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

Zhanjie Zhang[1*], Quanwei Zhang[1*], Huaizhong Lin[1†], Wei Xing[1†], Juncheng Mo[1], Shuaicheng Huang[2], Jinheng Xie[3], Guangyuan Li[1], Junsheng Luan[1], Lei Zhao[1†], Dalong Zhang[1], Lixia Chen[4†]

[1] Zhejiang University
[2] Anhui University
[3] National University of Singapore
[4] Zhejiang Gongshang University

{cszzj, cszqw, linhz, wxing, csmjc, cslgy, l.junsheng121,cszhl}@zju.edu.cn,
Z02114187@stu.ahu.edu.cn, xiejinheng2020@email.szu.edu.cn, zdlxing@126.com, clixia@163.com

Abstract

Artistic style transfer aims to transfer the learned artistic style onto an arbitrary content image, generating artistic stylized images. Existing generative adversarial network-based methods fail to generate highly realistic stylized images and always introduce obvious artifacts and disharmonious patterns. Recently, large-scale pre-trained diffusion models opened up a new way for generating highly realistic artistic stylized images. However, diffusion model-based methods generally fail to preserve the content structure of input content images well, introducing some undesired content structure and style patterns. To address the above problems, we propose a novel pre-trained diffusion-based artistic style transfer method, called LSAST, which can generate highly realistic artistic stylized images while preserving the content structure of input content images well, without bringing obvious artifacts and disharmonious style patterns. Specifically, we introduce a Step-aware and Layer-aware Prompt Space, a set of learnable prompts, which can learn the style information from the collection of artworks and dynamically adjusts the input images' content structure and style pattern. To train our prompt space, we propose a novel inversion method, called Step-ware and Layer-aware Prompt Inversion, which allows the prompt space to learn the style information of the artworks collection. In addition, we inject a pre-trained conditional branch of ControlNet into our LSAST, which further improved our framework's ability to maintain content structure. Extensive experiments demonstrate that our proposed method can generate more highly realistic artistic stylized images than the state-of-the-art artistic style transfer methods. Code is available at https://github.com/Jamie-Cheung/LSAST.

[*] Both authors contributed equally to this research.
[†] Corresponding authors.

1 Introduction

Artistic style transfer has recently attracted widespread attention in academia and industry since the seminal work of CycleGAN [Zhu *et al.*, 2017]. Existing artistic style transfer methods can be divided into generative adversarial network-based approaches (GAN-based approaches) and large-scale pre-trained diffusion model-based approaches (Diffusion-based approaches).

More specifically, GAN-based methods [Zhu *et al.*, 2017; Sanakoyeu *et al.*, 2018; Kim *et al.*, 2019; Park *et al.*, 2020; Chen *et al.*, 2023; Zhang *et al.*, 2023d; Li *et al.*, 2023c] generally utilize generative adversarial network and a training set of aligned/unaligned image pairs to learn the mapping between an input image and an output image. For example, Zhu et al. [Zhu *et al.*, 2017] used two mirror generative adversarial network to learn and improve the mapping between the input image and output image, synthesizing artistic stylized images. Sanakoyeu *et al.* [Sanakoyeu *et al.*, 2018] proposed a style-aware content loss, to improve the quality of artistic stylized images by capturing how style patterns affect content structure. However, GAN-based methods are limited by the instability of adversarial training and the scarcity of training data, failing to generate highly realistic artistic stylized images and introducing the obvious artifacts and disharmonious patterns on the stylized images (Please see in Fig. 1 (e-h)).

Large-scale pre-trained diffusion model-based approaches [Nichol *et al.*, 2021; Wu, 2022; Ho *et al.*, 2020; Hu *et al.*, 2023; Li *et al.*, 2024] use massive parameters to learn and store the information from the large-scale training data, possessing the ability to generate highly realistic images. This opens up a new way for generating highly realistic artistic stylized images. For example, Zhang *et al.* [Zhang *et al.*, 2024b] proposed a global prompt space, a learnable parameter matrix, to learn and store the style information from the collection of artworks and condition pre-trained large-scale diffusion model to generate artistic stylized image. Zhang *et al.* [Zhang *et al.*, 2023b] introduced a step-aware prompt space, a set of learnable parameter matrixes, for the whole diffusion process to generate desired stylized image. Although these approaches could generate highly realistic stylizied images, they failed to preserve the

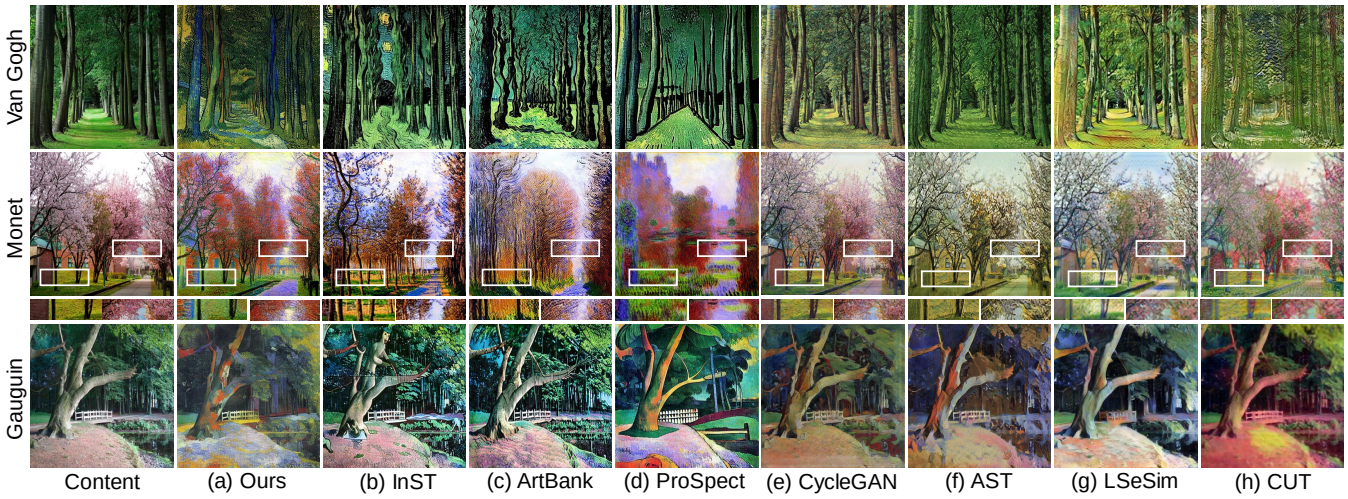

Figure 1: Stylization examples with three different styles (i.e., Van Gogh, Monet, Gauguin). Compared to existing state-of-the-art large-scale pre-trained diffusion model-based methods (b-d) and generative adversarial network-based methods (e-h), our proposed method (a) generates highly realistic artistic stylized images and preserves the content structure of input content images well.

content structure of input content image well, bringing some undesired content structure and style patterns (Please see in Fig. 1 (b-d)).

To summarize, the main contribution of this paper is as follows:

- We propose a novel pre-trained diffusion-based artistic style transfer framework, which can generate highly realistic stylized images and preserve the content structure of input content images well without introducing obvious artifacts and disharmonious style patterns.
- We design a novel Step-aware and Layer-aware Prompt Space and conduct a Step-aware and Layer-aware Prompt Inversion to train prompt space, which can learn the style information from the collection of artworks and dynamically adjusts the input images' content structure and style pattern.
- Extensive quantitative and qualitative experiments demonstrate that our proposed LSAST outperforms the state-of-the-art GAN-based and Diffusion-based methods.

2 Related Work

Generative Adversarial Network-based Methods. Generative adversarial network-based methods refer to the use of discriminator networks [Zhu *et al.*, 2017] to train a well-designed forward network that can bridge a mapping between the input image and output image. As the seminal work of artistic style transfer, Zhu *et al.* [Zhu *et al.*, 2017] expanded adversarial loss and proposed Cycle-Consistent loss to improve the quality of artistic stylized images. The cycle-consistent loss inspired a lot of researchers to explore a more effective way to enhance further the quality of artistic stylized images, including [Wang *et al.*, 2022; Park *et al.*, 2020; Zheng *et al.*, 2021; Fu *et al.*, 2019; Zuo *et al.*, 2023; Zhang *et al.*, 2023d; Zhang *et al.*, 2021]. For example, Park *et al.* [Park *et al.*, 2020] proposed a method to preserve the content structure of input content image by maximizing the mutual information between input image and output image via contrastive learning [Zhang *et al.*, 2024a]. Sanakoyeu *et al.* [Sanakoyeu *et al.*, 2018] proposed a style-aware content loss to improve the stylization of images by capturing how style affects content. Zheng *et al.* [Zheng *et al.*, 2021] exploited the spatial patterns of self-similarity to capture spatial relationships within an image rather than domain appearance to preserve the content structure. While generative adversarial network-based methods effectively transfer the learned style onto an arbitrary content image, they fail to generate highly realistic artistic stylized images, introducing obvious artifacts and disharmonious patterns.

Acknowledgments

This work was supported in part by National Social Science Foundation Major Project "Research on Virtual Restoration of Tang and Song Painting Colors and Construction of Traditional Color Resource Library from a Digital Perspective" (19ZDA046), Zhejiang Province Program (2022C01222, 2023C03199, 2023C03201), the National Program of China (62172365, 2021YFF0900604, 19ZDA197), Ningbo Science and Technology Plan Project (2022Z167, 2023Z137), and MOE Frontier Science Center for Brain Science & Brain-Machine Integration (Zhejiang University).

References

[Agarwal *et al.*, 2023] Aishwarya Agarwal, Srikrishna Karanam, Tripti Shukla, and Balaji Vasan Srinivasan. An image is worth multiple words: Multi-attribute inversion for constrained text-to-image synthesis. *arXiv preprint arXiv:2311.11919*, 2023.

[Agustsson and Timofte, 2017] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE*

conference on computer vision and pattern recognition workshops, pages 126–135, 2017.

[Chen *et al.*, 2023] Jiafu Chen, Boyan Ji, Zhanjie Zhang, Tianyi Chu, Zhiwen Zuo, Lei Zhao, Wei Xing, and Dongming Lu. Testnerf: text-driven 3d style transfer via cross-modal learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5788–5796, 2023.

[Cui *et al.*, 2022] Xuelian Cui, Zhanjie Zhang, Tao Zhang, Zhuoqun Yang, and Jie Yang. Attention graph: Learning effective visual features for large-scale image classification. *Journal of Algorithms & Computational Technology*, 16:17483026211065375, 2022.

[Fu *et al.*, 2019] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2427–2436, 2019.

[Gal *et al.*, 2022] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Hu *et al.*, 2023] Teng Hu, Jiangning Zhang, Liang Liu, Ran Yi, Siqi Kou, Haokun Zhu, Xu Chen, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Phasic content fusing diffusion model with directional distribution consistency for few-shot model adaption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2406–2415, 2023.

[Kim *et al.*, 2019] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019.

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

[Li *et al.*, 2023a] Guangyuan Li, Wei Xing, Lei Zhao, Zehua Lan, Jiakai Sun, Zhanjie Zhang, Quanwei Zhang, Huaizhong Lin, and Zhijie Lin. Self-reference image super-resolution via pre-trained diffusion large model and window adjustable transformer. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7981–7992, 2023.

[Li *et al.*, 2023b] Guangyuan Li, Wei Xing, Lei Zhao, Zehua Lan, Zhanjie Zhang, Jiakai Sun, Haolin Yin, Huaizhong Lin, and Zhijie Lin. Dudoinet: Dual-domain implicit network for multi-modality mr image arbitrary-scale super-resolution. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7335–7344, 2023.

[Li *et al.*, 2023c] Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Zhanjie Zhang, Jiafu Chen, Zhijie Lin, Huaizhong Lin, and Wei Xing. Rethinking multi-contrast mri super-resolution: Rectangle-window cross-attention transformer and arbitrary-scale upsampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21230–21240, 2023.

[Li *et al.*, 2024] Guangyuan Li, Chen Rao, Juncheng Mo, Zhanjie Zhang, Wei Xing, and Lei Zhao. Rethinking diffusion model for multi-contrast mri super-resolution. *arXiv preprint arXiv:2404.04785*, 2024.

[Nichol *et al.*, 2021] 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.

[Nichol, 2016] K. Nichol. Painter by numbers, wikiart. https://www.kaggle.com/c/painter-by-numbers, 2016. Accessed: 2016-5.

[Park *et al.*, 2020] 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.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[Radford *et al.*, 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*, pages 8748–8763. PMLR, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[Sanakoyeu *et al.*, 2018] 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.

[Sun *et al.*, 2023] Jiakai Sun, Zhanjie Zhang, Jiafu Chen, Guangyuan Li, Boyan Ji, Lei Zhao, and Wei Xing. Vgos: voxel grid optimization for view synthesis from sparse inputs. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 1414–1422, 2023.

[Sun *et al.*, 2024] Jiakai Sun, Han Jiao, Guangyuan Li, Zhanjie Zhang, Lei Zhao, and Wei Xing. 3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos. *arXiv preprint arXiv:2403.01444*, 2024.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[Wang *et al.*, 2022] Zhizhong Wang, Zhanjie Zhang, Lei Zhao, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Aesust: Towards aesthetic-enhanced universal style transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1095–1106, 2022.

[Wu, 2022] Xianchao Wu. Creative painting with latent diffusion models. *arXiv preprint arXiv:2209.14697*, 2022.

[Xie *et al.*, 2023] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023.

[Yang *et al.*, 2022] Fan Yang, Haibo Chen, Zhanjie Zhang, Lei Zhao, and Huaizhong Lin. Gating patternpyramid for diversified image style transfer. *Journal of Electronic Imaging*, 31(6):063007, 2022.

[Zhang *et al.*, 2021] Tao Zhang, Zhanjie Zhang, Wenjing Jia, Xiangjian He, and Jie Yang. Generating cartoon images from face photos with cycle-consistent adversarial networks. *Computers, Materials and Continua*, 2021.

[Zhang *et al.*, 2023a] 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.

[Zhang *et al.*, 2023b] 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.

[Zhang *et al.*, 2023c] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based creativity transfer with diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[Zhang *et al.*, 2023d] Zhanjie Zhang, Jiakai Sun, Jiafu Chen, Lei Zhao, Boyan Ji, Zehua Lan, Guangyuan Li, Wei Xing, and Duanqing Xu. Caster: Cartoon style transfer via dynamic cartoon style casting. *Neurocomputing*, 556:126654, 2023.

[Zhang *et al.*, 2024a] Zhanjie Zhang, Jiakai Sun, Guangyuan Li, Lei Zhao, Quanwei Zhang, Zehua Lan, Haolin Yin, Wei Xing, Huaizhong Lin, and Zhiwen Zuo. Rethink arbitrary style transfer with transformer and contrastive learning. *Computer Vision and Image Understanding*, page 103951, 2024.

[Zhang *et al.*, 2024b] Zhanjie Zhang, Quanwei Zhang, Guangyuan Li, Wei Xing, 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[Zheng *et al.*, 2021] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. The spatially-correlative loss for various image translation tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16407–16417, 2021.

[Zhu *et al.*, 2017] 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.

[Zuo *et al.*, 2023] Zhiwen Zuo, Lei Zhao, Ailin Li, Zhizhong Wang, Zhanjie Zhang, Jiafu Chen, Wei Xing, and Dongming Lu. Generative image inpainting with segmentation confusion adversarial training and contrastive learning. *arXiv preprint arXiv:2303.13133*, 2023.