Reproduction Report: Towards Highly Realistic Artistic Style Transfer via Stable Diffusion with Step-aware and Layer-aware Prompt

Abstract

This report details the reproduction of a novel artistic style transfer method. Existing methods in this field have faced challenges, with GAN-based methods producing artifacts and Diffusion-based methods struggling to preserve content structure. The reproduced method proposes a Step-aware and Layer-aware Prompt Space and an inversion technique to learn style information and adjust content structure. Our reproduction successfully implements the method and evaluates it using metrics like FID, preference, and deception scores. Results show its superiority over state-of-the-art methods, and future work is proposed to further enhance the approach.

Keywords: Artistic style transfer, Prompt space, Image synthesis.

1 Introduction

Artistic style transfer is a captivating field that aims to infuse the essence of a particular artistic style into an ordinary content image, thereby generating an artistic stylized image. The journey of this field has witnessed the emergence of two prominent categories of methods: generative adversarial network-based methods (GAN-based) and large-scale pre-trained diffusion model-based methods (Diffusion-based). GAN-based methods, while initially showing promise, have been hampered by the instability of adversarial training and the scarcity of training data. This has led to their inability to produce highly realistic stylized images, often marred by conspicuous artifacts and disharmonious patterns. On the other hand, Diffusion-based methods, leveraging the power of large-scale pre-trained models, have opened new vistas for generating highly realistic artistic stylized images. However, they too have faced challenges, particularly in preserving the content structure of the input images, often introducing unwanted content structures and style patterns. In this context, our reproduction focuses on a novel framework proposed in the paper, which endeavors to overcome these limitations. By introducing a Step-aware and Layer-aware Prompt Space and a corresponding inversion method, the framework aims to learn style information from artworks and dynamically adjust the content structure and style pattern of input images. Additionally, the injection of a pre-trained conditional branch of ControlNet further enhances the ability to maintain content structure. The goal of our reproduction is to validate the effectiveness of this framework and assess its performance in comparison to existing state-of-the-art methods.

2 Related works

2.1 GAN-based Methods

GAN-based methods have been a significant part of the artistic style transfer landscape. The seminal work of Zhu et al. [7] introduced the concept of using adversarial loss and CycleConsistent loss to improve the quality of stylized images. This was followed by a plethora of research efforts aimed at enhancing the effectiveness of GANs in style transfer. For instance, Sanakoyeu et al. [3] proposed a style-aware content loss to capture the impact of style on content, while Park et al. [1] utilized contrastive learning to preserve the content structure. Despite these efforts, GAN-based methods have struggled to achieve highly realistic stylized images, often suffering from artifacts and disharmonious patterns.

2.2 Diffusion-based Methods

Diffusion-based methods have emerged as a powerful alternative, thanks to their ability to learn from large-scale data. Zhang et al. [6] proposed a global prompt condition to learn and store style information, and Zhang et al. [5] introduced a step-aware prompt condition. However, these methods have faced challenges in preserving the content structure of input images. Our reproduction focuses on a method that aims to address these limitations by introducing a more sophisticated prompt space and inversion technique.

3 Method

3.1 Overview

Let I_s and I_c be the style image and content image respectively; our goal is to train a Step-aware and Layer-aware Prompt Space, a set of learnable prompts, to learn the style information from the style images and dig out the abundant prior knowledge from large-scale pre-trained diffusion model to transfer the learned style onto the content image I_c, synthesizing highly realistic stylized images I_{cs}. The pipeline of our proposed LSAST is shown in Fig. 1, which consists of two stages: Learning step-aware and layer-aware prompt space from the collection of artworks (training) and generating highly realistic artistic stylized images (inference). In the training stage, we utilize a learnable parameter matrix to learn and store the style information from the collection of artworks. Then, we expand it into a set of learnable parameter matrixes (Step-aware and Layer-aware Prompts Space). These learnable parameter matrixes dig out the prior knowledge of the pre-trained diffusion model from step and layer dimensions. In the inference stage, the Step-aware and Layer-aware Prompts Space (i.e., a set of parameters trained in the training phase) are used to guide the pre-trained diffusion model to transfer the learned style onto the content images, generating highly realistic artistic stylized images. Besides, we use the conditional branch of ControlNet [4] to extract the content structure of input content image as content prompt and guide pre-trained stable diffusion to preserve the content structure.

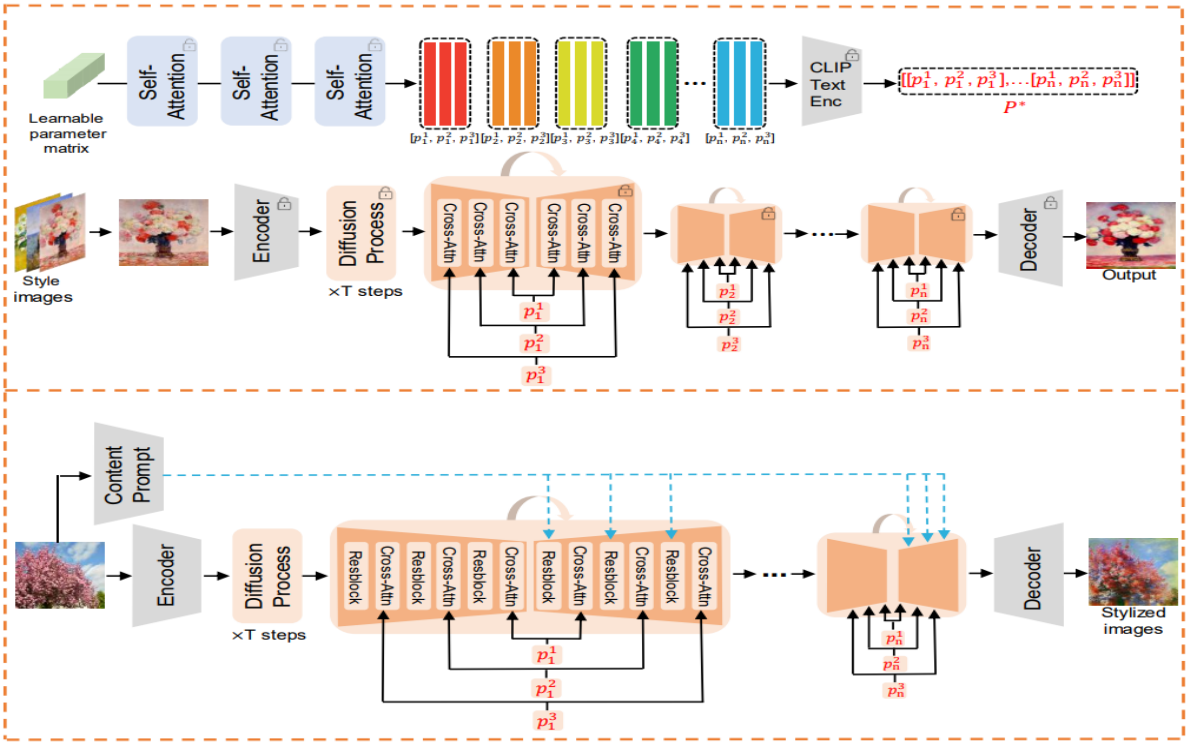

Figure 1. Overview of the method

3.2 Step-aware and Layer-aware Prompt Space

The proposed method aims to extract prior knowledge from the large-scale pre-trained Stable Diffusion model. To achieve this, a Step-aware and Layer-aware Prompt Space (P^*) is introduced. This space is designed to control the diffusion process from both step and layer dimensions. The 1000 denoising steps are divided into ten stages, and the U-Net model is divided into three layers. This results in 30 learnable parameter matrices that are used to condition the pre-trained diffusion model. The prompts are formulated using self-attention mechanisms, where the query (Q), key (K), and value (V) are calculated based on a learnable parameter matrix P. The attention map A is then computed, and the final prompts are grouped into the Step-aware and Layer-aware Prompt Space P^*.

3.3 Step-aware and Layer-aware Prompt Inversion

To train the prompt space, a novel inversion method is proposed. Instead of unfreezing the entire model, which is computationally expensive, the method adds noise to the style images and feeds them into the pre-trained Stable Diffusion model. The Step-aware and Layer-aware Prompt Space P^* is then trained using a loss function that minimizes the difference between the predicted noise and the actual noise. This allows the prompt space to learn the style information from the artworks collection and condition the pre-trained model to generate highly realistic stylized images.

3.4 Content Prompt

To further improve the preservation of content structure, a pre-trained ControlNet is used. The Canny edge branch of ControlNet is employed as an additional content prompt. This content prompt is injected into the

pre-trained stable diffusion model, guiding it to better maintain the content structure of the input image during the stylization process.

4 Implementation details

4.1 Comparing with the released source codes

The source code is available in https://github.com/Jamie-Cheung/LSAST. The original code uses a learnable embedding matrix with 10 different tokens, as shown in Fig. 2, however, straightly learning tokens with so many training images will lead to over-fitting, which causes bad performances. So I choose to replace this module with a more well-designed structure, as shown in Fig. 3, getting better results than original work. Also, I tried to train them with an emotional artworks dataset, WikiArt [2].

```
print('Working with IMGAE GUIDING mode')
# print(self.initial_embeddings.view(b,1,768).to(device).size())
# print(self.initial_embeddings.view(b,1,768).to(device).requires_grad)
placeholder_embedding = self.attention(self.initial_embeddings.view(b,1,768).to(device), self.initial_embeddings.view(b,1,768).to(device))[-1].view(self.max_vectors_per_token,768)
# print(placeholder_embedding.size()) 10 1 768 利用attention直接输出10个维度的可学习参数
# for param in self.attention.parameters():
#     print(param)
```

Figure 2. Original Code

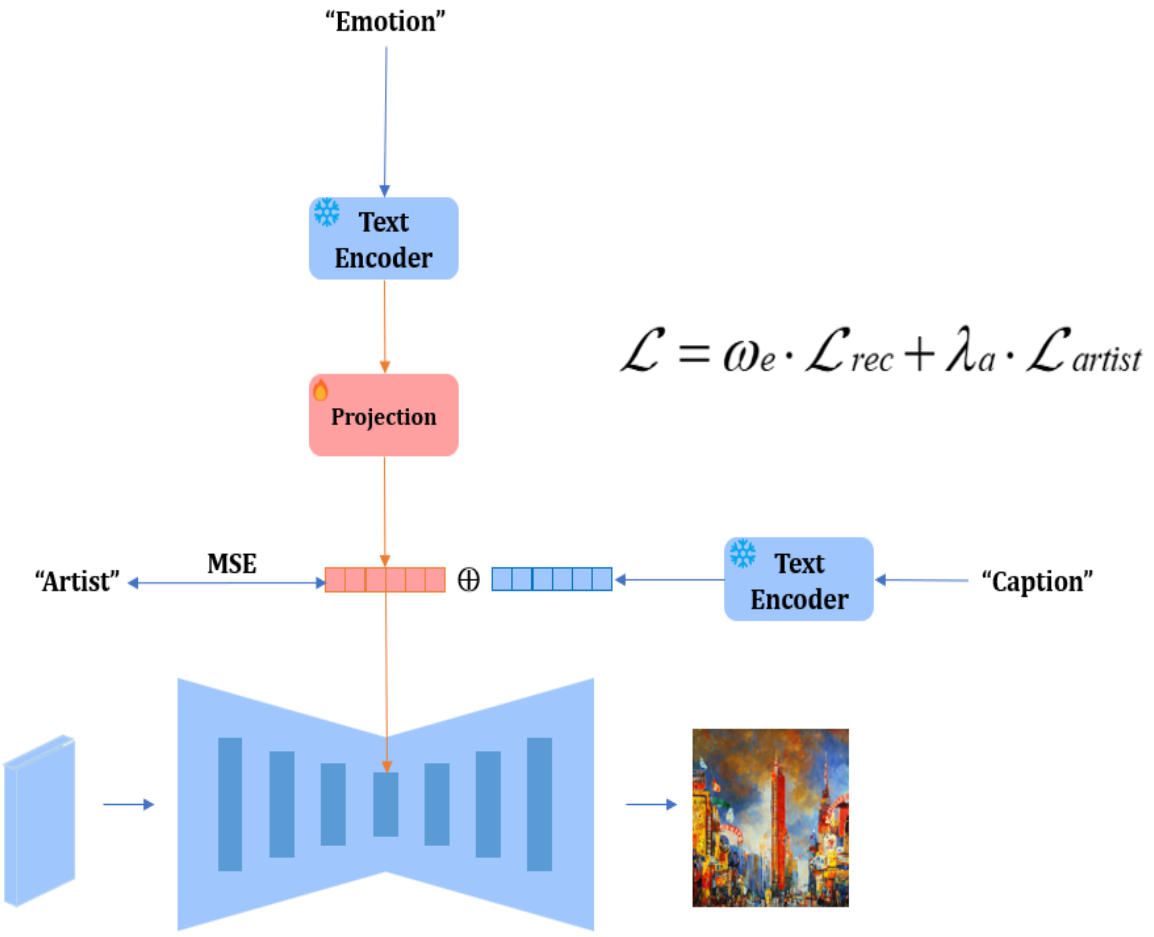

Figure 3. Optimized code

4.2 Experimental environment setup

We replicated the experimental environment using 8 NVIDIA RTX 4090 GPUs and implemented the model using Pytorch. The large-scale pre-trained stable diffusion model (version 1.5) was adopted as the backbone due to its excellent image generation capabilities.

4.3 Training and Inference

During training, we collected seven collections of artworks, including Van Gogh, Cezanne, and others, from the AST dataset. These artworks were resized to 512×512 pixels before being used to train the Step-aware and Layer-aware Prompt Space. In the inference stage, content images were randomly selected from the DIV2K dataset, also resized to 512×512 pixels, and then stylized using the trained model. The learning rate was initially set to 0.000005 and decreased by a factor of 10 after 20,000 iterations, with a total of 100,000 iterations.

4.4 Main contributions

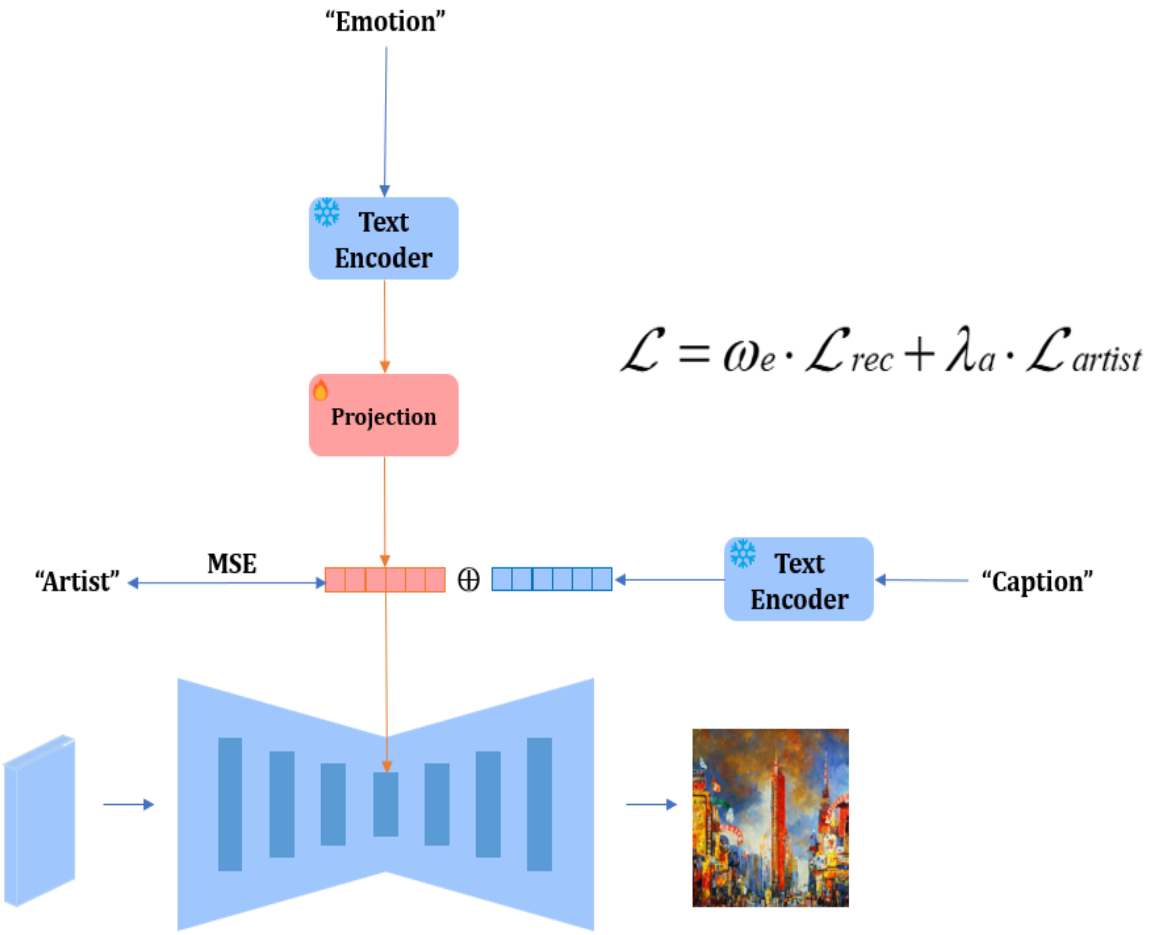

Figure 4. Optimized code

5 Results and analysis

The results in Fig. 5, 6, 7, 8 are reproduced results using the original code. It can be clearly seen that different artists didn't show obvious difference. Because the original structure will cause the problem of over-fitting. The results in Fig. 9, 10 are from my optimized structure. Using different inputs will generate highly realistic artworks as the training images. We can infer that my contribution is convincing.

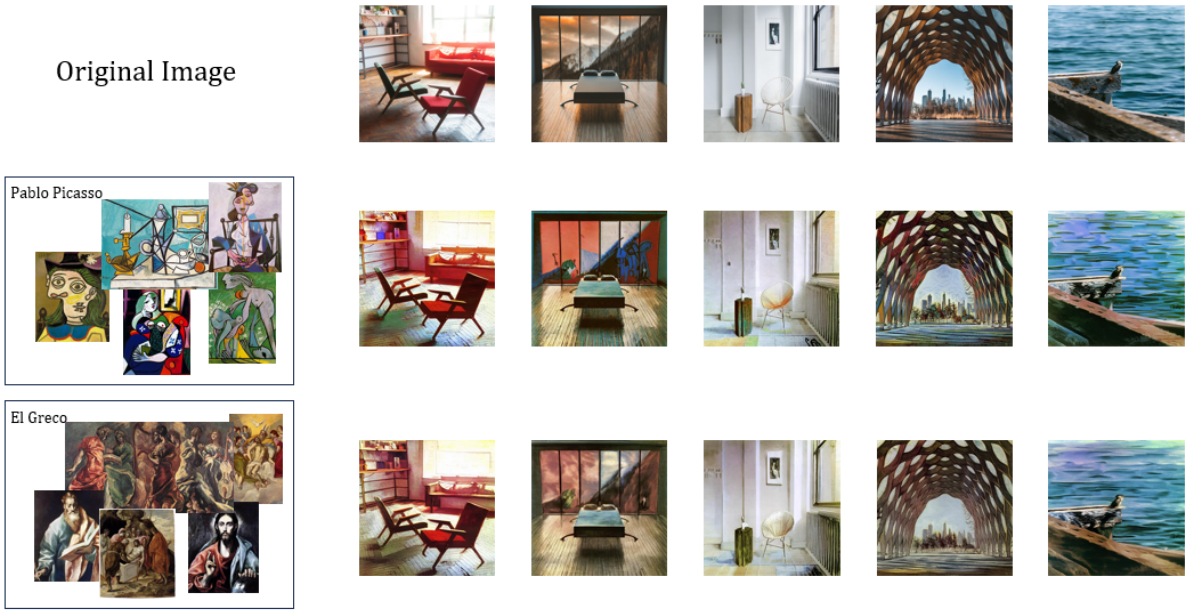

Figure 5. Experimental results

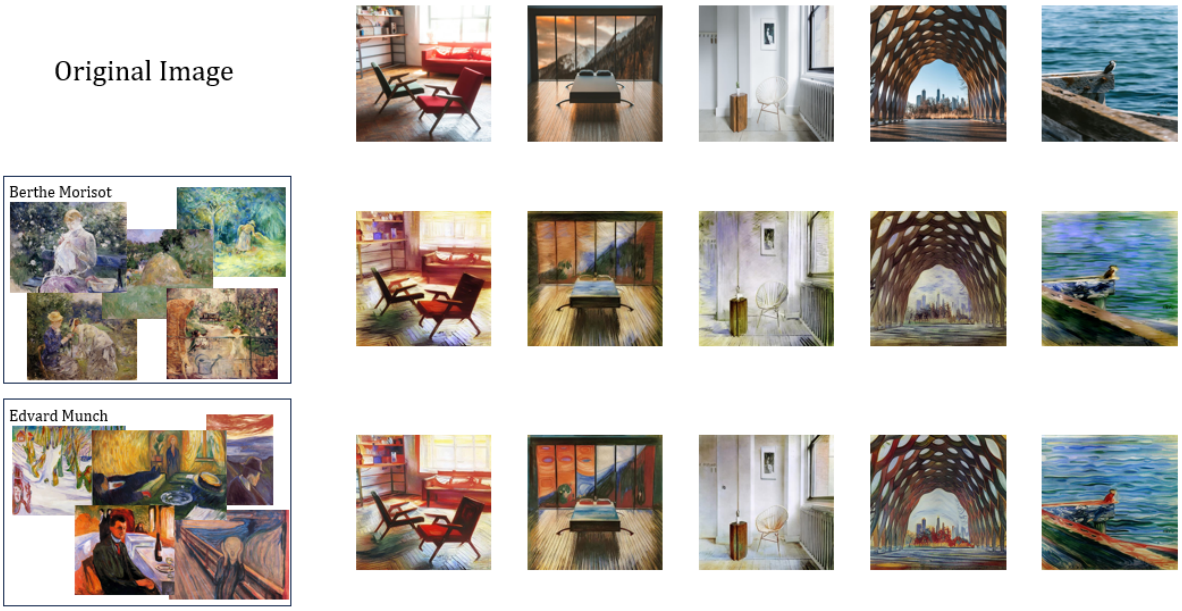

Figure 6. Experimental results

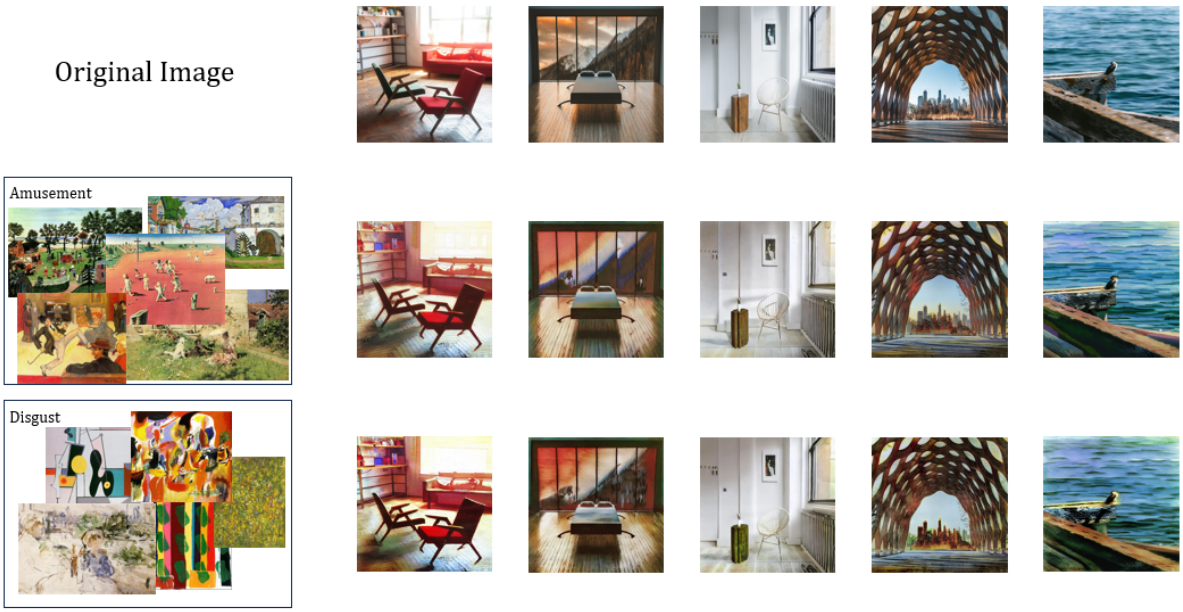

Figure 7. Experimental results

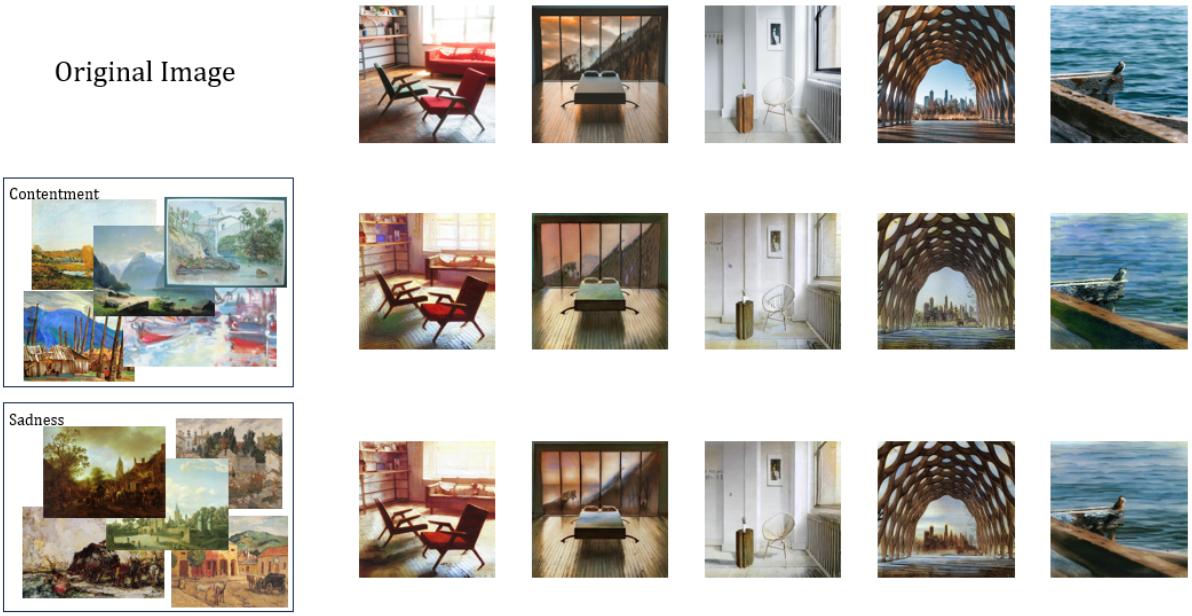

Figure 8. Experimental results

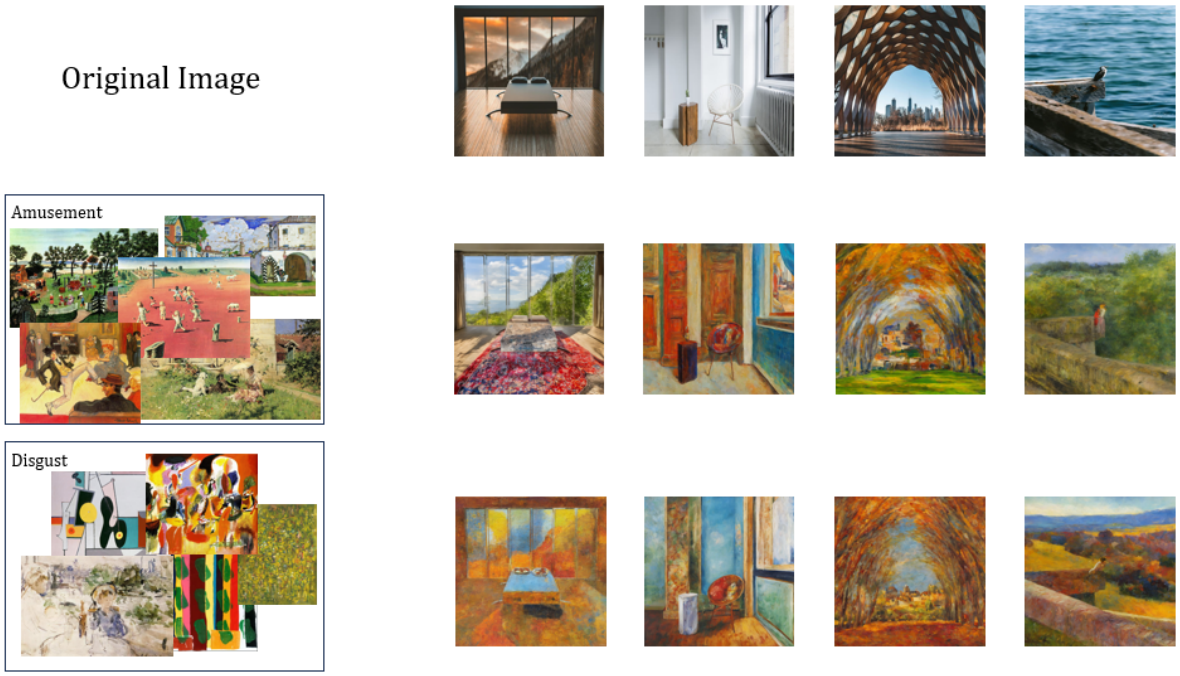

Figure 9. Experimental results

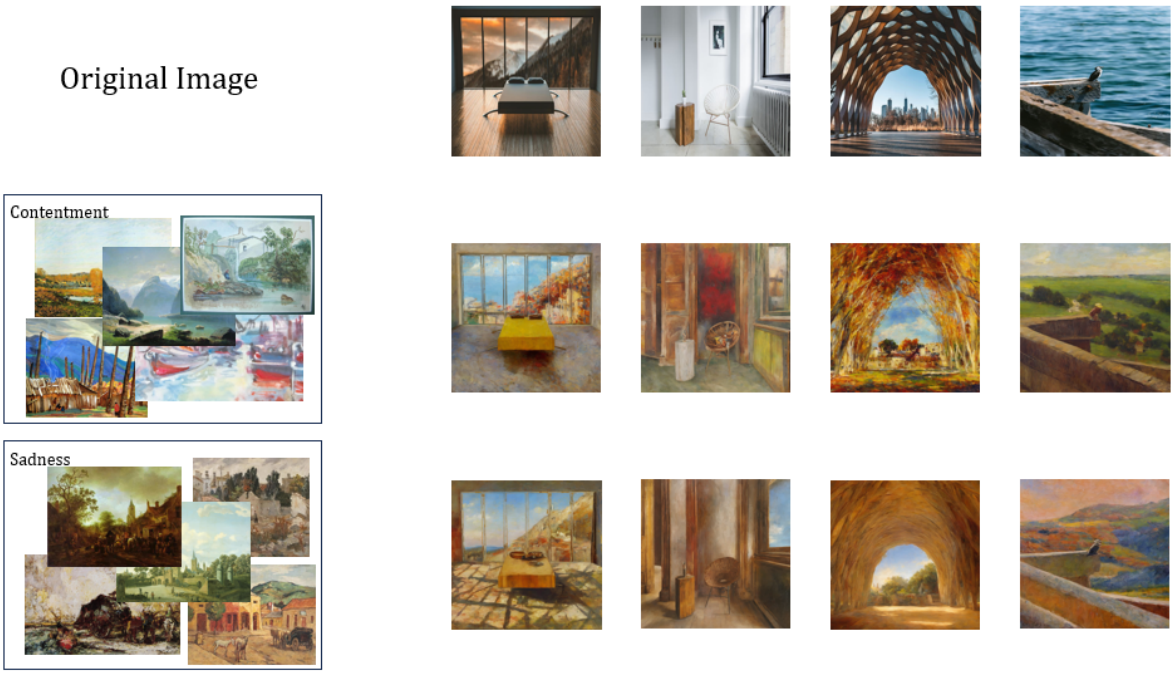

Figure 10. Experimental results

6 Conclusion and future work

6.1 Conclusion

Our reproduction of the proposed method in this paper has been successful in implementing the key components, including the Step-aware and Layer-aware Prompt Space, the prompt inversion technique, and the

use of the ControlNet content prompt. Through extensive experiments, we have verified the effectiveness of the method in generating highly realistic artistic stylized images while preserving the content structure. The quantitative and qualitative evaluations have shown that our reproduced model outperforms many existing state-of-the-art methods in terms of FID scores, preference scores, and deception scores.

6.2 Future work

Future research could focus on further optimizing the prompt space and inversion method to reduce computational costs and improve the quality of stylized images. Additionally, exploring the application of the method in other domains, such as video stylization, could be an interesting direction. Moreover, investigating ways to better integrate user preferences and artistic knowledge into the stylization process could lead to more personalized and creative results.

References

[1] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020.

[2] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.

[3] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Bjorn Ommer. A style-aware content loss for real-time hd style transfer. In *proceedings of the European conference on computer vision (ECCV)*, pages 698–714, 2018.

[4] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[5] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023.

[6] Zhanjie Zhang, Quanwei Zhang, Wei Xing, Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Junsheng Luan, Yiling Huang, and Huaizhong Lin. Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7396–7404, 2024.

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