Diffusion-based Style Transfer Guided by Existing Image

Zixu Lin 23320231154449[1], Yixin Chen 23320231154443[1], Renjie Zhao 36920231153266[2], Senmao Cheng 36920231153186[2], Yingtong Gan 36920231153190[2],

[1]School of Informatics, Xiamen University
[2]Artificial Intelligence Research Institute, Xiamen University
{23320231154449, 23320231154443, 36920231153266, 36920231153186, 36920231153190}@stu.xmu.edu.cn

Abstract

The objective of image style transfer is to generate an image that preserves the original content while incorporating artistic elements inspired by a reference style. Existing neural style transfer methods require a large style dataset for training. The transferable styles are limited by the presence of the style within a style library. Recently, Text-driven diffusion models have been widely used in the fields of style transfer and image edit due to its powerful generation capabilities. However, diffusion models edit the image by modifying its corresponding prompt. In many practical situations, users may want to transfer style using existing style images, which may lack the corresponding prompt and include unseen styles that not in the style library. In this paper, we proposed a training-free framework that enables a style transfer without any large dataset. Guided by the provided style image, the method we proposed can effectively transfer images into the reference style, not just limited to a few styles. Finally, the method we proposed demonstrates excellent performance in generating results and style transfer.

Introduction

Style transfer is an important task in which the style of a source image is mapped onto that of a target image. It holds immense significance in the realms of art, design, and visual communication, enabling artists and designers to explore diverse creative expressions by blending different visual styles. Automated style transfer software facilitates the conversion of real-world images into the appropriate style to form the background in cartoons, simulations, and other renderings. The method is also useful for generating derivative works of a particular artist or painting. Meanwhile, it is used in our daily life, such as various filters in the beauty tool, including the conversion between real people and secondary characters (Song et al. 2021), and the transfer of makeup (Li et al. 2018). It is also often used as an aid to improve the performance of other computer vision tasks, such as Pedestrian re-identification (Deng et al. 2018).

The seminal work (Gatys, Ecker, and Bethge 2016) proposed an image optimization method that iteratively minimizes the joint content and style loss in the feature space of a pre-trained deep neural network. This time-consuming optimization process has motivated researchers to explore more efficient approaches. AdaIN (Huang and Belongie 2017) transfers global mean and variance of a style image to a content image in the feature space to support arbitrary input style image. To enhance the locality awareness of arbitrary style transfer models, recently, attention mechanism is adopted in multiple works (Park and Lee 2019) for this task. Their common intuition is that a model should pay more attention to those feature-similar areas in the style image for stylizing a content image region. Unfortunately, it fails to totally solve this problem and the local distortions still occur. Diffusion-based methods (Dhariwal and Nichol 2021; Huang et al. 2022; Kawar et al. 2022) generate high-quality and diverse artistic images based on 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, which may be still difficult to reproduce the key idea of a specific painting in the result.

Through literature research, we found that existing style transfer methods often only allow transfer to a few fixed styles and cannot accommodate arbitrary style selection. Therefore, we proposed a method that eliminates the constraints imposed by fixed styles, enabling the generation of images with the desired style specified by the using arbitrary style images. Given style image I_s and content image I_c to be transferred, we obtain the latent codes of two images and concatenate them for denoising, thereby leveraging the semantic information from the style image more effectively. During the generation process of the text-driven diffusion model, we employ a well-designed prompt to guide the style generation towards convergence with the style image.

The contributions of our work are as follows:

- We proposed a training-free method that eliminates the need for any style dataset, which avoids incurring costly training or fine-tuning. It allows the transfer not just limited to a few styles. Moreover, it possesses the capability to undertake style transfer tasks across a wide range of diverse domains.
- By utilizing image inversion, we obtain latent representations corresponding to style images I_s and content images I_c, which greatly enhance our ability to perform image reconstruction tasks. This approach allows us to

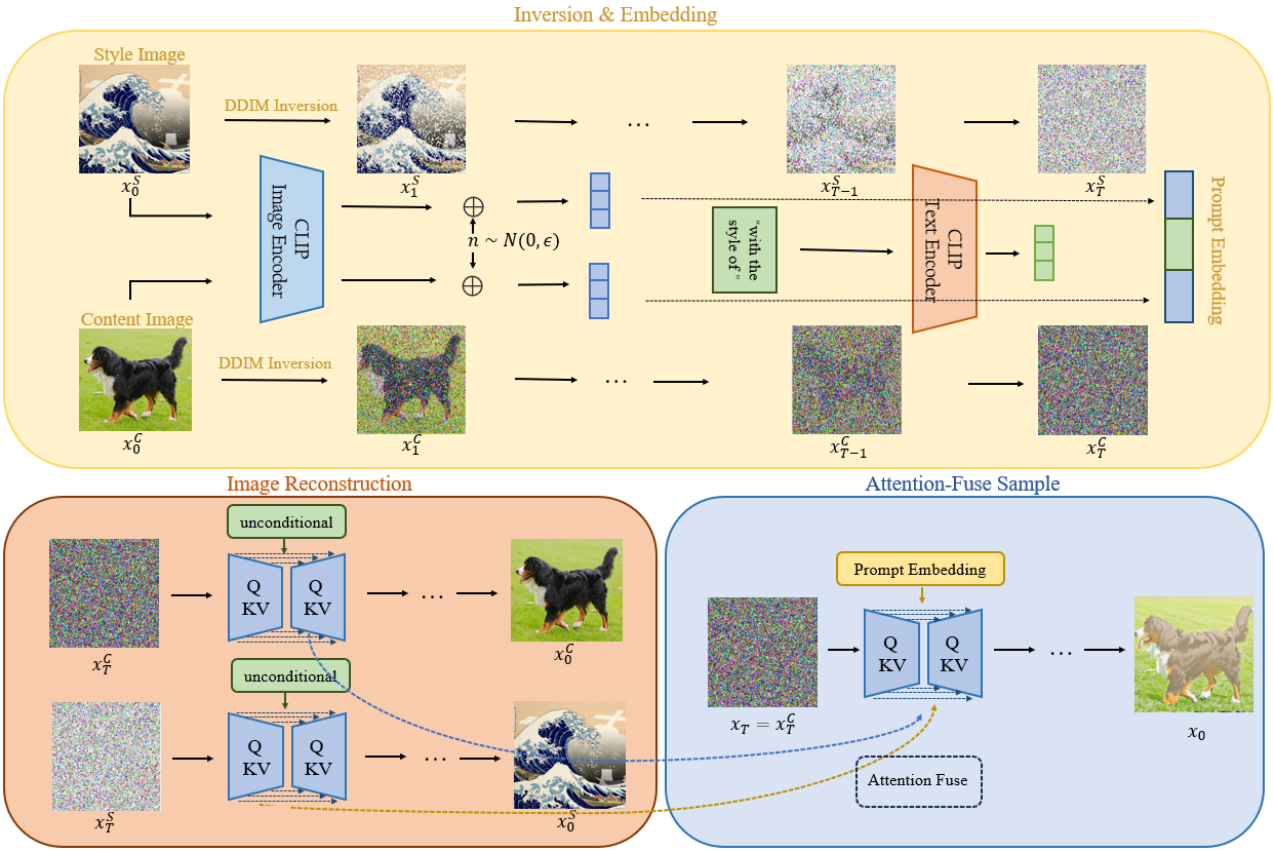

Figure 1: Overview of our method, we use DDIM inversion to obtain noise representations and apply noise injection to bridge the domain gap between image and text features(yellow area). Then unconditional sampling reconstructs images, and content image noise is sampled (blue area) guided by prompt embedding and attention fusion from unconditional samples(orange area).

leverage the abundant semantic knowledge present in the style images.

- We propose a novel prompt design method that effectively guides the generation of content image I_c towards the direction of the style image I_s.

Related Work

Style Transfer

Gatys et al. (Gatys, Ecker, and Bethge 2016)find that hierarchical layers in CNNs can be used to extract image content structures and style texture information and propose an optimization-based method to generate stylized images iteratively. More generally, arbitrary style transfer gains more attention in recent years. Huang et al. (Huang and Belongie 2017) propose an adaptive instance normalization (AdaIN) to replace the mean and variance of content with that of style. AdaIN is widely adopted in image generation tasks (An et al. 2020; Wang et al. 2020) to fuse the content and style features. (Lu, Liu, and Kong 2023)employs a transformer-inspired approach with three key modules: style bank generation for compact style pattern extraction, transformer-driven style composition for content-aware global styling, and parametric content modulation for flexible stylization.Synthesis-based approaches often require large amounts of training data and large models, which limits their usability in real-world applications.

Diffusion Models

Diffusion models have been widely used in image synthesis during the past few years, leading to impressive advancements in the area (Ho, Jain, and Abbeel 2020). Thanks to an iterative denoising approach, this model has shown to be quite successful in producing pictures from Gaussian noise. It should be noted that this model's application is based on strict physical principles (Sohl-Dickstein et al. 2015; Song and Ermon 2019), which include both a diffusion process and a reversal process. Many contemporary diffusion techniques concentrated on the visual style transfer challenge. For instance, (Kwon and Ye 2023) used contrastive loss, the target image's circumstances, and fine-tuning the classic Diffusion model to accomplish style transfer. (Zhang et al. 2022) extracted feature conditions from the pictures using CLIP encoding, and then added an extra Attention layer for optimization to further optimize the image transfer outcomes. (Zhang et al. 2023a) then suggested breaking down the condition into a number of different control vectors, which allowed for more precise style transmission and

guidance. The concepts of the three approaches mentioned above are similar in that style transfer is achieved by adjusting the diffusion model parameters based on the target domain image's state. In addition, (Mou et al. 2023) accomplishes style migration by adding an Adapter that provides more access to condition without requiring the model's parameters to be adjusted.Like the Adapter in StyleGAN, the role of Adapter is to learn how to construct a control vector condition that satisfies the requirements of the diffusion model. (Parmar et al. 2023) adds contrast loss to strengthen the produced model's resilience and uses ChatGPT to improve the textual cues.

Method

Overview

In this work, we propose a training-free framework that utilizes off-the-shelf stable diffusion models(Rombach et al. 2022) for style transfer guided by a single style image. As shown in Figure 1, In the inversion and embedding stage, we obtain the noise representation x_T^S and x_T^C of input images through DDIM inversion(Song, Meng, and Ermon 2020). The style-content image pair is encoded using the clip image encoder. Then noise injection(Nukrai, Mokady, and Globerson 2022) is applied to the obtained embeddings to narrow the domain gap between image and text features, merging them with the embeddings derived from the clip text encoder to incorporate the style information. In the sampling stage, we first perform image reconstruction tasks by unconditionally sampling the style-content noise pairs to obtain reconstructed images while preserving features at different sampling time steps. Then we sample the noise from the content image noise x_T^C to maintain its original layout and entities. During the sampling process, we leverage the prompt embedding and attention fusion from the unconditional samples to guide the image's style.

DDIM Inversion

In the context of Denoising Diffusion Implicit Models Inversion (DDIM Inversion), the process involves transforming a clear image progressively into noise. This is represented by a series of conditional probabilities $p(x_{t-1}|x_t)$, where x_t is the noisy image at time step t. Mathematically, this process is described as $x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon$, where ϵ is noise sampled from a standard normal distribution, and α_t represents the noise level at step t.

As the time steps increase, the image gradually loses its original structure and eventually becomes entirely noise, denoted as x_T, with T being the total number of steps in the forward process. This transformation forms the foundation for the inverse process, which aims to reconstruct the original image from its noisy state. In the DDIM model, this reverse transition from noise to a clear image is achieved through carefully designed denoising steps, progressively reducing noise and restoring the image's original features.

For accurate reconstruction of a given real image, deterministic DDIM sampling is employed:

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \varepsilon_\theta \left(z_t, t, \mathcal{C} \right)$$

For definitions of α_t and additional details, refer to Appendix E. Diffusion models typically operate in the image pixel space, where z_0 is a sample of a real image.

A simple inversion technique for DDIM sampling, based on the assumption that the ODE process can be reversed in the limit of small steps, is:

$$z_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \varepsilon_\theta \left(z_t, t, \mathcal{C} \right)$$

In other words, the diffusion process is performed in reverse, i.e., $z_0 \rightarrow z_T$ instead of $z_T \rightarrow z_0$, with z_0 set as the encoding of the given real image.

Clip-Based Embedding

The goal of clip-based embedding is to generate captions for a given image I . Unlike supervised methods, we only have access to a set τ of texts during training. These can be obtained by harvesting text corpora. We first introduce the notation of Clip-Based models.Given an image I let $\phi(I) \in R^d$ be its embedding, and given a text T let $\psi(T) \in R^d$ be its embedding.For converting a vector $v \in R^d$ into a caption, we use a textual decoder $C(\upsilon)$ consisting of a lightweight mapping network and a pretrained auto-regressive language model, as suggested in (Mokady, Hertz, and Bermano 2021). We let each text $T \in \mathcal{T}$ is first mapped to CLIP space via $\psi(T)$ and then decoded back into a text via $C(\psi(T))$. We would like this decoding to be similar to the original text T. Namely, our training objective is a reconstruction of the input text from Clip-Based's textual embedding. At inference, given an image I we simply apply the decoder to $\phi(I)$, returning the caption $C(\phi(I))$.

We assume that the visual embedding corresponding to a text embedding lies somewhere within a ball of small radius ϵ around the text embedding. We would like all text embeddings in this ball to decode to the same caption, which should also correspond to the visual content mapped to this ball. We implement this intuition by adding zero-mean Gaussian noise of STD ϵ to the text embedding before decoding it. The value of ϵ is calculated by estimating the spread of captions corresponding to the same image.

Our overall training objective is thus to minimize:

$$\sum_{T \in \mathcal{T}} \ell(C(\psi(T) + n), T)$$

where $n \in R^d$ is a random standard Gaussian noise with standard-deviation ϵ and ℓ is an autoregressive cross-entropy loss for all tokens in T. We train just the parameters of the textual decoder C, while the encoder $\psi()$ is kept frozen. The noise is sampled independently at each application of the encoder.

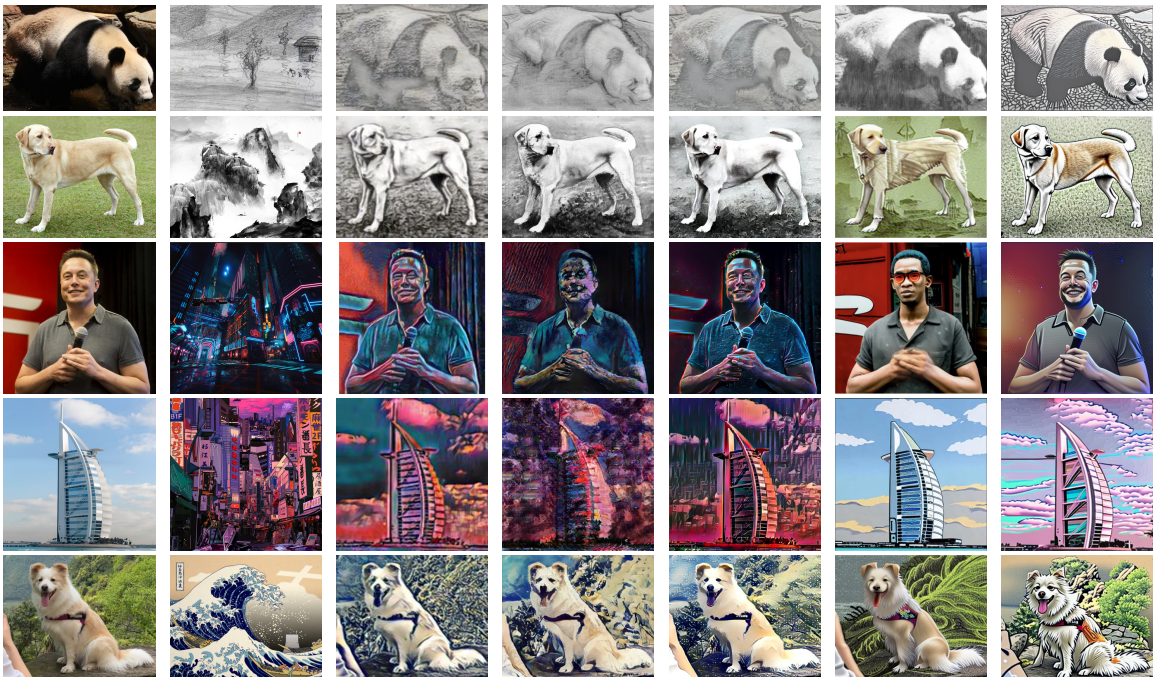

Figure 2: Qualitative comparison with several state-of-the-arts image style transfer methods.

Attention Fuse

The content noise x_T^C is employed as the starting point for DDIM sampling from T to 0 with the prompt embedding P to ultimately transfer the style. P aims to guide the generation style of images through semantic information. However, relying solely on prompt embedding, the pretrained text-to-image model cannot preserve the appearance of the content images effectively, Inspired by feature injection method in prompt-to-prompt(Hertz et al. 2022) and TF-ICON(Lu, Liu, and Kong 2023), we propose fusing self-attention maps in a specially designed manner.

The self-attention maps of style image and content image in the unconditional sampling process is respectively $A_{l,t}^s$ and $A_{l,t}^c$, which are calculated using self-attention modules of the pretrained Stable Diffusion model. Typically, a self-attention module at layer l contains three projection matrices w_l^q, w_l^k, w_l^v in the same dimension $R^{d\times d}$. Denote the feature embeddings of the style and content image at timestep t and layer l as $\mathbf{f}_{l,t}^{\mathrm{s}} \in R^{(h\times w)\times d}$ and $\mathbf{f}_{l,t}^{\mathrm{c}} \in R^{(h\times w)\times d}$ .The queries, keys, and values for each self-attention module are obtained as:

$$\mathbf{q}_{l,t}^{\mathrm{s}} = \mathbf{f}_{l,t}^{\mathrm{s}}\mathbf{W}_l^q, \quad \mathbf{k}_{l,t}^{\mathrm{s}} = \mathbf{f}_{l,t}^{\mathrm{s}}\mathbf{W}_l^k, \quad \mathbf{v}_{l,t}^{\mathrm{s}} = \mathbf{f}_{l,t}^{\mathrm{s}}\mathbf{W}_l^v,$$

$$\mathbf{q}_{l,t}^{\mathrm{c}} = \mathbf{f}_{l,t}^{\mathrm{c}}\mathbf{W}_l^q, \quad \mathbf{k}_{l,t}^{\mathrm{c}} = \mathbf{f}_{l,t}^{\mathrm{c}}\mathbf{W}_l^k, \quad \mathbf{v}_{l,t}^{\mathrm{c}} = \mathbf{f}_{l,t}^{\mathrm{c}}\mathbf{W}_l^v$$

where $\mathbf{q}_{l,t}^{\mathrm{s}}, \mathbf{k}_{l,t}^{\mathrm{s}}, \mathbf{v}_{l,t}^{\mathrm{s}} \in R^{(h\times w)\times d}$ and $\mathbf{q}_{l,t}^{\mathrm{c}}, \mathbf{k}_{l,t}^{\mathrm{c}}, \mathbf{v}_{l,t}^{\mathrm{c}} \in R^{(h\times w)\times d}$. Then $\mathbf{A}_{l,t}^{\mathrm{s}}$ and $\mathbf{A}_{l,t}^{\mathrm{c}}$ are calculated.

$$\mathbf{A}_{l,t}^{\mathrm{s}} = Softmax\left(\mathbf{q}_{l,t}^{\mathrm{s}} \cdot \left(\mathbf{k}_{l,t}^{\mathrm{s}}\right)^{\top} / \sqrt{d}\right),$$

$$\mathbf{A}_{l,t}^{\mathrm{c}} = Softmax\left(\mathbf{q}_{l,t}^{\mathrm{c}} \cdot \left(\mathbf{k}_{l,t}^{\mathrm{c}}\right)^{\top} / \sqrt{d}\right),$$

where $\mathbf{A}_{l,t}^{\mathrm{m}}$, $\mathbf{A}_{l,t}^{\mathrm{r}} \in R^{(h\times w)\times(h\times w)}$, analogously we calculate the self-attention map $\mathbf{A}_{l,t}$ in prompt embedding sampling process and fused $\mathbf{A}_{l,t}^{\mathrm{s}}$, $\mathbf{A}_{l,t}^{\mathrm{c}}$ and $\mathbf{A}_{l,t}$ as $\mathbf{A}_{l,t}^{*}$.

$$\mathbf{A}_{l,t} = \begin{cases} \mathbf{A}_{l,t}^{\mathrm{c}}, & if \quad t < \tau_a \cdot T \\ \mathbf{A}_{l,t}, & otherwise. \end{cases}$$

$$\mathbf{A}_{l,t}^{*} = inject \cdot \mathbf{A}_{l,t}^{\mathrm{s}} + (1 - inject) \cdot \mathbf{A}_{l,t}$$

where $inject$ is the feature injection coefficient of style image, and τ_a is the feature injection step of content image.

Experiments

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

Comparison with Style Transfer Methods

We compare our method with the state-of-the-arts image style transfer methods, including AdaAttN(Liu et al. 2021), StyTr2(Deng et al. 2022), UCAST(Zhang et al. 2023c) and InST(Zhang et al. 2023b) to show the effectiveness of our

method. As shown in Figure 2, we can see apparent advantages of our method on transferring the semantics and artistic techniques of the reference images to the content images over traditional style transfer methods. For example, such as the facial forms and eyes (the 4th row), the animal(the 1st, 3rd, 6th rows), the building and the cloud (the 5th row). Our method can capture some special semantics of the reference images and reproduce the visual effects in the results, such as the ground on the background (the 1st row) and the microphone(the 4th row). Those effects are very difficult for traditional style transfer methods to achieve. As shown in Table 1, our method outperforms (Liu et al. 2021), (Zhang et al. 2023c), (Deng et al. 2022) in $accuracy$.

Table 1: CLIP-based evaluations.

	AdaAttn	**UCAST**	**StyTR2**	**InST**	**Ours**
Acc (%)	65.5627	66.5467	66.4216	68.0280	67.8684

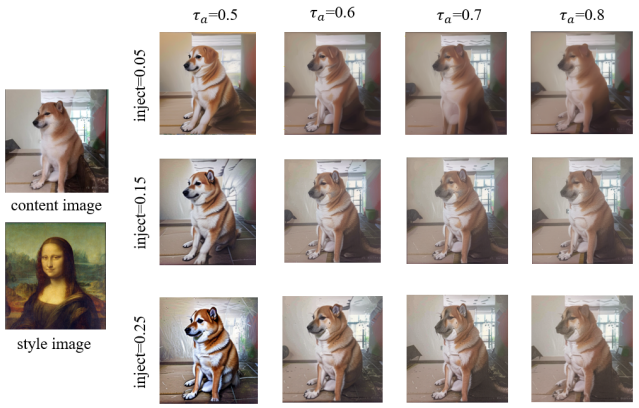

Figure 3: The results of the ablation study on hyperparameters $inject$ and τ_a

Ablation Study

For image synthesis, the most relevant hyperparameters are $inject$ and τ_a. Their effects are illustrated in figure 3. $Inject$ involves incorporating coefficients into the characteristics of the style image. The greater the $inject$ value, the more closely the tone and style of the image align with those of the style image. τ_a denotes the number of steps for injecting features from the content image, and a larger τ_a value results in images closer to the content. A smaller τ_a fosters greater creativity in the content of the generated images.

User Study

We compare our method with several SOTA image style transfer methods(AdaAttn, UCAST, StyTR2, InST).All the baselines are trained using publicly available implementations with default configurations. For each participant, 26 content-reference pairs are randomly selected and the generated results of ours and one of the other methods are displayed randomly. Participants were suggested that the artistic consistency between the generated image and the reference image was the main metric. Then, they were invited to select the better result of each content-reference pair. Finally, we collect 962 votes from 37 participants. The percentage of votes for each approach is shown in Table 2, demonstrating that our method achieves the best visual characteristics transfer results.

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**.

	Preference↑	**Ours**
AdaAttn	0.269	**0.731**
UCAST	0.161	**0.839**
StyTR2	0.298	**0.702**
InST	0.442	**0.558**

Furthermore, we conducted a survey of 30 participants on the preferences of the content image guidance strength and artistic visual effects. In the case of a content image existing, users tend to consider that "To depict the artistic style, the details of the content should be embellished appropriately". We then invite the participants to rank the factors of their expected visual effect. The average comprehensive score of the options in the sorting question is automatically calculated based on the ranking of the options by all the participants. The higher the score, the higher the comprehensive ranking. The scoring rule is:

$$score = \frac{\sum frequency \times weight}{participantes} \tag{1}$$

where $score$ denotes the average comprehensive score of the options, $participantes$ denotes the number of people who complete this question, $frequency$ denotes the frequency that the option is selected by users, $weight$ denotes the weight which is determined by the option's ranking. The ranking results (rank by score from highest to lowest): (1) Similar artistic effect on semantic corresponding subjects ($score$=5.7); (2) With the same paint material ($score$=3.79); (3) Having similar brushstrokes ($score$=3.36); (4) Having typical shapes ($score$=2.85); (5) With the same decorative elements ($score$=2.37); (6) Sharing the same color ($score$=1.7).

Conclusion

In this work, we propose a training-free framework that utilizes off-the-shelf stable diffusion models for style transfer guided by a single style image, which allows the transfer not just limited to a few styles. We fuse the attention acquired in the unconditional sampling and prompt embedding to guide the style of the generated image. The experimental results demonstrate that our method achieves excellent image-to-image and generation results compared with state-of-the-art approaches. I believe our work will better help users generate images in specific styles as they want.

References

An, J.; Li, T.; Huang, H.; Shen, L.; and Luo, J. 2020. Real-time Universal Style Transfer on High-resolution Images via Zero-channel Pruning.

Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; and Jiao, J. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 994–1003.

Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; and Xu, C. 2022. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11326–11336.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. *Neural Information Processing Systems,Neural Information Processing Systems*.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

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*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. volume 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*.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2022. Imagic: Text-Based Real Image Editing with Diffusion Models.

Kwon, G.; and Ye, J. C. 2023. Diffusion-based Image Translation using Disentangled Style and Content Representation.

Li, T.; Qian, R.; Dong, C.; Liu, S.; Yan, Q.; Zhu, W.; and Lin, L. 2018. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, 645–653.

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.

Lu, S.; Liu, Y.; and Kong, A. W.-K. 2023. TF-ICON: Diffusion-Based Training-Free Cross-Domain Image Composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2294–2305.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Mou, C.; Wang, X.; Xie, L.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models.

Nukrai, D.; Mokady, R.; and Globerson, A. 2022. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*.

Park, D. Y.; and Lee, K. H. 2019. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5880–5888.

Parmar, G.; Kumar Singh, K.; Zhang, R.; Li, Y.; Lu, J.; and Zhu, J.-Y. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

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.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, G.; Luo, L.; Liu, J.; Ma, W.-C.; Lai, C.; Zheng, C.; and Cham, T.-J. 2021. Agilegan: stylizing portraits by inversion-consistent transfer learning. *ACM Transactions on Graphics (TOG)*, 40(4): 1–13.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. volume 32.

Wang, H.; Li, Y.; Wang, Y.; Hu, H.; and Yang, M.-H. 2020. Collaborative distillation for ultra-resolution universal style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1860–1869.

Zhang, Y.; Dong, W.; Tang, F.; Huang, N.; Huang, H.; Ma, C.; Lee, T.-Y.; Deussen, O.; and Xu, C. 2023a. ProSpect: Expanded Conditioning for the Personalization of Attribute-aware Image Generation.

Zhang, Y.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; and Xu, C. 2022. Inversion-based creativity transfer with diffusion models.

Zhang, Y.; Huang, N.; Tang, F.; Huang, H.; Ma, C.; Dong, W.; and Xu, C. 2023b. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10146–10156.

Zhang, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Lee, T.-Y.; and Xu, C. 2023c. A Unified Arbitrary Style Transfer Framework via Adaptive Contrastive Learning. *ACM Transactions on Graphics*.