

A THE SCHRÖDINGER BRIDGE PROBLEM

In this section, we will provide a detailed solution for the two problems proposed in Section 3.5 of the original article.

The First Process. In the 1st process of our approach, we can rewrite it through the perspective of SBP:

Problem 1. *The 1st process is to find a distribution from $\mathcal{D}(\alpha_0, \alpha_{N/2})$ that minimizes the KL-divergence: $\mathbb{Q}^{1*} := \arg \min \{D_{KL}(\mathbb{Q}^1 \|\mathbb{W}^1) \mid \mathbb{Q}^1 \in \mathcal{D}(\alpha_0, \alpha_{N/2})\}$, where $\alpha_0 = p_d(X_A)$, $\alpha_{N/2} = p_\theta(v)$, \mathbb{W}^1 is a prior reference measure.*

Proof. Let \mathbb{D} denote the set of all probability measures on Ω which are absolutely continuous with respect to stationary Winener measure. By Girsanov’s theorem any $\Pi \in \mathbb{D}$ has a forward drift ($\mu(t)$), and a backward drift ($\lambda(t)$), the canonical process has Itô differential such that:

$$\text{forward: } dx(t) = \mu(t)dt + d\mathcal{W}^+(t) \quad (19)$$

$$\text{backward: } dx(t) = \lambda(t)dt + d\mathcal{W}^-(t) \quad (20)$$

where $\mathcal{W}^+(t), \mathcal{W}^-(t)$ are standard Wiener processes adapted to the forward and reverse time diffusion. By defining $b(t, x(t)) = \mu(t) - \nabla \ln \hat{\phi}_t(x)$ (Pavon & Wakolbinger, 1991), where $\hat{\phi}_t \cdot \hat{\phi}_t = q_t$ and q_t represents the density of $x(t)$ that satisfies the Fokker-Planck (FPK) equation for the process of the form $dx(t) = b(t, x(t))dt + d\mathcal{W}(t)$, and referring to e.g. (Pavon & Wakolbinger, 1991) (Lemma 3.8) and (Léonard, 2014) (Theorem 2.4), the KLD between \mathbb{Q}^1 and \mathbb{W}^1 can be expressed as a decomposition:

$$\begin{aligned} D_{KL}[\mathbb{Q}^1 \|\mathbb{W}^1] &= \overbrace{D_{KL}[\mathbb{Q}_0^1 \|\mathbb{W}_0^1]}^{\text{constant}} \\ &+ \mathbb{E}_{\mathbb{Q}^1} \left[\int_0^{\frac{N}{2}} \frac{1}{2} \|\mu(t) - b(t, x(t))\|^2 d(t) \right] \end{aligned} \quad (21)$$

where \mathbb{Q}_0^1 and \mathbb{W}_0^1 denote the initial densities of \mathbb{Q}^1 and \mathbb{W}^1 , and the first term is constant. By the Theorem 3.9 (Pavon & Wakolbinger, 1991), we can obtain the forward equivalent objective for SBP of the 1st process such that:

$$F(\mathbb{Q}^1) := \min_{\mathbb{Q}^1 \in \mathcal{D}(\alpha_0, \alpha_{N/2})} \mathbb{E}_{\mathbb{Q}^1} \left[\int_0^{\frac{N}{2}} \frac{1}{2} \|\mu(t) - b(t, x(t))\|^2 d(t) \right] \quad (22)$$

Using reverse diffusion, we can also obtain the backward equivalent objective for SBP of the 1st process:

$$B(\mathbb{Q}^1) := \min_{\mathbb{Q}^1 \in \mathcal{D}(\alpha_0, \alpha_{N/2})} \mathbb{E}_{\mathbb{Q}^1} \left[\int_0^{\frac{N}{2}} \frac{1}{2} \|\lambda(t) - b_-(t, x(t))\|^2 d(t) \right] \quad (23)$$

Then, there holds:

$$\begin{aligned} D_{KL}[\mathbb{Q}^1 \|\mathbb{W}^1] &= D_{KL}[\mathbb{Q}_0^1 \|\mathbb{W}_0^1] + F(\mathbb{Q}^1) \\ &= D_{KL}[\mathbb{Q}_{N/2}^1 \|\mathbb{W}_{N/2}^1] + B(\mathbb{Q}^1) \end{aligned} \quad (24)$$

where $\mathbb{Q}_{N/2}^1$ and $\mathbb{W}_{N/2}^1$ denote the initial densities of \mathbb{Q}^1 and \mathbb{W}^1 at time index $t = \frac{N}{2}$.

Half Bridge Problem. To simplify the numerical solution of the iterative algorithms, we set α_0 as initial value and force it into single-constraint problems, which transforms the original problem into a half bridge problem (Pavon et al., 2021). Then the forward and backward half bridge of the 1st process is given by:

$$\begin{aligned} \text{forward: } \mathbb{Q}^{1*} &= \inf_{\mathbb{Q}^1 \in \mathcal{D}(\alpha_0, \cdot)} D_{KL}(\mathbb{Q}^1 \|\mathbb{W}^1) \\ \text{backward: } \mathbb{P}^{1*} &= \inf_{\mathbb{P}^1 \in \mathcal{D}(\cdot, \alpha_{N/2})} D_{KL}(\mathbb{P}^1 \|\mathbb{W}^1) \end{aligned}$$

Using e.g. (Pavon et al., 2021) and (Vargas, 2021)(Theorem 9&10), the optimal solution of static forward bridge holds:

$$q^*(x, y) = p^{\mathbb{W}}(x, y) \frac{\alpha_0(x)}{p^{\mathbb{W}}(x)} \quad (25)$$

where $p^{\mathbb{W}}(x, y) = p_0^{\mathbb{W}}(x)p^{\mathbb{W}}(y|x)$ with marginal prior $p_0^{\mathbb{W}}(x)$, the joint distribution $q(x, y) \in \mathcal{D}(\alpha_0(x), \alpha_{N/2}(y))$, $\alpha_0(x) = \int q(x, y)dy$, $\alpha_{N/2}(y) = \int p(x, y)dx$. The optimal solution of static backward bridge holds:

$$p^*(x, y) = p^{\mathbb{W}}(x, y) \frac{\alpha_{N/2}(y)}{p^{\mathbb{W}}(y)} \quad (26)$$

Now, we have completed the description of the 1st process from the perspective of SBP and provided the objectives. Half bridge's solutions can be considered "closed-form" to some extent, they can also be used to remove constraints by including them as an initial value problem, which provides simplification objectives for solving SBPs problems using iterative methods.

The Second Process. Similar to the 1st process, the 2nd process can also be represented as a description in terms of SBP.

Problem 2. The 2nd process is to find a distribution from $\mathcal{D}(\alpha_{N/2}, \alpha_N)$ that minimizes the KL-divergence: $\mathbb{Q}^{2*} := \arg \min\{D_{KL}(\mathbb{Q}^2 \|\mathbb{W}^2) \mid \mathbb{Q}^2 \in \mathcal{D}(\alpha_{N/2}, \alpha_N)\}$, where $\alpha_{N/2} = p_\theta(v)$, $\alpha_N = p_d(X_B)$, \mathbb{W}^2 is a prior reference measure.

Similar to the discussion of the 1st process, the forward and backward half bridge of the 2nd process is given by:

$$\begin{aligned} \text{forward: } \mathbb{Q}^{2*} &= \inf_{\mathbb{Q}^2 \in \mathcal{D}(\alpha_{N/2}, \cdot)} D_{KL}(\mathbb{Q}^2 \|\mathbb{W}^2) \\ \text{backward: } \mathbb{P}^{2*} &= \inf_{\mathbb{P}^2 \in \mathcal{D}(\cdot, \alpha_N)} D_{KL}(\mathbb{P}^2 \|\mathbb{W}^2) \end{aligned}$$

So far, the formulation of the 2nd process towards SBP and the objectives are given.

B CONTROL FACTOR

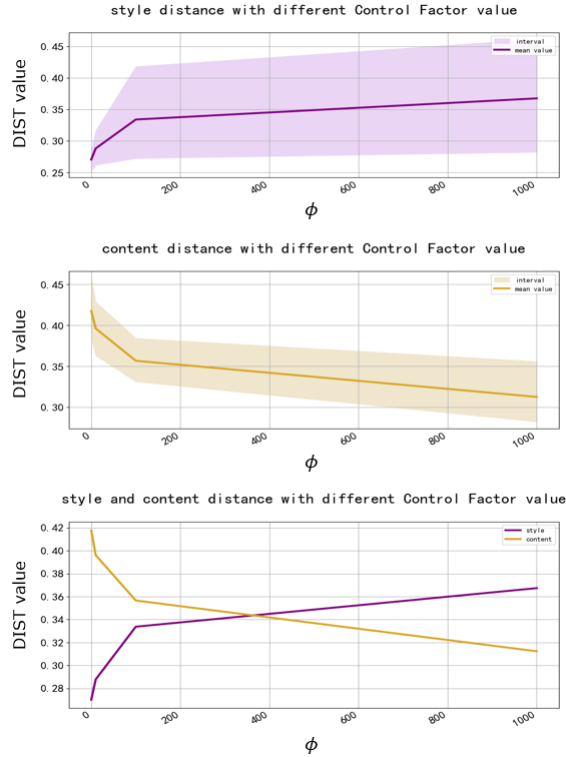
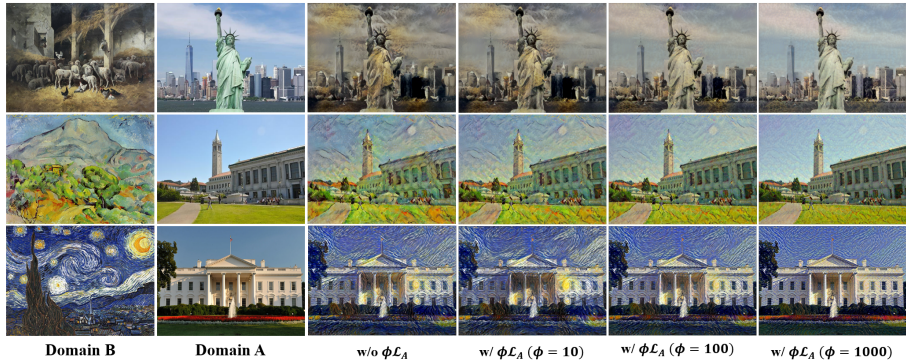


Figure 4: The DIST value with different control factor ϕ .

Figure 5: The outputs with different control factor ϕ .Table 3: 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

The distinct characteristic of our model in comparison to other methods is the conditioning mechanism, which allows for the flexibility to modify the degree of texture rendering while maintaining semantic structure preservation. Our model can produce a wide range of stylized results by adjusting the Control Factor (CF), ϕ , to control the balance between structure and texture, as illustrated in Figure 5.

Qualitative evaluation. By altering the value of ϕ , the level of stylization and the semantic structure in the generated image can be adjusted. As depicted in Figure 5, as the value of ϕ increases, the semantic structure of the image becomes more defined (e.g. windows and doors on buildings, outlines of statues, etc.), while at the same time the level of stylization decreases (e.g. brushstrokes and textures in the image, etc.). It is important to note that one cannot excessively reduce the value of ϕ in an effort to achieve a stronger stylistic transition, as this can result in certain areas of the image becoming overwhelmed (e.g. when $\phi = 1$ and 10 in sample 1 and 2). Similarly, the value of ϕ should not be increased excessively in an attempt to obtain a sharper semantic structure, as this can result in the generated image being insufficiently stylized (e.g. the stylized strokes and textures from Domain B are weak at $\phi = 1000$ in samples 1 and 3).

Quantitative evaluation. Table 3 records the DISTS values (Ding et al., 2020) of the generated images in Figure 5 in comparison to the target style image and the original content image. This data can then be used to generate the line graph depicted in Figure 4. Through quantitative analysis, it is evident that there is a clear trade-off between style DIST and content DIST, indicating that an enhancement in stylization is accompanied by a loss of semantic structural information. Furthermore, it can be observed that as the semantic structure becomes sharper, the stylization is weakened.

C PROGRESSIVE RENDERING IMPLEMENTATION

As described in the original text, we employed VGG as an encoder to extract corresponding feature from the content and style images as priors. The content feature target vector is extracted from

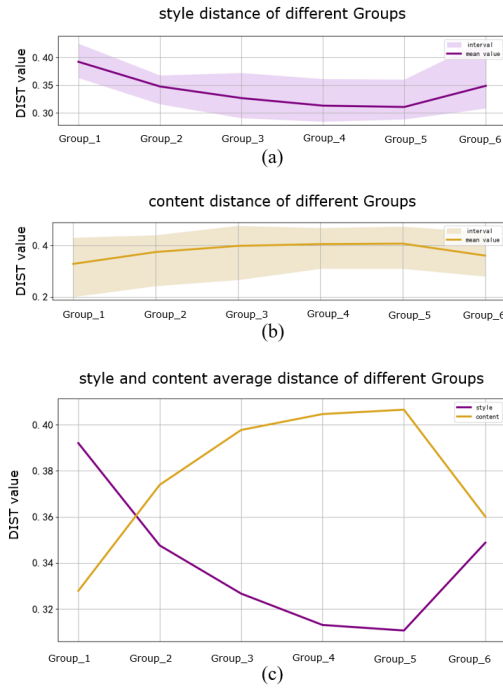


Figure 6: The DISTs value varies over different Groups.

layer ‘conv4_2’, and the style feature target vectors are extracted from layers ‘conv1_1’, ‘conv2_1’, ‘conv3_1’, ‘conv4_1’ and ‘conv5_1’ (Gatys et al., 2016). In this section, we will discuss the impact of using varying numbers of features on the final output results.

Group Design. In Group_1, we only utilized the style feature (‘conv1_1’) from the style image. In Group_2, we employed style features (‘conv1_1’ and ‘conv2_1’) from the style image. In Group_3, we utilized style features (‘conv1_1’, ‘conv2_1’ and ‘conv3_1’) from the style image. In Group_4, we employed style features (‘conv1_1’, ‘conv2_1’, ‘conv3_1’ and ‘conv4_1’) from the style image. In Group_5, we utilized style features (‘conv1_1’, ‘conv2_1’, ‘conv3_1’, ‘conv4_1’ and ‘conv5_1’) from the style image. In Group_6, we employed style features (‘conv1_1’, ‘conv2_1’, ‘conv3_1’, ‘conv4_1’ and ‘conv5_1’) from the style image and content feature (‘conv4_2’) from the content image. Furthermore, we only impose the CF term in group 6 among all the groups.

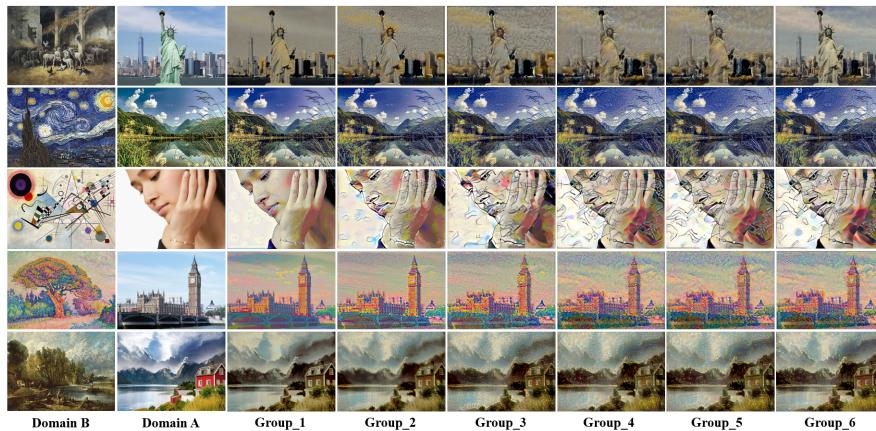


Figure 7: The progressive changes of each Group.

Table 4: Quantitative DISTs value with different Groups.

	Group_1	Group_2	Group_3	Group_4	Group_5	Group_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

Qualitative evaluation. As shown in Figure 7, as the number of style feature targets imposed on the images increases from Group_1 to Group_5, the stylization in the images becomes more prominent (e.g. brush strokes and textures). However, we can observe that simply increasing the style features could result in certain critical semantic structures in the image becoming increasingly blurred. With the reference of ϕ and the content feature target, the images in Group_6 not only successfully achieve stylization but also make the semantic structures clearer than the previous outputs.

Quantitative evaluation. The Table 4 presents the DISTs values (Ding et al., 2020) of the output images compared to the target images, and it can be observed that as more style feature targets are progressively imposed on each Group, the style DISTs value is getting smaller, indicating 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 stylized, resulting in a gradual blurring of contours from the content image. However, with the guidance of the content feature target and the CF ϕ in Group_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. At the same time, the style DISTs value becomes larger, indicating a drop in performance at the stylized level compared to the previous Group.

This analysis and Figure 6 reveal that in the process of style transfer, there is a trade-off between preserving semantic structure and stylization, and it is difficult to preserve both at the same time. However, our model’s unique CF mechanism allows for greater flexibility in terms of controlling the extent to which semantic structure is preserved, enabling the user to select results from Group_6 when a clearer content image is desired, or results from Group_5 when stronger stylization is needed.

D SEMANTIC STRUCTURE PRESERVATION

When performing style transfer from a content image to an abstract painting, it particularly tests the model’s ability to preserve semantic structures. As shown in Figure 8, our model exhibits strong performance in preserving image semantic structures. In the original image, the woman is wearing a bracelet on her wrist, which is not present in the rendered result in Group_3 due to the absence of the content target or the ϕ that adjusts the sharpness of the original semantic structure in the output image. Furthermore, because Group_3 only imposes lower-level vectors (‘conv1_1’, ‘conv2_1’ and

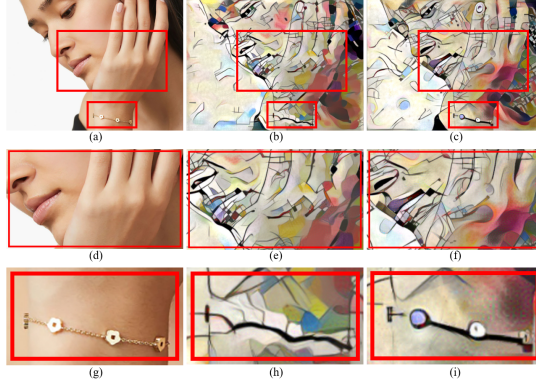


Figure 8: Evaluation of semantic structure preservation. The original image is denoted by (a), the image from Group_3 is denoted by (b), and the image from Group_6 is denoted by (c). Detailed display images corresponding to (d) through (i) are also provided.

‘conv3_1’) of CNN, it is more biased towards simulating local small structure textures from Domain B, resulting in a cluster of small structures near the nose and mouth. In contrast, Group_6 imposes more high-level features (‘conv4_1’ and ‘conv5_1’), the content target and ϕ , which controls the extent of semantic structure preservation. As a result, the bracelet and the hand curve, two semantic structures in the source image, are nicely preserved in the final rendering results in Group_6, further demonstrating the superior performance of our method.

E ABLATION STUDY

In this section, we conduct several ablation studies on the number of reference feature vectors $\{\sum_{i=1}^5 (s_i, c_1)\}$, the distance of Metric Space L_D , loss term and the control factor (CF) ϕ .

Reference Feature Vectors. In the experiment, by utilizing Equation (14) (Section 3.3) as our loss function, setting ϕ to 358, and using the Euclidean distance L_D , we impose a varying number of $\{\sum_{i=1}^5 (s_i, c_1)\}$ as prior references and obtain different output images, as shown in Figure 9. Based on the outputs’ performance, with a small number of style features as references, the style

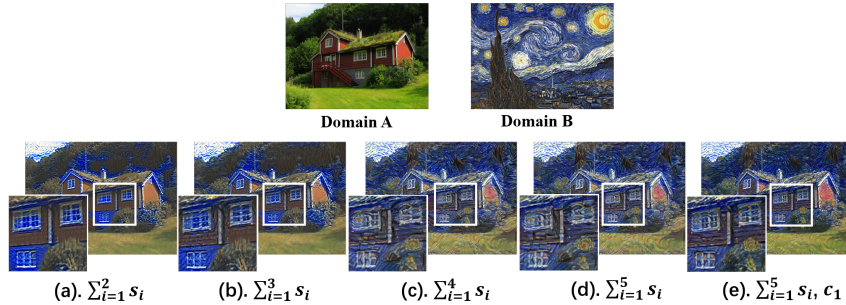


Figure 9: Ablation study of different number of target feature vectors. More detailed results and evolution are illustrated in supplementary materials.

transfer results are not satisfactory. With more style features without content feature, the stylization is improved, but the semantic structure is not well preserved. With all content and style features, the output image can preserve clear semantic content structure and achieve extraordinary style transfer.

Metric Space Distance. In order to verify the effectiveness of distances in different metric spaces on model performance, we compared the images generated using the three distances defined in Section 3.3, as shown in Figure 10. The results demonstrate that using the \mathbb{C} space cannot complete the style transfer task smoothly; using the l^p space results in poorer style transfer effect, and some semantic

content is not achieved in style transfer (as shown in the red box in the figure); using the *Euclidean space* \mathbb{R}^n can achieve satisfactory performance.

Loss Term and the Control Factor. To validate the necessity and rationality of the control factor term $\phi\mathcal{L}_A$ in Equation (14) (Section 3.3), we conducted comparative experiments as shown in Figure 11. In the experiments, we observed the influence of the $\phi\mathcal{L}_A$ on the experimental results, and obtained corresponding output images by modifying the hyperparameter ϕ . The results show that when the $\phi\mathcal{L}_A$ is not introduced, or the ϕ value is small, the style transfer will cause the semantic structure of the content image to be blurred and lose some semantic content. Moreover, serious overflow will occur in the originally clean background. Increasing ϕ will alleviate the above problems, but if ϕ is too large, the degree of stylization will be low, affecting the effect of style transfer.

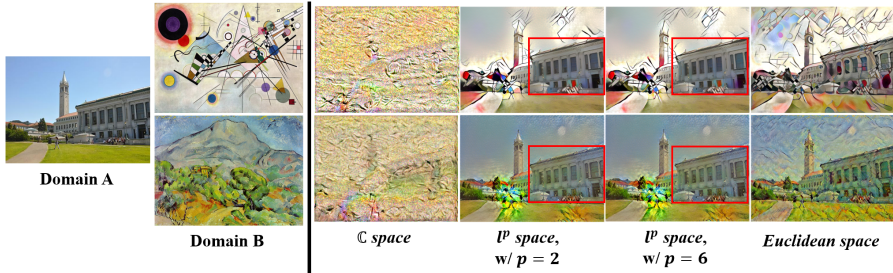


Figure 10: Ablation study of the Metric space distance.

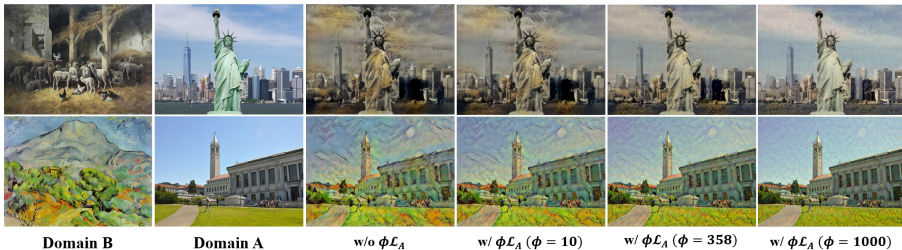


Figure 11: Ablation study of the Loss term and the control factor. We compared the results generated by using or not using the CF term in loss function, and observed the impact of adjusting the ϕ value on the images. More quantitative analysis and further discussions on this are presented in the supplementary material.

F EXTRA TEST SAMPLES

Due to the page limit, we only showcased four samples in Figure 3 in the paper. Figure 12 shows the remaining samples used in our tests.

G USER STUDY

User study has been discussed in Section 4.4. In this section, we provide an example illustrated in Figure 13 to demonstrate the formatting of the questions and the description of the options in our questionnaire. The participants in this user study have been informed of the content and objectives of our research.

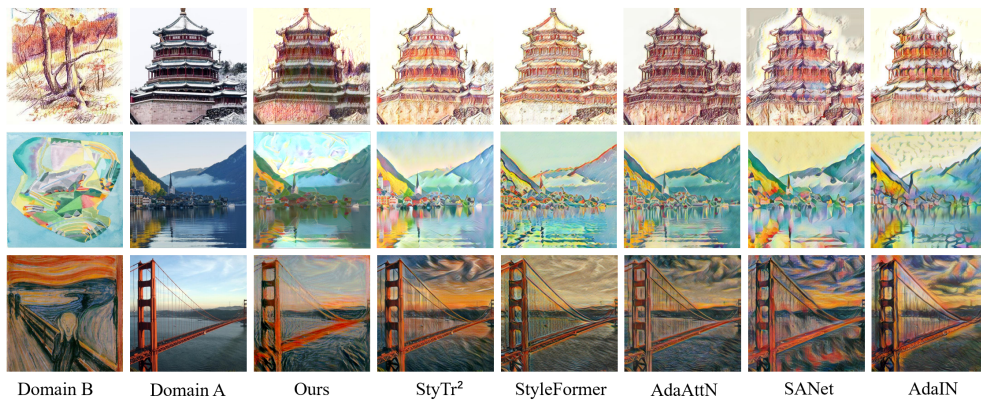





Figure 12: This figure presents the extra samples.

From left to right are the target style image, the image to be rendered and the image to be evaluated

*** 01** Please select the score in the checkbox, the higher the score the better the rating.

General: How well the image is stylized
 Texture: How well the texture strokes are imitated
 Semantic structure: Whether the image preserves its content structure after the stylization
 Please take at least 20 seconds to give the following scores, thank you very much for your participation!

	1	2	3	4	5
General	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Texture	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
semantic structure	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

*** 02** Your gender

female
 male

Figure 13: Example of questionnaire