding distributions. Then the aggregated distribution of content and style codes of sampled data can be written as:

$$q_\theta(v_c) = \int\int p_\theta(v_c|x_t)q(v_c|x_A)p_\theta(x_A)dx_t dx_A \tag{19}$$

$$q_\theta(v_s) = \int\int p_\theta(v_s|x_t)q(v_s|x_B)p_\theta(x_B)dx_t dx_B \tag{20}$$

Different from the VAE (Kingma & Welling, 2013; Rezende et al.), whose goal is to reconstruct data x by the multivariate Gaussian distribution $q_{vae}(z|x)$, we utilize diffusion model to generate data samples with condition v. The optimal function of dual-domain target codes loss has been expressed in Eq. (10)

B CONTROL FACTOR

Further, the distinctive feature of our model compared to other methods is the conditioning mechanism which enable the liberty to modify the degree of texture rendering over semantic structure preservation. Our model is allowed to freely obtain various extent of stylized results by modifying the Control Factor (CF) ϕ to control the trade-off between structure and texture as shown in Figure 8.

Qualitative evaluation. By modifying the ϕ value, the semantic structure and level of stylisation in the generated image is adjusted. Figure 8 shows that as the ϕ value increases, the semantic structure of the image becomes sharper (e.g. windows and doors of buildings, outlines of statues, etc.), while at the same time the extent of stylisation decreases (e.g. brush strokes and textures in the image,

Table 4: DISTS value with different ϕ

	$\phi = 0$	$\phi = 10$	$\phi = 100$	$\phi = 1000$
sample_1				
D_S	0.2825	0.3161	0.4179	0.4613
D_C	0.4574	0.4286	0.3544	0.2994
sample_2				
D_S	0.2760	0.2859	0.3111	0.3424
D_C	0.3859	0.3630	0.3312	0.2822
sample_3				
D_S	0.2521	0.2617	0.2723	0.2986
D_C	0.4098	0.3969	0.3841	0.3558

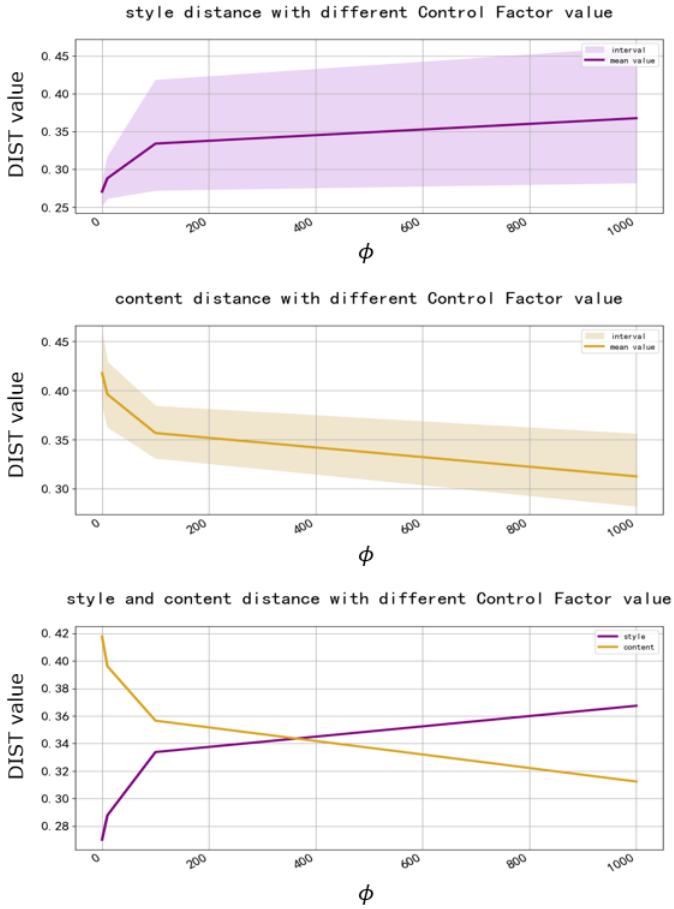

Figure 9: The DIST value with different control factor ϕ.

etc.). It is noticeable that we cannot aggressively reduce the ϕ value in pursuit of a stronger stylistic transition, as too strong a stylisation can cause some areas of the image to overflow (e.g. when ϕ = 1 and 10 in sample 1 and 2). Similarly, the ϕ value should not be increased too large in an attempt to obtain a sharper semantic structure, as this would make the generated image insufficiently stylised (e.g. the stylised strokes and textures from Domain B are weak at ϕ = 1000 in samples 1 and 3).

Quantitative evaluation. Table 4 records the DISTS value Ding et al. (2020) of generated images in Figure 8 in compare with the target style image and original content image, then can obtain the line graph as shonw in Figure 9. From the quantitative analysis we can observe a clear trade-off between style DIST and content DIST, which means that stylisation enhancement is accompanied by a loss

of semantic structural information, with the sharper the semantic structure leading to a weakening of stylisation.

C UNPAIRED IMAGE-TO-IMAGE TRANSLATION

We evaluated our model to using a set of classic art paintings as Domain B, from which we extracted textures and styles to render the contents of different nature landscape pictures in Domain A. Figure 10 shows the fantastic results from our D3PSR model.

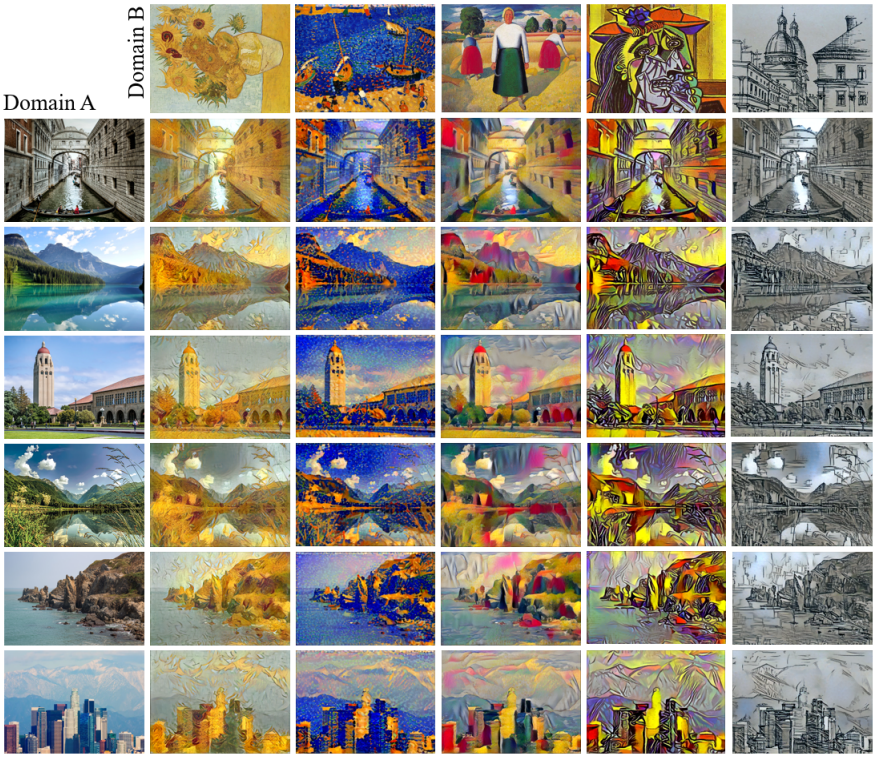

Figure 10: The rendering results over several typical artistic styles from our D3PSR model.

In Figure 10, we show the results obtained by rendering different nature scenes using different painting styles. The resulting images are taken from set_6 from the midst of the 'U' shape structure. The styles used are, in chronological order from left to right, Sunflower (Van Gogh) representing Post-Impressionism, Boats at Collioure (André Derain) representing Fauvism/Neo-Impressionism, Reapers (Kazimir Severinovich Malevich) representing Neo-Suprematism, The Weeping Woman (Pablo Ruiz Picasso) representing Cubism and architecture pen art painting from Rumeysa Şahin.

From the results we can observe that the output images consist of both the style and texture from Domain B while preserving the semantic contents in Domain A, retaining the contour features of the large structures inheriting from Domain A. Moreover, as shown in Figure 11, the clouds in the black frame maintain the the semantic structure of clouds after the stylistic rendering. Meanwhile, the output image inherits the dot painting characteristics of Domain B (shown by the white circle) in the painting style and texture of the small structures.

D USER STUDY

In this section, we give an example as shown in Figure 12 to illustrate the setting of the questions and the description of the options in our questionnaire. Each participant in this user study understands the content and purpose of our research.

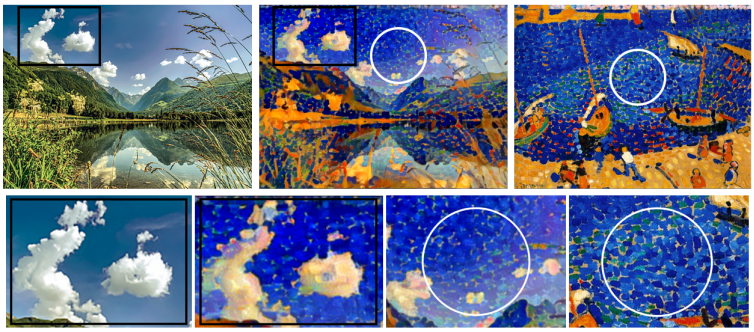

Figure 11: The rendered global and local features from two domains using our D3PSR model

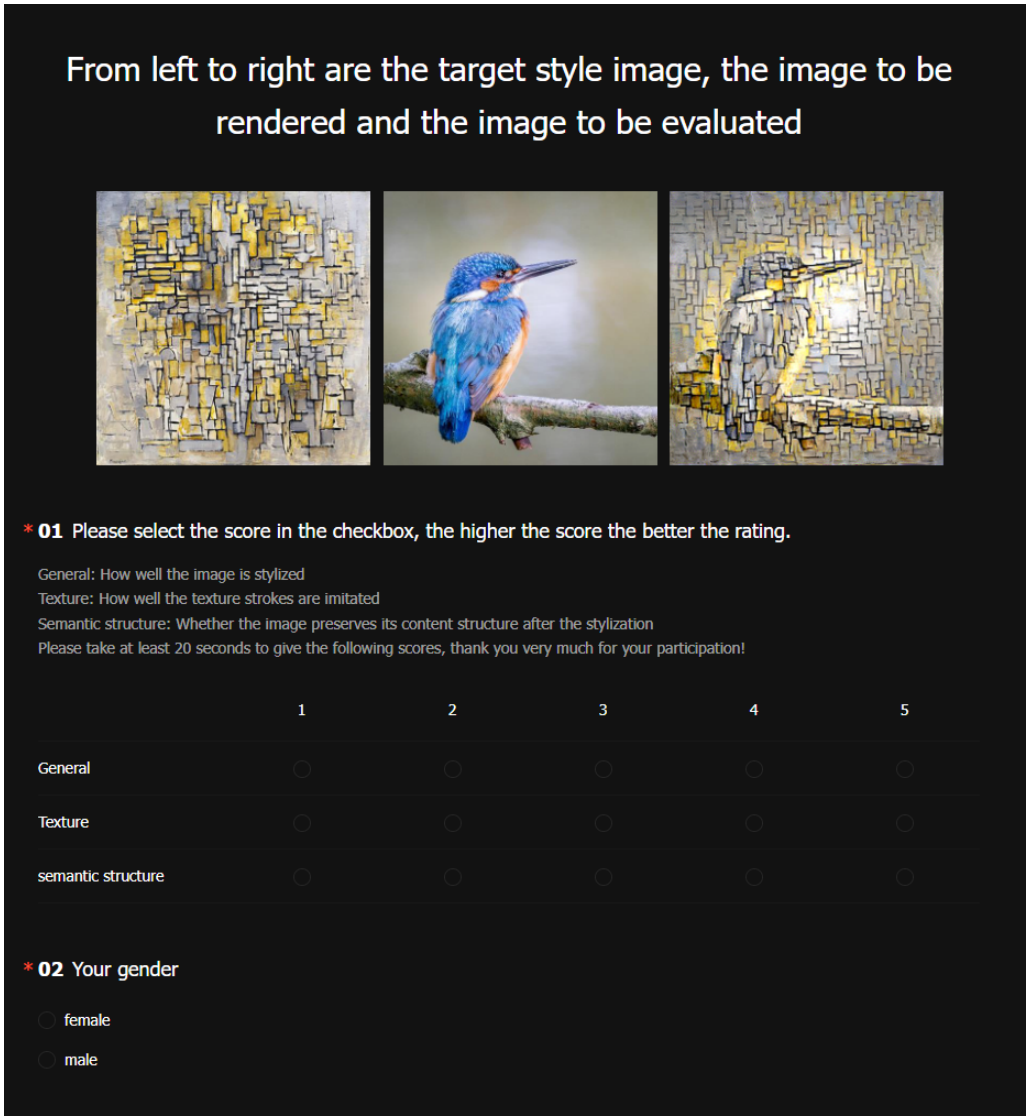

Figure 12: Example of questionnaire