Inversion-based Style Transfer with Diffusion Models

Yuxin Zhang[1,2] Nisha Huang[1,2] Fan Tang[3] Haibin Huang[4]
Chongyang Ma[4] Weiming Dong[1,2*] Changsheng Xu[1,2]
[1]MAIS, Institute of Automation, Chinese Academy of Sciences [2]School of AI, UCAS
[3]Institute of Computing Technology, Chinese Academy of Sciences [4]Kuaishou Technology

Figure 1. Style transfer results using the proposed method. Given only a single input painting, our method can accurately transfer the style attributes such as semantics, material, object shape, brushstrokes and colors of the references to a natural image with a very simple learned textual description "[C]".

Abstract

The artistic style within a painting is the means of expression, which includes not only the painting material, colors, and brushstrokes, but also the high-level attributes, including semantic elements and object shapes. Previous arbitrary example-guided artistic image generation methods often fail to control shape changes or convey elements. Pre-trained text-to-image synthesis diffusion probabilistic models have achieved remarkable quality but often require extensive textual descriptions to accurately portray the attributes of a particular painting. The uniqueness of an artwork lies in the fact that it cannot be adequately explained with normal language. Our key idea is to learn the artistic style directly from a single painting and then guide the synthesis without providing complex textual descriptions. Specifically, we perceive style as a learnable textual description of a painting. We propose an inversion-based style transfer method (InST), which can efficiently and accurately learn the key information of an image, thus capturing and transferring the artistic style of a painting. We demonstrate the quality and efficiency of our method on numerous paintings of various artists and styles. Codes are available at https://github.com/zyxElsa/InST.

*Corresponding author: weiming.dong@ia.ac.cn

1. Introduction

If a photo speaks 1000 words, then every painting tells a story. A painting contains the engagement of an artist's own creation. The artistic style of a painting can be the personalized textures and brushstrokes, the portrayed beautiful moment, or some particular semantic elements. All these artistic factors are difficult to describe with words. Therefore, when we wish to utilize a favorite painting to create new digital artworks that can imitate the original idea of the artist, the task turns into example-guided artistic image generation.

Generating artistic image(s) from given example(s) has attracted great interest in recent years. A typical task is style transfer [1, 6, 8, 17, 34, 55, 59], which can create a new artistic image from an arbitrary input pair of natural images and painting image, by combining the content of the natural image and the style of the painting image. However, particular artistic factors such as object shape and semantic elements are difficult to be transferred (see Figures 2(b) and 2(e)). Text guided stylization [14, 16, 29, 38] produces an artistic image from a natural image and a text prompt, but the text prompt for the target style can usually be a rough description only of the material (e.g., "oil", "watercolor" or "sketch"), art movement(e.g. "Impressionism" or "Cubism"), artist (e.g., "Vincent van Gogh" or "Claude Monet") or a famous artwork (e.g., "Starry Night" or "The Scream").

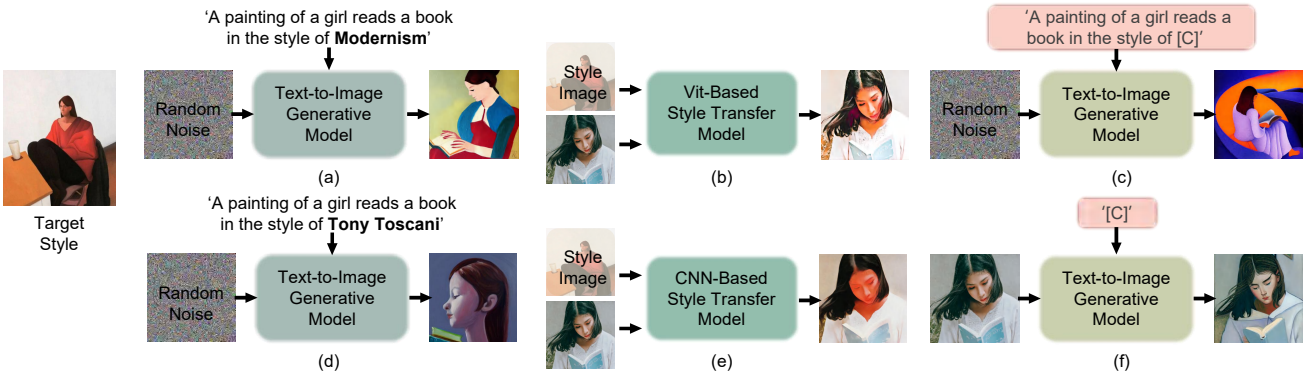

Figure 2. Concept differences between text-to-image synthesis [50], classical style transfer [7, 59] and our InST.

Diffusion-based methods [9, 22, 23, 26, 36, 54] generate high-quality and diverse artistic images on the basis of a text prompt, with or without image examples. In addition to the input image, a detailed auxiliary textual input is required to guide the generation process if we want to reproduce some vivid contents and styles. However, the creative idea of a specific painting may still be difficult to reproduce in the result.

In this paper, we propose a novel example-guided artistic image generation framework (i.e., inversion-based style transfer, InST) which related to style transfer and text-to-image synthesis, to mitigate all the above problems. Given only a single input painting image, our method can learn and transfer its style to a natural image with a very simple text prompt (see Figures 1 and 2(f)). The resulting image exhibits very similar artistic attributes to those of the original painting, including material, brushstrokes, colors, object shapes and semantic elements, without losing diversity. Furthermore, the content of the resulting image can also be controlled by providing a text description (see Figure 2(c)).

To achieve this goal, we need to obtain the representation of the image style, which refers to the set of attributes that appear in the high-level textual description of the image. We define the textual descriptions as "new words" that do not exist in the normal language and obtain the embeddings via inversion method. We exploit the recent success of diffusion models [42, 50] and inversion [2, 15]. We adapt diffusion models in our work as the backbone to be inverted and as a generator in image-to-image and text-to-image synthesis. Specifically, we propose an efficient and accurate textual inversion based on the attention mechanism, which can quickly learn key features from an image, and a stochastic inversion to maintain the semantic of the content image. We use CLIP [39] image encoder and learn key information of the image through multi-layer cross-attention. Taking an artistic image as a reference, the attention-based inversion module is fed with its image embedding and then gives its textual embedding. The diffusion models conditioning on the textual embedding can produce new images with the learned style of the reference.

To demonstrate the effectiveness of InST, we conduct comprehensive experiments on numerous images of various artists and styles. The experiments show that InST produces outstanding results, generating artistic images that imitate the style attributes to a high degree and with a content that is consistent with that of the input natural images or text descriptions. InST demonstrates much improved visual quality and artistic consistency compared with state-of-the-art approaches. These outcomes indicate the generality, precision, and adaptability of the proposed method.

2. Related Work

Image style transfer Image style transfer has been widely studied as a typical mechanism of example-guided artistic image generation. Traditional style transfer methods use low-level handcrafted features to match the patches between the content image and the style image [51, 57]. In recent years pre-trained deep convolutional neural networks have been used to extract the statistical distribution of features, which can effectively capture style patterns [18, 19, 27, 33, 52, 60]. Arbitrary style transfer methods use unified models to handle arbitrary inputs by building feed-forward architectures [8, 24, 28, 30–32, 37, 49, 53, 55, 58]. Liu et al. [34] learn the spatial attention score from both shallow and deep features using an adaptive attention normalization module (AdaAttN). An et al. [1] alleviate the content leak using reversible neural flows and an unbiased feature transfer module (ArtFlow). Chen et al. [3] apply an internal-external scheme to learn feature statistics (mean and standard deviation) as style priors (IEST). Zhang et al. [59] learn the style representation directly from image features via contrastive learning to achieve domain enhanced arbitrary style transfer (CAST). In addition to CNN, visual transformer has also been used for style transfer tasks. Wu et al. [55] perform content-guided global style composition by a transformer-driven style composition module

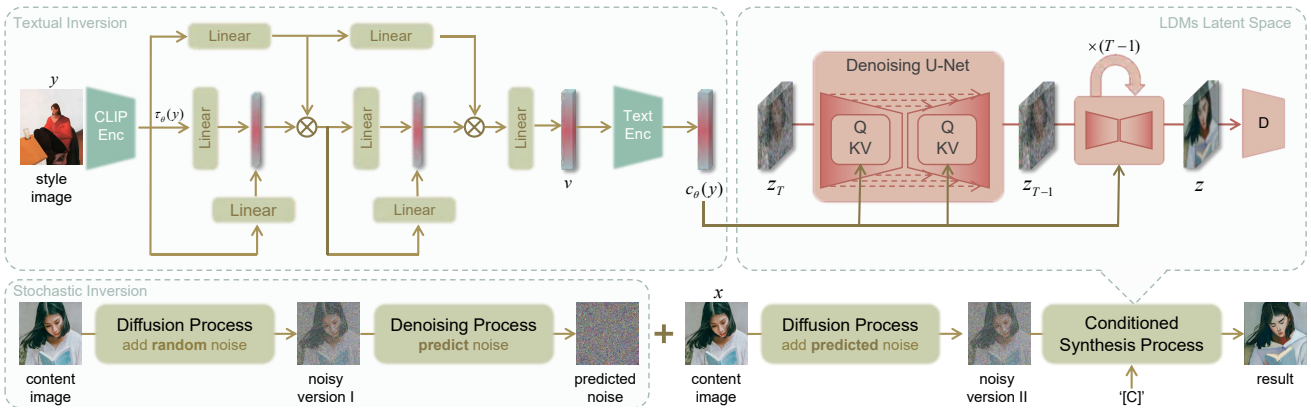

Figure 3. Overview of our InST. We apply the Stable Diffusion Models (SDMs) [42, 50] as the generative backbone and propose an attention-based textual inversion module. During image synthesis, the inversion module takes the CLIP [39] image embedding $\tau_\theta(y)$ of an artistic image y, and gives learned corresponding text embedding v, which is then encoded into the standard form of SDMs's caption conditioning $c_\theta(y)$. Through these, the SDMs can generate new images with encoded latent or random noise z_T conditioned on $c_\theta(y)$.

(StyleFormer). Deng et al. [7] propose a transformer-based method (StyTr^2) to avoid the biased content representation in style transfer by considering the long-range dependencies of input images. Image style transfer methods mainly focus on learning and transferring colors and brushstrokes and have difficulty on other artistic factors such as object shape and decorative elements.

Text-to-image synthesis Text-guided synthesis methods can also be used to generate artistic images [11, 40, 41, 45, 56]. CLIPDraw [13] synthesizes artistic images from text by using CLIP encoder [39] to maximize the similarity between the textual description and the generated drawing. VQGAN-CLIP [5] uses the CLIP-guided VQGAN [12] to generate artistic images of various styles from text prompts. Rombach et al. [42] train diffusion models [21, 47] in the latent space to reduce the complexity and generate high quality artistic images from texts. These models only use text guidance to generate an image, without fine-grained content or style control. Some methods add an image prompt to the content of the generate image to increase controllability. CLIPstyler [29] transforms the input image to a desired style with a text description by using CLIP loss and PatchCLIP loss. StyleGAN-NADA [16] uses CLIP to adaptively train the generator, which can transfer a photo to the artistic domain using the text description of the target style. Huang et al. [22] a diffusion-based artistic image generation approach by utilizing multimodal prompts as a guide to control the classifier-free diffusion model. Hertz et al. [20] change an image to artistic style by using the text prompt with style description and injecting the source attention maps. Images with complex or special artistic characteristics that cannot be described by normal texts are still difficult to generate using these methods. StyleCLIPDraw [46] jointly optimizes the text description and style image for artistic image generation. Liu et al. [35] extract the style description from CLIP model by a contrastive training strategy, which enables the network to perform style transfer between a content image and a textual style description. These methods utilize the aligned image and the text embedding of CLIP to achieve style transfer via narrowing the distance between the generated image and the style image whereas we obtain the image representation straight from the artistic image.

Inversion of diffusion models In the inversion of diffusion models, a noise map and a conditioning vector corresponding to a generated image are sought. This approach is a potential way of improving the quality of example-guided artistic image generation. However, naively adding noise to an image and then denoising it may yield an image with a significantly different content. Choi et al. [4] perform inversion by using the noised low-pass filter data from the target image as the basis for the denoising process. Dhariwal et al. [10] invert the deterministic DDIM [48] sampling process in closed form to obtain a latent noise map that will produce a given real image. Ramesh [40] develop a text-conditional image generator based on the diffusion models and the inverted CLIP. New instances of a given example are difficult to generate while maintaining fidelity using the above methods. Gal et al. [15] present a textual inversion method to find a new pseudo-word to describe the visual concept of a specific object or artistic style in the embedding space of a fixed text-to-image model. They use optimization-based methods to directly optimize the embedding of the concept. Ruiz et al. [44] implant a subject into the output domain of a text-to-image diffusion model so that it can be synthesized in novel views with a unique

identifier. Their inversion method is based on the fine-tuning of diffusion models, which requires high computational resources. Both methods learn concepts from pictures through textual inversion, and they need a small (*3-5*) image set to depict the concept. The concept they aim to learn is always an object. Our method can learn the corresponding textual embedding from a single image and use it as a condition to guide the generation of artistic images without fine-tuning the generative model.

3. Method

3.1. Overview

In this work, we use inversion as the basis of the InST framework and Stable Diffusion Models (SDMs) [42,50] as the generative backbone. Our framework is not restricted to a specific generative model. As shown in Figure 3, our method involves the pixel, latent, and textual spaces. During training, image x is the same as image y. The image embedding of image x is obtained by the CLIP image encoder and then sent to the attention-based inversion module. The key information of the image embedding is learned through multi-layer cross-attention. The inversion module provides text embedding v, which is converted into the standard format of caption conditioning SDMs. Conditioned on the input textual information, the generative model obtains a series of latent codes z_t through the sequence denoising process from to the random noise z_T and finally provides the latent code z corresponding to the artistic image. The inversion module is optimized by the simple loss of LDMs [42] computed on the "latent noise" of the forward process and the reverse process (see Sec. 3.2). In the inference process, x is the content image, and y is the reference image. The textual embedding v of reference image y guides the generative model in generating a new artistic image.

3.2. Textual Inversion

We aim to obtain the intermediate representation of a pre-trained text-to-image model for a specific painting. SDMs utilize CLIP text embedding as the condition in text-to-image generation. CLIP text encoding contains two processes, namely, tokenization and parameterization. An input text is first transformed into a token, which is an index in a pre-defined dictionary, for each word or sub-word. Then, each token is associated with a distinct embedding vector that can be located using an index. We set the concept of a picture as a placeholder "[C]", and its corresponding tokenized text embedding as a learnable vector $\hat{v}$. $[C]$ is in the normal language domain, and $\hat{v}$ is in the textual space. By assuming a $[C]$ that does not exist in real language, we create a "new word" for a certain artistic image that cannot be expressed in normal language. To obtain $\hat{v}$, we need to design constraints as supervision that relies on a single image. An instinctive way to learn $\hat{v}$ is by direct optimization [15], which minimizes the LDMs loss of a single image:

$$\hat{v} = \arg\min_{v} \mathbb{E}_{z,x,y,t} \left[\|\epsilon - \epsilon_\theta \left(z_t, t, v_\theta(y) \right)\|_2^2 \right], \quad (1)$$

where y denotes the artistic image, $v_\theta(y)$ is a learnable vector, $z \sim E(x), \epsilon \sim \mathcal{N}(0,1)$. However, this optimization-based approach is inefficient, and accurate embeddings are difficult to obtain without overfitting with a single image as training data.

We propose a learning method based on multi-layer cross attention. The artistic image input is first sent to the CLIP image encoder and provides image embeddings. By performing multi-layer attention on these image embeddings, the key information of the image can be quickly obtained. The image encoder τ_θ projects y to an image embedding $\tau_\theta(y)$. The multi-layer cross attention starts with $v_0 = \tau_\theta(y)$. Then each layer implements $\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V$ as follows:

$$\begin{gathered} Q_i = W_Q^{(i)} \cdot v_i, K = W_K^{(i)} \cdot \tau_\theta(y), V = W_V^{(i)} \cdot \tau_\theta(y), \\ v_{i+1} = \mathrm{Attention}(Q_i, K, V). \end{gathered} \quad (2)$$

During training, the model is conditioned by the corresponding text embedding only. To avoid overfitting, we apply a dropout strategy in each cross-attention layer, which is set to 0.05.

Our optimization goal can finally be defined as follows:

$$\hat{v} = \arg\min_{v} \mathbb{E}_{z,x,y,t} \left[\|\epsilon - \epsilon_\theta \left(z_t, t, \mathrm{MultiAtt}(\tau_\theta(y)) \right)\|_2^2 \right], \quad (3)$$

where $z \sim E(x), \epsilon \sim \mathcal{N}(0,1)$. τ_θ and ϵ_θ are fixed during training. In this manner, $\hat{v}$ is efficiently optimized to the target area. Note that, it is not strictly necessary to use CLIP image embedding when the model is optimized on a single image, we aim to utilize the high-quality feature space of CLIP to pave the way for diverse tasks. The key of our approach is the multi-layer attention mechanism.

3.3. Stochastic Inversion

In addition to the text description, the random noise controlled by the random seed is also important for the representation of the image. As demonstrated in [20], the changes in random seed result in obvious changes in visual differences. We divide pre-trained text-to-image diffusion model-based image representation into two parts: holistic representation and detail representation. Holistic representation involves text conditions, and the detail representation is controlled by random noise. We define the process from an image to noise maps as an inversion problem and propose stochastic inversion to preserve the semantics of the content image. We first add random noise to the content image and

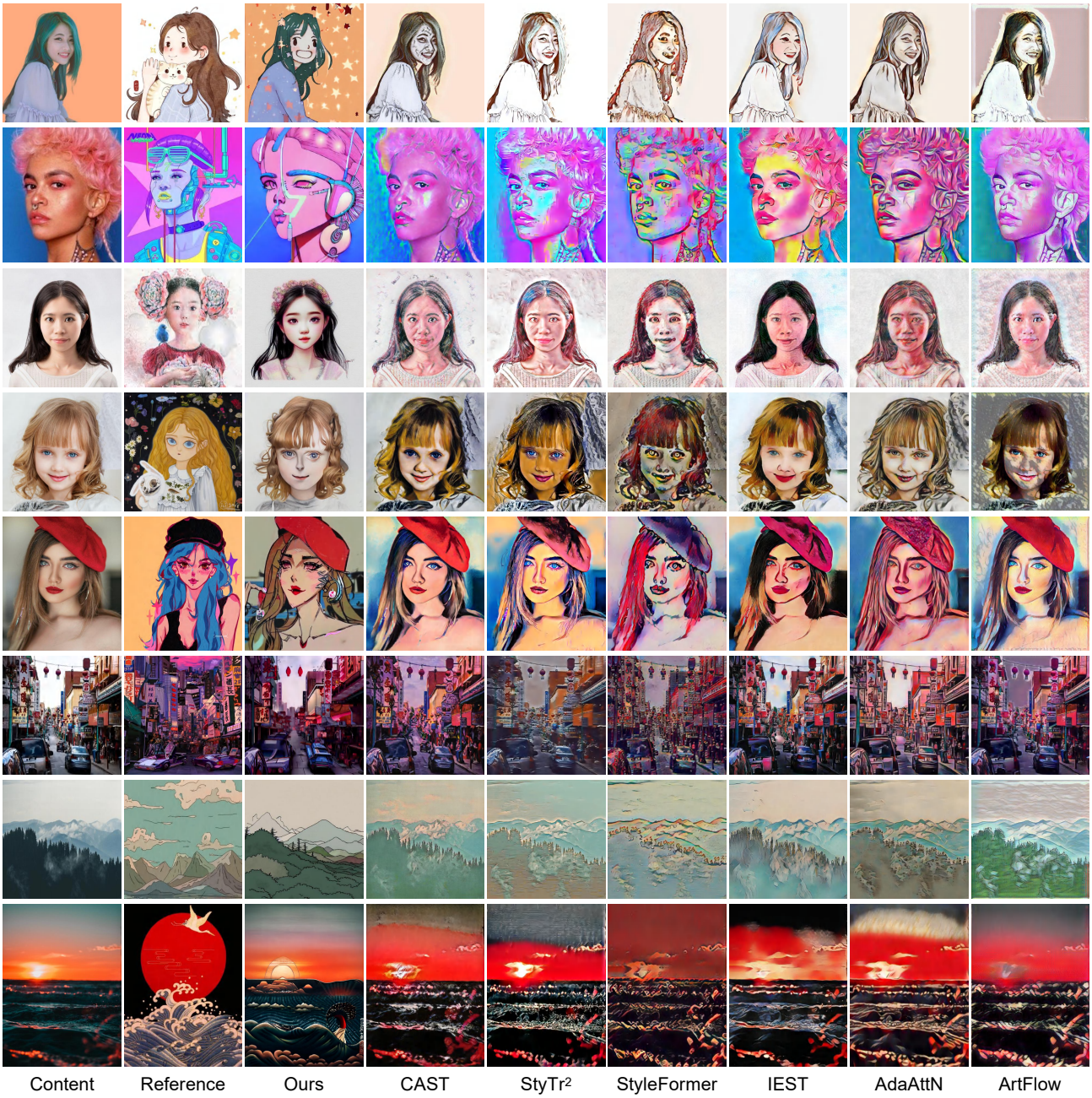

Figure 4. Qualitative comparison with several state-of-the-art image style transfer methods.

then use the denoising U-Net [43] in the diffusion model to predict the noise in the image. The predicted noise is used as the initial input noise during generation to preserve the content. Specifically, for each image z, the stochastic inversion module takes the image latent code $z = E(y)$ as input. Then z_t, the noisy version of z, is set as computable parameters, and ϵ_t is obtained as follows:

$$\hat{\epsilon}_t = (z_{t-1} - \mu_T(z_t, t))\sigma_t. \tag{4}$$

We illustrate the stochastic inversion in Figure 3.

4. Experiments

In this section, we provide visual comparisons and applications to demonstrate the effectiveness of the proposed approach.

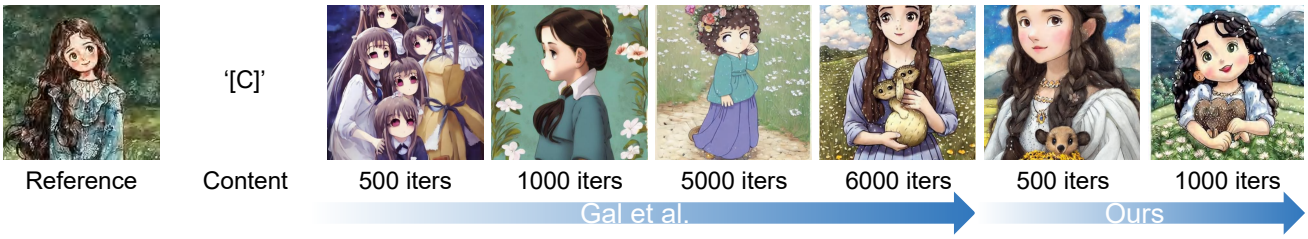

Figure 5. Qualitative comparison with textual inversion [15]. Our inversion based on the attention mechanism can converge fast to the text feature space region corresponding to the artistic image.

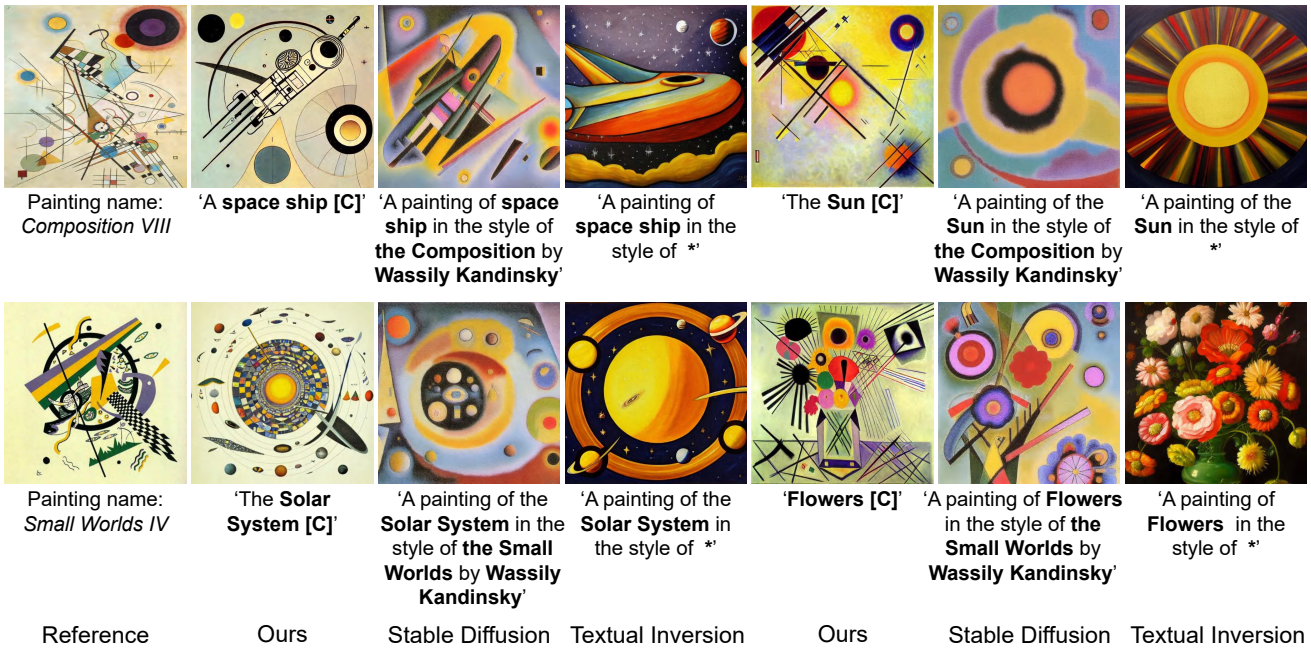

Figure 6. Qualitative comparison with textual inversion [15] and SDMs [42] in text-to-image generation. The content of each image depends on the caption below. "[C]" refers to our learned description of the corresponding painting image shown in the left column. "*" refers to the pseudo word optimized by textual inversion [15]. InST performs well in terms of editability. Our method can interact with additional text to generate artistic images with corresponding styles. However, when [15] is trained based on a single image, it is seriously affected by the additional text and cannot maintain the style.

Implementation details We retain the original hyper-parameter choices of SDMs. The training process takes approximately 20 minutes on each image in NVIDIA GeForce RTX3090 with a batch size of 1. The base learning rate is set to 0.001. The synthesis process takes the same time as SDMs, which depends on the steps.

4.1. Comparison with Style Transfer Methods

We compare our method with the state-of-the-art image style transfer methods, including ArtFlow [1], AdaAttN [34], StyleFormer [55], IEST [3], StyTr^2 [7] and CAST [59] to show the effectiveness of our method. The results show the apparent advantages of our method over traditional style transfer methods in transferring the semantics and artistic techniques of the reference images to the content images. For example, our method can transfer the shapes of important objects, such as the facial forms and eyes (the 1st, 3rd, 4th and 5th rows), the mountain (the 7th row) and the sun (the 8th row) better than the other methods. Our method can capture some special semantics of the reference images and reproduce the visual effects in the results, such as the stars on the background (the 1st row), the flower headwear (the 3rd row), and the roadsters (the 6th row, the cars in the content image are changed to roadsters). These effects are very difficult to achieve using traditional style transfer methods.

4.2. Comparison with Text-Guided Methods

Guided by a human caption, we compare our method with TexIn [15] and SDMs [42]. Following TexIn [15], we measure *accuracy* and *editability* using the similarities be-

Figure 7. Qualitative comparison with SDM [42] conditioned on human captions. Our method can accurately represent the target style image, while describing artistic attributes of a specific painting in words for SDM is more difficult.

Table 1. CLIP-based evaluations.

	Ours	TexIn [15]	SDM [42]	MLP	w/o drop
Accuracy(%)↑	**78.92**	68.84	63.83	71.03	72.25
Editability(%)↑	**80.72**	69.03	-	66.99	66.53

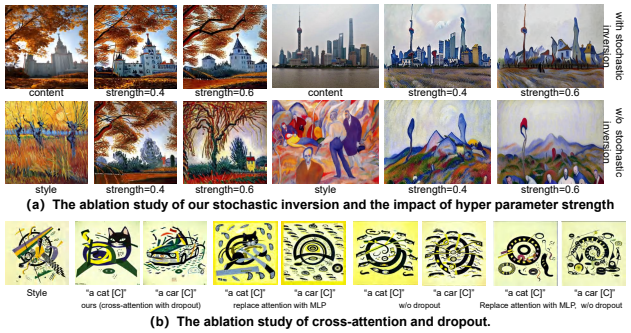

Figure 8. (a) Results of the ablation study on our stochastic inversion, hyper-parameter *Strength*, (b) attention module and dropout.

tween the CLIP embeddings of style and generated images and those of guide texts and generated images, respectively.

Comparison with Textual Inversion We begin by demonstrating the effectiveness of our attention-based inversion in learning and transferring style. In Figure 5, we show the optimization process of TexIn and ours. Our method can quickly optimize to the target text embedding in approximately 1000 iterations, while textual inversion usually takes 10 times the number of iterations using InST due to its simple optimization-based scheme. In Figure 6, we demonstrate our method's superior generality and editability by giving additional semantic descriptions that do not appear in the reference image. Our method is more robust to additional descriptions than textual inversion and can generate results that match the textual description and the reference image. However, textual inversion loses adaptability to these texts and cannot depict the specific artistic visual effect. As shown in Table. 1, our method outperforms TexIn in both *accuracy* and *editability*.

Comparison with SDMs We compare InST with the state-of-the-art text-to-image generative model SDMs [42]. SDMs can generate high-quality images from text descriptions. However, the style of a specific painting with only text as a condition is difficult to describe, so satisfactory results cannot be obtained. As shown in Figures 6 and 7, our method captures the unique artistic attributes of the reference images better than SDMs. As shown in Table 1, our method outperforms SDMs in *accuracy*.

4.3. Ablation Study

Stochastic inversion As shown in Figure 8(a), the buildings in the content images are turned into trees or mountains without stochastic inversion, while the full model can maintain the content information and reduce the impact of the style image's semantic.

Hyper-parameter *Strength* For image synthesis, the most related hyper-parameter is the *strength* of changes and its impacts are shown in Figure 8(a). The larger the *Strength*, the stronger the influence of the style image on the generated result, vice versa, the generated image is closer to the content image

Attention module We show the ablation study of multi-layer attention in Figures 8(b). The multi-layer attention helps the content of the generated image be better controlled by the input text conditions and improves *editability*, as shown in Table. 1.

Dropout Dropout is added in the linear layer of the attention module to prevent overfitting. As shown in Figure 8(b) and Table 1, by dropping the parameters of the latent embeddings, both the *accuracy* and the *editability* are improved.

4.4. User Study

We compare our method with several SOTA image style transfer methods (i.e., ArtFlow [1], AdaAttN [34], StyleFormer [55], IEST [3], StyTr^2 [7], and CAST [59]), and a text-to-image generation method (i.e., TexIn [15]). All the baselines are trained using publicly available implementations with default configurations.

For each participant, 26 content–reference pairs were

Table 2. Quantitative evaluation. The results show the average percentage of cases in which the result of the corresponding method is preferred compared with ours. The best results are in **bold**.

	CAST [59]	StyTr^2 [7]	StyleFormer [55]	IEST [3]	AdaAttN [34]	ArtFlow [1]	TexIn [15]	
							img2img	txt2img
Preference↑	0.368	0.310	0.218	0.161	0.310	0.276	0.379	0.121
Ours	**0.632**	**0.690**	**0.782**	**0.839**	**0.690**	**0.724**	**0.621**	**0.879**

Table 3. Importance ranking results of the visual factors (rank by $score$ from highest to lowest).

Visual factors	$score$
Similar artistic effect on semantic corresponding subjects	5.40
With the same paint material	3.65
Having similar brushstrokes	3.20
Having typical shapes	2.65
With the same decorative elements	2.10
Sharing the same color	1.40

randomly selected, and the generated results using InST and one of the other methods are displayed randomly. The participants were informed that the artistic consistency between the generated and reference images is the main metric. Then, they were invited to select which result for each content–reference pair is better. Finally, we collected 2,262 votes from 87 participants. The percentage of votes for each approach is shown in Table 2, demonstrating that our method achieves the best visual characteristic transfer results.

Furthermore, we conducted a survey of 60 participants on their preferences in content image guidance strength and artistic visual effects. In the case where a content image exists, users tended to consider that "To depict the artistic style, the details of the content should be embellished appropriately." We then invited the participants to rank the factors of their expected visual effects. The average comprehensive score of the options in the sorting question was automatically calculated based on the ranking of the options by all the participants. The higher the score is, the higher the comprehensive ranking is. The scoring rule is:

$$score = \frac{(\sum frequency \times weight)}{participantes}, \tag{5}$$

where $score$ denotes the average comprehensive score of the options, $participantes$ denotes the number of people who answered this question, $frequency$ denotes the frequency that the option is selected by users, and $weight$ denotes the weight which is determined by the option's ranking. The results are shown in Table 3.

4.5. Discussions and Limitations

Although our method can transfer typical colors to some extent, when a significant difference exists between the colors of the content and reference images, our method may fail to semantically transfer the color in a one-to-one correspondence. For example, the green hair of the content images in the 1st row of Figure 4 is not transferred into brown. As shown in Figure 9, we employ an additional tone transfer module [25] to align the color of content and reference images. However, different users have different preferences on retaining the colors of the content image. We believe that the colors of a photograph are crucial, so we opt to respect the tone of the original content image in some conditions.

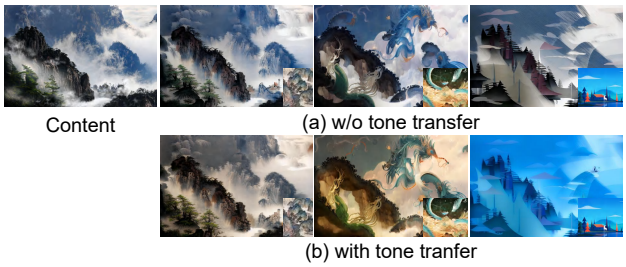

Figure 9. For different art forms, the tone of natural images and of artistic images have their own advantages.

5. Conclusion

We introduce a novel example-guided artistic image generation framework called InST, which refers to learning the high-level textual descriptions of a single painting image and then guiding the text-to-image generative model in creating images of specific artistic appearance. We propose an attention-based textual inversion method to invert a painting into the corresponding textual embeddings. The extensive experimental results demonstrate that our method achieves superior image-to-image and text-to-image generation results compared with state-of-the-art approaches. Our approach is intended to pave the way for upcoming unique artistic image synthesis tasks.

Acknowledgment This work was supported in part by National Key R&D Program of China under no. 2020AAA0106200, by National Natural Science Foundation of China under nos. 61832016, U20B2070, and 62102162, and in part by Beijing Natural Science Foundation under no. L221013.

References

[1] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. ArtFlow: Unbiased image style transfer via reversible neural flows. In *IEEE/CVF Conferences on Computer Vision and Pattern Recognition (CVPR)*, pages 862–871, 2021. 1, 2, 6, 7, 8

[2] Hila Chefer, Sagie Benaim, Roni Paiss, and Lior Wolf. Image-based clip-guided essence transfer. In *European Conference on Computer Vision (ECCV)*, pages 695–711. Springer, 2022. 2

[3] Haibo Chen, Lei Zhao, Zhizhong Wang, Zhang Hui Ming, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Artistic style transfer with internal-external learning and contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 6, 7, 8

[4] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14347–14356, 2021. 3

[5] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. VQGAN-CLIP: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision (ECCV)*, pages 88–105. Springer, 2022. 3

[6] Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. Arbitrary video style transfer via multi-channel correlation. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1210–1217, 2021. 1

[7] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. StyTr^2: Image style transfer with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11326–11336, 2022. 2, 3, 6, 7, 8

[8] Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. Arbitrary style transfer via multi-adaptation network. In *ACM International Conference on Multimedia*, pages 2719–2727, 2020. 1, 2

[9] Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances Neural Information Processing Systems (NeurIPS)*, pages 8780–8794, 2021. 2

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:8780–8794, 2021. 3

[11] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. CogView2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 3

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, 2021. 3

[13] Kevin Frans, Lisa Soros, and Olaf Witkowski. CLIPDraw: Exploring text-to-drawing synthesis through language-image encoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[14] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven artistic style transfer. In *European Conference on Computer Vision (ECCV)*, pages 717–734, 2022. 1

[15] 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. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 4, 6, 7, 8

[16] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics*, 41(4):141:1–141:13, 2022. 1, 3

[17] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE/CVF Conferences on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. 1

[18] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. 2

[19] Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3730–3738, 2017. 2

[20] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3, 4

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 3

[22] Nisha Huang, Fan Tang, Weiming Dong, and Changsheng Xu. Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion. In *ACM International Conference on Multimedia*, page 1085–1094, 2022. 2, 3

[23] Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Yong Zhang, Weiming Dong, and Changsheng Xu. DiffStyler: Controllable dual diffusion for text-driven image stylization. *arXiv preprint arXiv:2211.10682*, 2022. 2

[24] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 2

[25] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1501–1510, 2017. 8

[26] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 2

[27] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10043–10052, 2019. 2

[28] Xiaoyu Kong, Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Yongyong Chen, Zhenyu He, and Changsheng Xu. Exploring the temporal consistency of arbitrary style transfer: A channelwise perspective. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2023. 2

[29] Gihyun Kwon and Jong Chul Ye. CLIPstyler: Image style transfer with a single text condition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 18062–18071, 2022. 1, 3

[30] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3804–3812, 2019. 2

[31] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances Neural Information Processing Systems (NeurIPS)*, pages 386–396, 2017. 2

[32] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *ACM Transactions on Graphics*, 36(4), 2017. 2

[33] Minxuan Lin, Fan Tang, Weiming Dong, Xiao Li, Changsheng Xu, and Chongyang Ma. Distribution aligned multimodal and multi-domain image stylization. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 17(3):96:1–96:17, jul 2021. 2

[34] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. AdaAttN: Revisit attention mechanism in arbitrary neural style transfer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6649–6658, 2021. 1, 2, 6, 7, 8

[35] Zhi-Song Liu, Li-Wen Wang, Wan-Chi Siu, and Vicky Kalogeiton. Name your style: An arbitrary artist-aware image style transfer. *arXiv preprint arXiv:2202.13562*, 2022. 3

[36] 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. In *International Conference on Machine Learning (ICML)*, 2022. 2

[37] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5880–5888, 2019. 2

[38] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of stylegan imagery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, 2021. 1

[39] 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 (ICML)*, pages 8748–8763, 2021. 2, 3

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[41] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3, 4, 6, 7

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015. 5

[44] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[46] Peter Schaldenbrand, Zhixuan Liu, and Jean Oh. StyleCLIPDraw: Coupling content and style in text-to-drawing translation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4966–4972. International Joint Conferences on Artificial Intelligence Organization, 7 2022. 3

[47] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265. PMLR, 2015. 3

[48] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[49] Jan Svoboda, Asha Anoosheh, Christian Osendorfer, and Jonathan Masci. Two-stage peer-regularized feature recombination for arbitrary image style transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13816–13825, 2020. 2

[50] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 2, 3, 4

[51] Bin Wang, Wenping Wang, Huaiping Yang, and Jiaguang Sun. Efficient example-based painting and synthesis of 2D

directional texture. *IEEE Transactions on Visualization and Computer Graphics*, 10(3):266–277, 2004. 2

[52] Qian Wang, Cai Guo, Hong-Ning Dai, and Ping Li. Stroke-gan painter: Learning to paint artworks using stroke-style generative adversarial networks. *Computational Visual Media*, Mar 2023. 2

[53] Huapeng Wei, Yingying Deng, Fan Tang, Xingjia Pan, and Weiming Dong. A comparative study of cnn- and transformer-based visual style transfer. *Journal of Computer Science and Technology*, 37(3):601–614, 2022. 2

[54] Xianchao Wu. Creative painting with latent diffusion models. *arXiv preprint arXiv:2209.14697*, 2022. 2

[55] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. StyleFormer: Real-time arbitrary style transfer via parametric style composition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14598–14607, 2021. 1, 2, 6, 7, 8

[56] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 3

[57] Wei Zhang, Chen Cao, Shifeng Chen, Jianzhuang Liu, and Xiaoou Tang. Style transfer via image component analysis. *IEEE Transactions on Multimedia*, 15(7):1594–1601, 2013. 2

[58] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8035–8045, June 2022. 2

[59] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 12:1–12:8, New York, NY, USA, 2022. Association for Computing Machinery. 1, 2, 6, 7, 8

[60] Yuxin Zhang, Fan Tang, Weiming Dong, Thi-Ngoc-Hanh Le, Changsheng Xu, and Tong-Yee Lee. Portrait map art generation by asymmetric image-to-image translation. *Leonardo*, 56(1):28–36, feb 2023. 2