ArtBank: Artistic Style Transfer with Pre-trained Diffusion Model and Implicit Style Prompt Bank

Zhanjie Zhang*, Quanwei Zhang*, Wei Xing†, Guangyuan Li, Lei Zhao†, Jiakai Sun, Zehua Lan, Junsheng Luan, Yiling Huang, Huaizhong Lin†

Intelligent Vision Lab, Zhejiang University
{cszzj,cszqw,cslgy,wxing,cszhl,csjk,zjucslzh,l.junsheng121,linhz}@zju.edu.cn, huangyiling@hotmail.com

Abstract

Artistic style transfer aims to repaint the content image with the learned artistic style. Existing artistic style transfer methods can be divided into two categories: small model-based approaches and pre-trained large-scale model-based approaches. Small model-based approaches can preserve the content strucuture, but fail to produce highly realistic stylized images and introduce artifacts and disharmonious patterns; Pre-trained large-scale model-based approaches can generate highly realistic stylized images but struggle with preserving the content structure. To address the above issues, we propose **ArtBank**, a novel artistic style transfer framework, to generate highly realistic stylized images while preserving the content structure of the content images. Specifically, to sufficiently dig out the knowledge embedded in pre-trained large-scale models, an **I**mplicit **S**tyle **P**rompt **B**ank (ISPB), a set of trainable parameter matrices, is designed to learn and store knowledge from the collection of artworks and behave as a visual prompt to guide pre-trained large-scale models to generate highly realistic stylized images while preserving content structure. Besides, to accelerate training the above ISPB, we propose a novel **S**patial-**S**tatistical-based self-**A**ttention **M**odule (SSAM). The qualitative and quantitative experiments demonstrate the superiority of our proposed method over state-of-the-art artistic style transfer methods. Code is available at https://github.com/Jamie-Cheung/ArtBank.

Introduction

Artistic style transfer aims to transfer the learned styles onto arbitrary content images to create a new artistic image. Existing artistic style transfer methods can be classified into small model-based methods and pre-trained large-scale model-based methods.

More specifically, small model-based methods (Zhang et al. 2021; Sanakoyeu et al. 2018; Kim et al. 2019; Park et al. 2020; Wang et al. 2022; Zhang et al. 2021; Sun et al. 2023; Yang et al. 2022; Zhang et al. 2023b; Chen et al. 2023; Zuo et al. 2023; Zhao et al. 2020; Chen et al. 2021a,b,c; Zhang et al. 2024) focus on training a well-designed forward network to learn style information from the collection of artworks. To train such forward networks, Zhu et al. (Zhu et al. 2017) first employed a cycle consistency loss to realize the mapping between the style domain and the content domain in the RGB pixel space. AST (Sanakoyeu et al. 2018) proposed a style-aware content loss to learn style from the collection of artworks for real-time and high-resolution artistic style transfer. GcGAN (Fu et al. 2019) designed a predefined geometric transformation to ensure that the stylized image maintains geometric consistency with the input content image. CUT (Park et al. 2020) used contrastive learning to push the stylized patch to appear closer to its corresponding input content patch and keep a better content structure. Based on CUT, LseSim (Zheng, Cham, and Cai 2021) introduced a more general-spatially-correlative map in contrastive learning which encourages homologous structures to be closer. ITTR (Zheng et al. 2022) utilized a transformer-based architecture to capture contextual information from different perceptions locally and globally. While these methods are capable of learning style information from a collection of artworks and preserving the content structure, they fail to generate highly realistic stylized images, and introduce disharmonious patterns and evident artifacts (e.g., 4^{th}

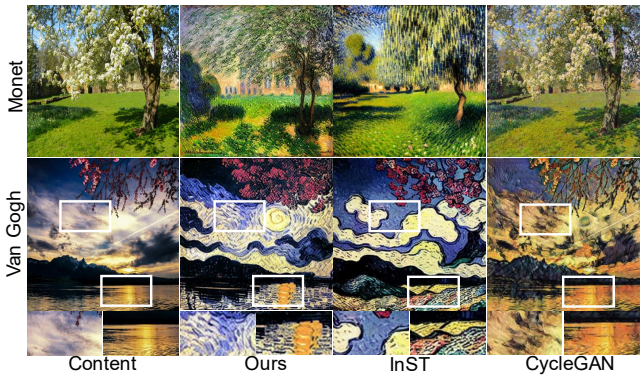

Figure 1: Stylized examples. (a) The 1^{st} column shows the content image. The 2^{nd} column shows the stylized image by our method. The other two columns show the stylized images produced by the pre-trained large-scale models (e.g., InST (Zhang et al. 2023a)) and small model-based methods (e.g., CycleGAN (Zhu et al. 2017)).

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

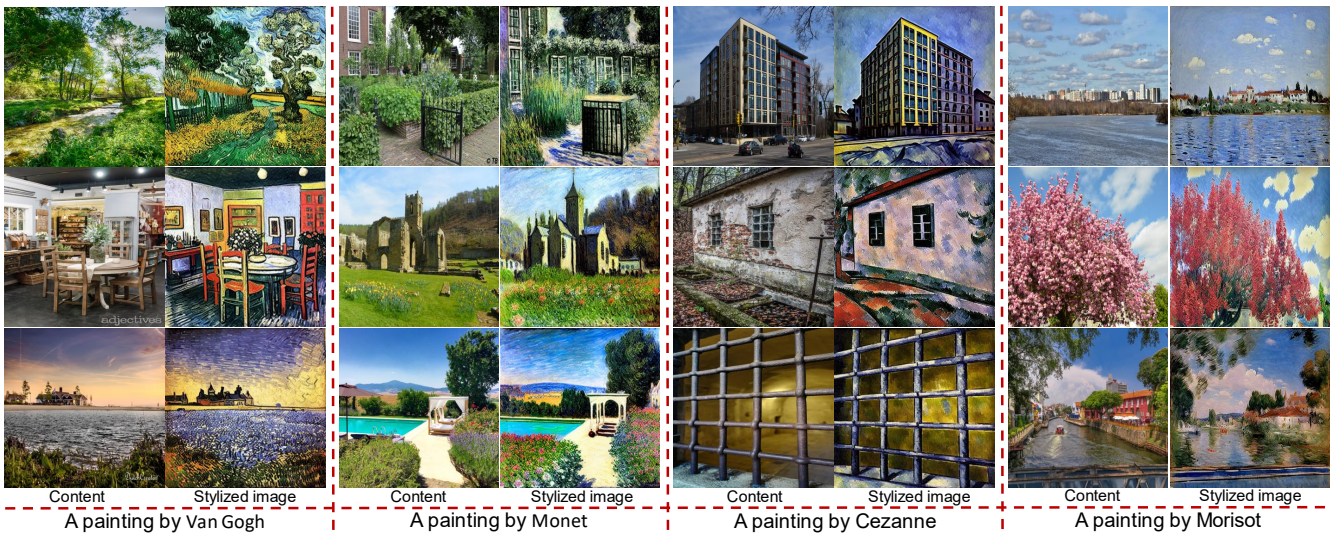

Figure 2: Stylized images generated by our proposed ArtBank. With a simple text prompt and a content image, ArtBank can generate highly realistic stylized images with preserving the structure of original content image.

col in Fig. 1).

The pre-trained large-scale model-based methods can generate highly realistic image since they are trained on large amounts of data and possess large-scale neural network parameters, which opens up the possibility to generate highly realistic stylized images. Recently, some methods (Dhariwal and Nichol 2021; Huang et al. 2022; Nichol et al. 2021; Wu 2022; Ho, Jain, and Abbeel 2020; Xie et al. 2023b,a) utilized a text prompt to synthesize highly realistic artistic images based on pre-trained large-scale model. The most representative method is Stable Diffusion (SD) (Rombach et al. 2022), which uses text prompts as guidance to generate stylized images. However, they struggle with preseving content structure. To this end, Ge et al. (Ge 2022) proposed to use a rich text editor to solve how to provide a detailed text prompt to constrain content structure; DiffuseIT (Kwon and Ye 2023) utilized a pre-trained ViT model (Tumanyan et al. 2022) to guide the generation process of DDPM models (Ho, Jain, and Abbeel 2020) in terms of preserving content structure. Zhang et al. (Zhang et al. 2023a) proposed a novel example-guided artistic image generation framework called InST related to artistic style transfer. Although these pre-trained large-scale model-based methods can generate highly realistic stylized images and attempt to preserve the content strcutre, they struggle with maintaining the structure of the original content image (e.g., 3^{rd} col in Fig. 1).

To address these problems, we focus on how to propose a more effective method that not only can generate highly realistic stylized images but also preserve the structure of the content image. Recently, the pre-trained SD has possessed the massive prior knowledge to generate highly realisc images. To exploit the prior knowledge in pre-trained SD, we first design a simple text prompt template with the artist's name (e.g., a painting by Van Gogh). Then, we use CLIP (Radford et al. 2021) to encode the text prompt template and obtain a text embedding space v_t. Next, we design an Implicit Style Prompt Bank (i.e., multiple learnable parameter matrices) that can learn and store style information from different collections of artworks. Besides, we propose Spatial and Statistical-based Self-Attention Module (SSAM) to project the learnable parameter matrix into the embedding tensor v_m. Then, v_t and v_m are concatenated, obtaining condition tensor $c_\theta(y)$. With condition tensor $c_\theta(y)$, our proposed ArtBank can fully use the prior knowledge in pre-trained large-scale models and the knowledge from the collection of artworks, generating highly realistic stylized images with preserving content structure (Please see Fig. 2 and Fig. 1). To demonstrate the effectiveness of proposed ArtBank, we build extensive experiments on different collections of artworks. All the experiments show our method outperforms the state-of-the-art artistic style transfer methods, including small model-based methods and pre-trained large-scale model-based methods. To summarize, our contributions are listed as follows:

- We propose a novel framework called ArtBank, which can generate highly realistic stylized images while preserving the content structure. This is realized by the Implicit Style Prompt Bank (ISPB), a set of trainable parameter matrices, which can learn and store the style information of multiple collections of artworks and dig out the prior knowledge of pre-trained large-scale models.

- We propose the Spatial-Statistical Self-Attention Module (SSAM), which focuses on spatial and statistical aspects, to accelerate the training of proposed ISPB.

- We have conducted extensive experiments on multiple collections of artworks and synthesized highly realistic stylized images compared to state-of-the-art methods.

Related Work

Small Model-based Methods. The small model-based method refers to training a small-scale forward neural network on a small amount of data. For example, Huang et al. (Huang and Belongie 2017) proposed an arbitrary artistic style transfer method that can transfer the style of a style image onto a content image. Li et al. (Li et al. 2017) conducted the whitening and coloring transforms (WCT) to endow the content features with the same statistical characteristics as the style features. However, these methods need a reference style image and fail to learn style information from the collection of artworks. To this end, CycleGAN (Zhu et al. 2017), DiscoGAN (Kim et al. 2017), and U-GAT-IT (Kim et al. 2019) adopt generative adversarial networks and a cycle consistency loss to realize the mapping between the style domain and content domain in the RGB pixel space. These methods can learn style information from the collection of artworks and preserve the structure of the content image, but the cycle consistency loss adds an extra computational burden. Some researchers (Sanakoyeu et al. 2018; Kim et al. 2019; Park et al. 2020) proposed to leverage the geometry consistency to preserve the structure of the content image. Although the aforementioned small model-based methods can generate stylized images with preserving the content structure, they fail to synthesize highly realistic stylized images.

Pre-trained Large-scale Model-based Methods. Large-scale models are trained on large amounts of data and can generate highly realistic images. For example, Stable Diffusion (Rombach et al. 2022) is a large-scale text-image generation model which can generate a new highly realistic image corresponding to a text prompt. Pix2pix-zero (Parmar et al. 2023) first proposed to automatically discover editing directions that reflect desired edits in the text embedding space, and condition diffusion model to generate expired image. Ramesh et al. (Ramesh et al. 2022) solve the problem of text-conditional image generation by inverted CLIP text embeddings. Zhang et al. (Zhang et al. 2023a) proposed an inversion-based artistic style transfer method to learn the corresponding textual embedding from a single image and use it as a condition to guide the generation of artistic images. DiffuseIT (Kwon and Ye 2023) utilized a pre-trained ViT model to guide the generation process of DDPM models (Ho, Jain, and Abbeel 2020) in terms of preserving content structure. Yang et al. (Yang, Hwang, and Ye 2023) proposed a zero-shot contrstive loss for diffusion models that doesn't require additional fine-tuning or auxiliary networks. These methods can perform artistic style transfer from accurate text description or examplar style image but fail to learn and store style information from the collection of artworks. Unlike these methods, our proposed approach learns style information from the collection of artworks based on the proposed ISPB. Our proposed approach does not require explicit text or images as a condition (See in Fig. 7) and can synthesize highly realistic artistic images while preserving the structure of the content image.

Method

Overview

Our proposed ArtBank includes an untrainable part (pre-trained large-scale models) and a trainable part (Implicit Style Prompt Bank, i.e., a set of learnable parameter matrices). The untrainable part utilizes a pre-trained large-scale model (Stable Diffusion, version 1.4) as a backbone, which can generate highly realistic images. The trainable part can learn and store the style information from the collection of artworks and condition pre-trained large-scale model to generate highly realistic stylized images while preserving the structure of the content image. Meanwhile, we propose the Spatial- Statistical Self-Attention Module (SSAM) to accelerate the training of ISPB. Once the training is completed, ArtBank can render arbitrary content images into highly realistic artistic stylized images while preserving the structure of the content image.

Implicit Style Prompt Bank

In order to learn the knowledge from a collection of artworks, an intuitive way is to unfreeze the parameter of pre-trained large-scale model with the following loss (Nichol and Dhariwal 2021):

$$\mathcal{L}_{diff} = \mathbb{E}_{z,x,t}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, t\right)\right\|_2^2\right], \tag{1}$$

where $z \sim E(x), \epsilon \sim \mathcal{N}(0,1)$. Once the loss function converges, the trained model can render an arbitrary content image into artistic style image. However, this naive approach will weaken pre-trained large-scale models' ability, learned from previous massive data, to generate highly realistic stylized images.

To dig out massive prior knowledge in pre-trained large-scale model and extract the knowledge from the collection of artworks, we freeze the parameter of pre-trained large-scale model and train an ISPB. ISPB comprises a series of trainable parameter matrices and each trainable parameter corresponds to a collection of artworks. The problem we need to solve is how to teach these learnable parameters to learn and store the style information from the collection of artworks and how to use these trainable parameters to condition the pre-trained large-scale model to generate highly realistic stylized images while preserving content structure. In this paper, we use pre-trained large-scale Stable Diffusion (SD) as backbone. We argue that SD relies on using CLIP-based codes, encoded by CLIP text/image encoder, to guide the denoising process and guide SD to generate desired image. CLIP text encoder can be divided into tokenizer and text transformer modules. If using the text encoder as an example, a text prompt is converted into continuous vector representations v_t. Although such v_t is effective in guiding SD to generate the desired image, it cannot fully dig out the knowledge of SD in style transfer. Based on the above analysis, we first design some coarse text prompt templates (e.g., a painting by Van Gogh *, * is only a meaningless placeholder). The coarse text prompt template is then converted into continuous vector representations: v_t and v_* (i.e., coarse text prompt is converted into v_t and $*$ is converted into v_*).

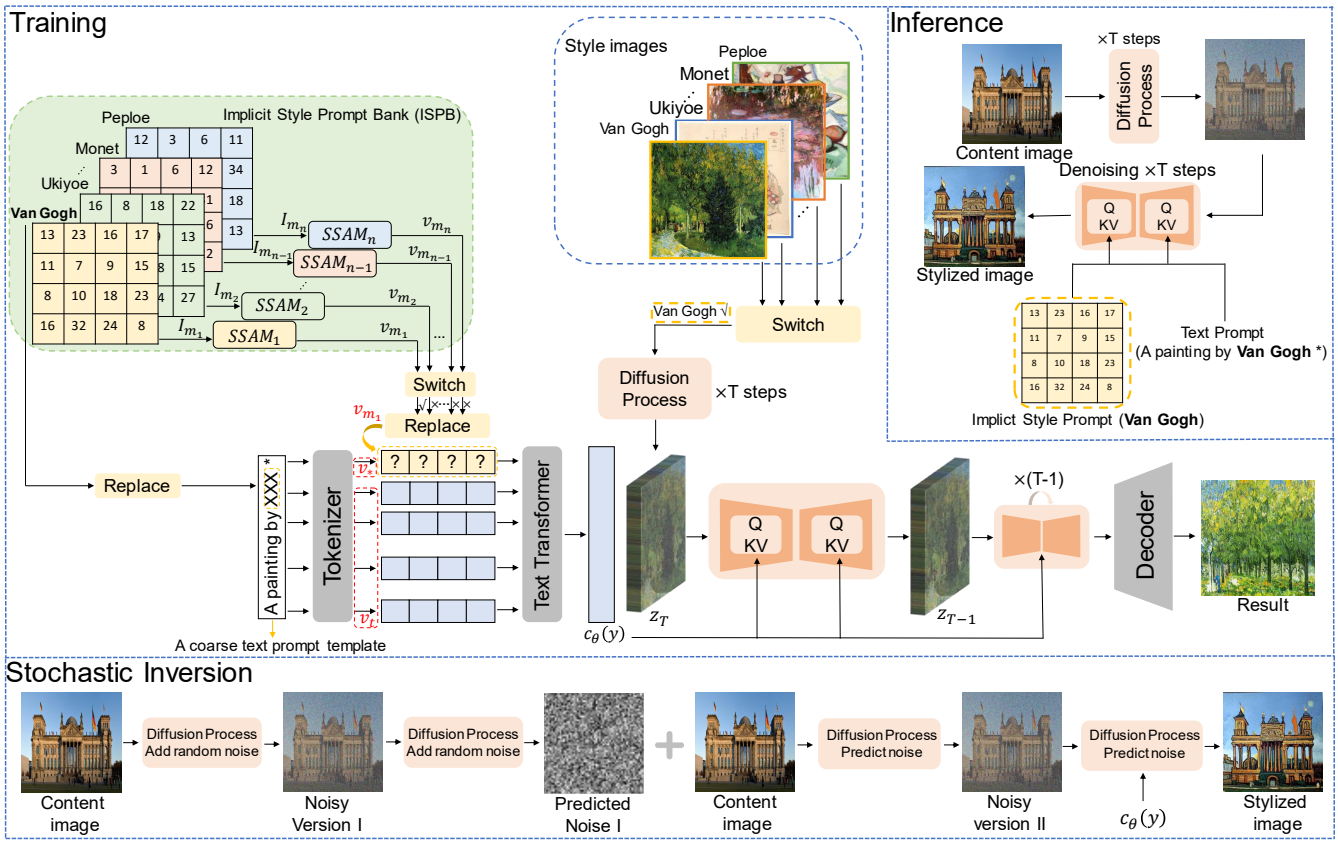

Figure 3: The overview of our proposed ArtBank which consists of two parts: an un-trainable module and a trainable module. The untrainable module is a pre-trained large-scale diffusion model (the model used in this paper is SD 1.4), which can generate highly realistic image. The trainable module is Implicit Style Prompt Bank (ISPB) which can learn and store style information from the collection of artworks. The stochastic inversion (bottom) is used in the diffusion process of inference stage (upper right).

In the meantime, the learnable parameter matrix I_m of ISPB is also projected into a continuous style representation vector v_m by our proposed SSAM (i.e., $v_m = SSAM(I_m)$. The SSAM will be illustrated in Fig. 4). Then, we replace embedding vector v_* with style representation v_m. Finally, the embedding vectors (v_t and v_m) are transformed into a single conditioning code $c_\theta(y)$. In the above process, only I_m and $SSAM$ need to be trained, and each collection of artworks need the corresponding I_{m_n} and $SSAM_n$. The $SSAM_n$ is primarily responsible for accelerating training I_{m_n}. We use the following loss function for training.

$$\mathcal{L}_{diff} = \mathbb{E}_{z,x,y,t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \text{SSAM}\left(I_m\right), v_t\right)\|_2^2\right], \quad (2)$$

where $z \sim E(x), \epsilon \sim \mathcal{N}(0,1)$ and v_t denotes text prompt. Once the I_m and $SSAM$ are trained, our proposed ArtBank supports arbitrary content images to generate highly realistic stylized images while preserving content structure.

Spatial-Statistical-Wise Self-Attention Module

Fig. 4 illustrates the architecture of our proposed Spatial-Statistical Self-Attention Module (SSAM), which differs from previous self-attention approaches (Park and Lee 2019; Liu et al. 2021; Li et al. 2023b,a,c; Cui et al. 2022). Our novel SSAM can learn and evaluate the value change of the parameter matrix from both spatial and statistical perspective. Specifically, we use row-column-wise attention from spatial aspects and mean-variance-wise attention from statistical aspects to extract parameter information. This approach can accelerate the convergence rate, reduce the volatility of parameter matrix updates, and dig out knowledge in SD. The SSAM starts with a trainable parameter matrix I_m, which is encoded into a query (Q), key (K), and value (V).

$$Q = W_Q \cdot I_m, K = W_K \cdot I_m, V = W_V \cdot I_m \quad (3)$$

where W_Q, W_K, W_V are learnable convolution layer. The attention map A can be calculated as:

$$A = Softmax\left(Q^\top \otimes K\right) \quad (4)$$

where $\otimes$ denotes matrix multiplication

For attention map A, we build col-wise weight matrix $W_{col} \in R^{H_c W_c \times 1}$ and row-wise weight matrix $W_{row} \in R^{1 \times H_c W_c}$. And to make it easier to calculate, we replicate W_{col} and W_{row} along with col and row, respectively. Then

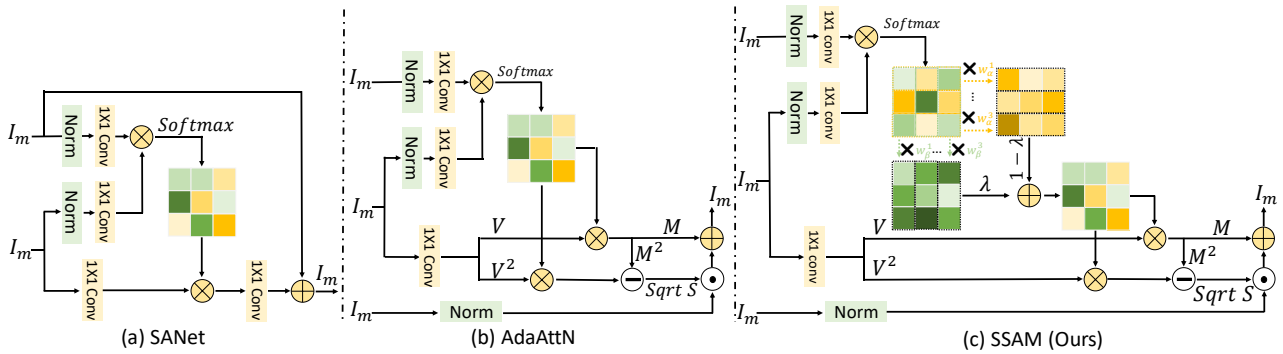

Figure 4: (a) The structure of SANet (Park and Lee 2019); (b) The structure of AdaAttN (Liu et al. 2021); (c) The structure of our proposed SSAM. Norm here denotes the mean-variance channel-wise normalization.

we can obtain col-wise and row-wise attention maps as below.

$$A_{\text{col}} = A \cdot W_{\text{col}}\,, A_{\text{row}} = A \cdot W_{\text{row}} \tag{5}$$

Then, $\hat{A} = \alpha \cdot A_{\text{col}} + (1-\alpha) \cdot A_{\text{row}}$ (i.e., α is learnable weight.). Furthermore,

$$\hat{M} = V \cdot \hat{A} \tag{6}$$

The attention-weighted standard deviation $\hat{S} \in R^{C \times H_c W_c}$ as:

$$\hat{S} = \sqrt{(V \cdot V) \otimes \hat{A}^{\top} - \hat{M} \cdot \hat{M}} \tag{7}$$

where $\cdot$ represents the element-wise product. Finally, corresponding scale $\hat{S}$ and shift $\hat{M}$ are used to generate a transformed parameter matrix as:

$$v_m = \hat{S} \cdot \text{Norm}\,(I_m) + \hat{M} \tag{8}$$

we redefine this process as: $v_m = SSAM(I_m)$.

Stochastic Inversion

In pre-trained large-scale model, random noise plays a crucial role in preserving the content structure of stylized images (Hertz et al. 2022). However, random noise is hard to predict, and incorrectly predicted noise can cause a content mismatch between the stylized image and the content image. To this end, we first add random noise to the content image and use the denoising U-Net in the diffusion model to predict the noise in the image. The predicted noise is used as the initial input noise during inference to preserve content structure (called stochastic inversion, as shown in bottom of Fig. 3). Based on this strategy, our proposed ArtBank can generate highly realistic stylized images while preserving better content structure.

Experiments

Implementation Details

We use pre-trained large-scale diffusion model (SD version 1.4) as our backbone. We train our proposed module for each collection of artworks using two NVIDIA GeForce RTX3090 GPUs. The training process requires about 200,000 iterations with a batch size of 1 and takes about two days to complete for each collection. We use a base learning rate of 0.001. The art images are chosen from the Wikiart (Nichol 2016) dataset and scaled to 512×512 pixels. The training set size varies for each class: 401 for Van Gogh, 130 for Morisot, 1433 for Ukiyoe, 1072 for Monet, 584 for Cezanne, 292 for Gauguin, and 204 for Peploe. During inference, we randomly select some content images from DIV2K (Agustsson and Timofte 2017) as the initial input images.

Qualitative Comparisons

Comparison With SOTA Style Transfer Methods. We compare our method with the sate-of-the-art artistic style transfer methods, including InST (Zhang et al. 2023a), DiffuseIT (Kwon and Ye 2023), SD (Rombach et al. 2022), AST (Sanakoyeu et al. 2018), CycleGAN (Zhu et al. 2017), LSeSim (Gao, Zhang, and Tian 2022) and CUT (Park et al. 2020). As the representative of small model-based methods, AST, CycleGAN, LSeSim and CUT can generate stylized images with better content structure; they also introduce artifacts and disharmonious patterns into stylized images. As shown in Fig. 5, as the representative of a pre-trained large-scale model, the InST is trained based on the diffusion model and can learn style information from a single style image. To make a fair comparison, we retrained InST and used the same collection of artworks and text prompts with our proposed method. In the inference, InST used the content image as the initial input image, the text prompt and the content image are used as conditional input. Fig. 5 shows that InST still has limiations in preserving content structure compared to our approach. DiffuseIT and SD have more limitations in preserving content structure.

Compared to the above methods, our proposed ArtBank can not only fully mine the knowledge in the pre-trained large-scale model but also learn and store the style information from the collection of artworks to generate highly realistic stylized images while preserving better content structure.

Quantitative Comparisons

To better demonstrate the superiority of our proposed method in artistic style transfer. We also compare our proposed method with other methods in the terms of CLIP

Figure 5: Qualitative comparisons with SOTA artistic style transfer methods.

Score, Preference Score and Timing Information.

CLIP Score. CLIP (Radford et al. 2021) is a cross-modal model pre-trained on 400M image-caption pairs and can be used for robust automatic evaluation of the accuracy between images and text prompt (Hessel et al. 2021). CLIP Score can measure the similarity between text prompt and the artistic style images. As shown in Tab. 1, "Ground Truth" denotes the similarity between text prompt and the collection of artworks. Taking the collection of artworks from Van Gogh as an example, we calculated the mean of similarity between 401 artistic images and a text (a painting by Van Gogh). We also calculate the mean of similarity between the 1,000 stylized image and a text prompt. We employed the same strategy to calculate the CLIP score for the other collection of artworks, such as Morisot, Ukiyoe, Monet, etc. As shown in Tab. 1, our method achieves a higher CLIP score compared to other state-of-the-art methods, and is even close to the ground truth score.

Timing Information. The 9^{th} row of Tab. 1 shows the run time comparisons on images with a scale of 512×512 pixels. Although our proposed method does not have the advantage of inference efficiency compared with the small method-based methods, it is significantly faster than the pre-trained large-scale model-based model.

Preference Score. (Chen et al. 2021a; Zhang et al. 2023a).To measure the popularity of stylized images generated by two artistic style transfer methods, preference score is commonly used. In this section, we randomly selected 100 content images as input for our proposed method and existing artistic style transfer methods, generating 100 stylized images. To ensure a fair and efficient calculation of the preference score, we asked each participant to select their preferred stylized image one at a time from a set of 10 images generated by our method and 10 images generated by one of the other methods. Participants were instructed to prioritize artistic authenticity and content continuity between stylized

	Ground truth	Ours	InST	SD	DiffuseIT	AST	CycleGAN	LSeSim	CUT
Van Gogh	0.7588	**0.7321**	0.7244	0.5440	0.6632	0.6736	0.6875	0.6727	0.7124
Morisot	0.8024	**0.7447**	0.6983	0.5013	0.6871	0.6659	0.7063	0.6730	0.7389
Ukiyoe	0.7495	**0.7384**	0.7272	0.5235	0.6953	0.6546	0.6504	0.6403	0.6553
Monet	0.7910	**0.7556**	0.7319	0.5031	0.6984	0.7249	0.7351	0.7125	0.7266
Cezanne	0.7760	**0.7646**	0.7332	0.5440	0.7216	0.7143	0.7363	0.7250	0.7563
Gauguin	0.8248	**0.8190**	0.7875	0.6231	0.7528	0.6839	0.7139	0.6730	0.7362
Peploe	0.7475	**0.7355**	0.7032	0.5396	0.6807	0.6677	0.6846	0.6327	0.6982
Time/sec	-	3.5725	4.0485	3.7547	32.352	0.0312	0.0312	0.0365	0.0312
Preference	-	0.679 *	**0.572**/0.428	**0.708**/0.292	**0.664**/0.336	**0.685**/0.315	**0.683**/0.317	**0.672**/0.328	**0.769**/0.231

Table 1: Quantitative comparisons with state-of-the-art methods. * denotes the average user preference.

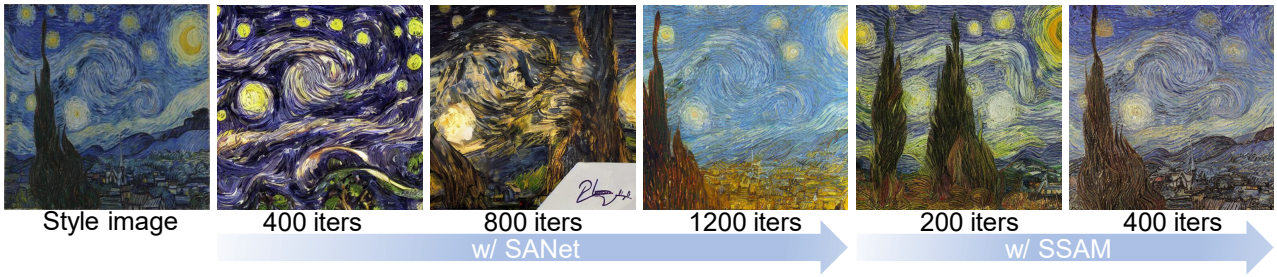

Figure 6: The optimization effiency comparison with SANet (Park and Lee 2019) and our proposed SSAM.

	Full Model	w/o Text	w/ SANet	w/ AdaAttN
Van Gogh	**0.7321**	0.7248	0.7267	0.7310
Morisot	**0.7447**	0.7383	0.7395	0.7425
Ukiyoe	**0.7384**	0.7256	0.7286	0.7288
Monet	**0.7556**	0.7419	0.7456	0.7462
Cezanne	**0.7646**	0.7532	0.7569	0.7572
Gauguin	**0.8190**	0.8025	0.8036	0.8139
Peploe	**0.7355**	0.7232	0.7285	0.7325

Table 2: The CLIP score between text prompt and stylized images.

Figure 7: The ablation study of attention module and text prompt.

images and content images. We recruited 200 participants to repeat the above process, and collected a total of 2,000 votes. Tab. 1 shows the percentage of votes for each artistic style transfer method, indicating that our proposed method was the most popular.

Ablation Study

Self-attention can faster optimization efficiency of diffusion model (Hessel et al. 2021). To demonstrate that our proposed method with the proposed SSAM can optimize the target style image faster, we show the optimization process (see in in Fig. 6) using our proposed SSAM or SANet (Park and Lee 2019). While using SANet requires 1400 iters to achieve incomplete convergence, SSAM only requires 400 iters. Besides, we retrained ISPB using SANet and AdaAttN with the same iterations. As shown in Fig. 7, we observe that stylized images with other attention mechanisms are less creative and exhibit fewer brushstroke details. The quality of stylized images also degraded in detail when the text prompt is removed. To further validate the effectiveness of our proposed module from quantitative evaluation, we also calculate CLIP score, as shown in Tab. 2.

Conclusion

We introduce a novel artistic style transfer framework, called as ArtBank, which can addresses the challenge of digging out the knowledge from pre-trained large models to generate highly realistic stylized images while preserving the content structure of the original content images. Extensive experiments demonstrate that our proposed method achieves state-of-the-art performance in artistic style transfer compared to existing SOTA methods. In the future, we lookforward to designing a more effective visual prompt to fully dig out the prior knowledge of pre-trained large-scale model in generating highly realisc stylized images.

Acknowledgments

This work was supported in part by Zhejiang Province Program (2022C01222, 2023C03199, 2023C03201, 2019007, 2021009), the National Program of China (62172365, 2021YFF0900604, 19ZDA197), Ningbo Program (022Z167), and MOE Frontier Science Center for Brain Science & Brain-Machine Integration (Zhejiang University).

References

Agustsson, E.; and Timofte, R. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 126–135.

Chen, H.; Wang, Z.; Zhang, H.; Zuo, Z.; Li, A.; Xing, W.; Lu, D.; et al. 2021a. Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems*, 34: 26561–26573.

Chen, H.; Zhao, L.; Wang, Z.; Zhang, H.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2021b. Dualast: Dual style-learning networks for artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 872–881.

Chen, H.; Zhao, L.; Zhang, H.; Wang, Z.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2021c. Diverse image style transfer via invertible cross-space mapping. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 14860–14869. IEEE Computer Society.

Chen, J.; Ji, B.; Zhang, Z.; Chu, T.; Zuo, Z.; Zhao, L.; Xing, W.; and Lu, D. 2023. TeSTNeRF: text-driven 3D style transfer via cross-modal learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 5788–5796.

Cui, X.; Zhang, Z.; Zhang, T.; Yang, Z.; and Yang, J. 2022. Attention graph: Learning effective visual features for large-scale image classification. *Journal of Algorithms & Computational Technology*, 16: 17483026211065375.

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

Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Zhang, K.; and Tao, D. 2019. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2427–2436.

Gao, X.; Zhang, Y.; and Tian, Y. 2022. Learning to Incorporate Texture Saliency Adaptive Attention to Image Cartoonization. *arXiv preprint arXiv:2208.01587*.

Ge, S. 2022. Expressive Text-to-Image Generation with Rich Text. *arXiv preprint arXiv:2304.06720*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, N.; Tang, F.; Dong, W.; and Xu, C. 2022. Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1085–1094.

Huang, X.; and Belongie, S. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 1510–1519.

Kim, J.; Kim, M.; Kang, H.; and Lee, K. 2019. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*.

Kim, T.; Cha, M.; Kim, H.; Lee, J. K.; and Kim, J. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, 1857–1865. PMLR.

Kwon, G.; and Ye, J. C. 2023. Diffusion-based Image Translation using disentangled style and content representation. In *The Eleventh International Conference on Learning Representations*.

Li, G.; Xing, W.; Zhao, L.; Lan, Z.; Sun, J.; Zhang, Z.; Zhang, Q.; Lin, H.; and Lin, Z. 2023a. 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*, 7981–7992.

Li, G.; Xing, W.; Zhao, L.; Lan, Z.; Zhang, Z.; Sun, J.; Yin, H.; Lin, H.; and Lin, Z. 2023b. DuDoINet: Dual-Domain Implicit Network for Multi-Modality MR Image Arbitrary-scale Super-Resolution. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7335–7344.

Li, G.; Zhao, L.; Sun, J.; Lan, Z.; Zhang, Z.; Chen, J.; Lin, Z.; Lin, H.; and Xing, W. 2023c. 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*, 21230–21240.

Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; and Yang, M.-H. 2017. Universal style transfer via feature transforms. In *Advances in neural information processing systems*, 386–396.

Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; and Ding, E. 2021. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6649–6658.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

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

Park, D. Y.; and Lee, K. H. 2019. Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5880–5888.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. 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*, 319–345. Springer.

Parmar, G.; Singh, K. K.; Zhang, R.; Li, Y.; Lu, J.; and Zhu, J.-Y. 2023. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Sanakoyeu, A.; Kotovenko, D.; Lang, S.; and Ommer, B. 2018. A style-aware content loss for real-time hd style transfer. In *proceedings of the European conference on computer vision (ECCV)*, 698–714.

Sun, J.; Zhang, Z.; Chen, J.; Li, G.; Ji, B.; Zhao, L.; and Xing, W. 2023. VGOS: Voxel Grid Optimization for View Synthesis from Sparse Inputs. In Elkind, E., ed., *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 1414–1422. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Tumanyan, N.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2022. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10748–10757.

Wang, Z.; Zhang, Z.; Zhao, L.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2022. AesUST: Towards Aesthetic-Enhanced Universal Style Transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1095–1106.

Wu, X. 2022. Creative painting with latent diffusion models. *arXiv preprint arXiv:2209.14697*.

Xie, J.; Li, Y.; Huang, Y.; Liu, H.; Zhang, W.; Zheng, Y.; and Shou, M. Z. 2023a. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7452–7461.

Xie, J.; Ye, K.; Li, Y.; Li, Y.; Lin, K. Q.; Zheng, Y.; Shen, L.; and Shou, M. Z. 2023b. Learning Visual Prior via Generative Pre-Training. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yang, F.; Chen, H.; Zhang, Z.; Zhao, L.; and Lin, H. 2022. Gating PatternPyramid for diversified image style transfer. *Journal of Electronic Imaging*, 31(6): 063007.

Yang, S.; Hwang, H.; and Ye, J. C. 2023. Zero-shot contrastive loss for text-guided diffusion image style transfer.

Zhang, T.; Zhang, Z.; Jia, W.; He, X.; and Yang, J. 2021. Generating cartoon images from face photos with cycle-consistent adversarial networks. *Computers, Materials and Continua*.

Zhang, Y.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; and Xu, C. 2023a. Inversion-Based Creativity Transfer with Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhang, Z.; Sun, J.; Chen, J.; Zhao, L.; Ji, B.; Lan, Z.; Li, G.; Xing, W.; and Xu, D. 2023b. Caster: Cartoon style transfer via dynamic cartoon style casting. *Neurocomputing*, 556: 126654.

Zhang, Z.; Sun, J.; Li, G.; Zhao, L.; Zhang, Q.; Lan, Z.; Yin, H.; Xing, W.; Lin, H.; and Zuo, Z. 2024. Rethink arbitrary style transfer with transformer and contrastive learning. *Computer Vision and Image Understanding*, 103951.

Zhao, L.; Mo, Q.; Lin, S.; Wang, Z.; Zuo, Z.; Chen, H.; Xing, W.; and Lu, D. 2020. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5741–5750.

Zheng, C.; Cham, T.-J.; and Cai, J. 2021. The spatially-correlative loss for various image translation tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16407–16417.

Zheng, W.; Li, Q.; Zhang, G.; Wan, P.; and Wang, Z. 2022. Ittr: Unpaired image-to-image translation with transformers. *arXiv preprint arXiv:2203.16015*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zuo, Z.; Zhao, L.; Li, A.; Wang, Z.; Zhang, Z.; Chen, J.; Xing, W.; and Lu, D. 2023. Generative Image Inpainting with Segmentation Confusion Adversarial Training and Contrastive Learning. *arXiv preprint arXiv:2303.13133*.