DiffStyler: Controllable Dual Diffusion for Text-Driven Image Stylization

Nisha Huang, Yuxin Zhang, *Graduate Student Member, IEEE*, Fan Tang, Chongyang Ma, Haibin Huang, Weiming Dong, *Member, IEEE*, and Changsheng Xu, *Fellow, IEEE*

***Abstract*— Despite the impressive results of arbitrary image-guided style transfer methods, text-driven image stylization has recently been proposed for transferring a natural image into a stylized one according to textual descriptions of the target style provided by the user. Unlike the previous image-to-image transfer approaches, text-guided stylization progress provides users with a more precise and intuitive way to express the desired style. However, the huge discrepancy between cross-modal inputs/outputs makes it challenging to conduct text-driven image stylization in a typical feed-forward CNN pipeline. In this article, we present DiffStyler, a dual diffusion processing architecture to control the balance between the content and style of the diffused results. The cross-modal style information can be easily integrated as guidance during the diffusion process step-by-step. Furthermore, we propose a content image-based learnable noise on which the reverse denoising process is based, enabling the stylization results to better preserve the structure information of the content image. We validate the proposed DiffStyler beyond the baseline methods through extensive qualitative and quantitative experiments. The code is available at https://github.com/haha-lisa/Diffstyler.**

***Index Terms*— Arbitrary image stylization, diffusion, textual guidance, neural network applications.**

I. Introduction

IMAGE stylization is appealing in practice as it allows amateurs to turn real-world photos into renderings that mimic the style of artwork without requiring any professional skills. Based on the extracted style textures, the style transfer methods [1], [2], [3], [4], [5], [6] migrate the semantic textures of the style images to the content images and generate vivid artwork. They are followed and improved by later works [7], [8] that have shown substantial value in creative visual design.

Manuscript received 15 March 2023; revised 24 October 2023; accepted 1 December 2023. Date of publication 10 January 2024; date of current version 6 February 2025. This work was supported in part by the National Science Foundation of China under Grant U20B2070, Grant 61832016, and Grant 62102162; and in part by the Beijing Natural Science Foundation under Grant L221013. *(Corresponding author: Weiming Dong.)*

Nisha Huang, Yuxin Zhang, Weiming Dong, and Changsheng Xu are with the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: huangnisha2021@ia.ac.cn; zhangyuxin2020@ia.ac.cn; weiming.dong@ia.ac.cn; csxu@nlpr.ia.ac.cn).

Fan Tang is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: tangfan@ict.ac.cn).

Chongyang Ma and Haibin Huang are with Kuaishou Technology, Beijing 100085, China (e-mail: chongyangm@gmail.com; jackiehuanghaibin@gmail.com).

Digital Object Identifier 10.1109/TNNLS.2023.3342645

While the above image stylization methods are capable of delivering attractive and stable results for artwork, they all require users to provide a style image as a reference for stylization, which leads to the following problems. First, providing an appropriate style image itself increases the complexity and redundancy of the user operation. In addition, style images themselves have strong limitations for color and texture. As a result, the user's need for stylization and artistic creation cannot be adequately expressed through a single image either. In contrast, expressing the user's artistic needs and esthetic preferences through text is more in line with the original intention of artistic creation. For illustration, the text conveys arbitrary artists, styles, and artistic movements. This more natural and intuitive way of textual guidance can create more imaginative and unrestricted digital artworks.

Current approaches [9], [10], [11] for text-driven stylization tasks are mainly based on GAN [12], [13] models, which results in limited modeling and generation capabilities between text and style. The GAN-based approaches, which rely on the patch style discriminator, are unable to generate satisfactory results in the case of inappropriate patches to sample. In our investigation, we found that textual information can easily guide the diffusion model for stable stepwise diffusion without relying on additional random style patches. Meanwhile, with the development of diffusion-based methods [14], [15], [16], [17], they have demonstrated phenomenal results on visual tasks, especially in generating artworks [18], [19], [20]. Therefore, we introduce diffusion into stylization tasks to guide the image global sampling process. This improves the problem that GAN-based methods [11], [21] repeatedly have the same stylized patches at different locations of the generated results, thus causing artifacts and failures to highlight the main content.

Moreover, the stylization task requires keeping the content image structure [22], [23], [24], but the current work of diffusion does not preserve the global content of the input image. There are several works [15], [17], [18], [25] that use the content image as the initial input to the diffusion model. However, the gradual addition of noise, which can be destructive to the image content, cannot achieve similar stylization applications. Specifically, DreamBooth [26] can maintain the content of a particular subject and embed it into the output domain. However, the method requires fine-tuning for each new set of content images. In addition, the forward diffusion entails introducing Gaussian noise to the image during the

2162-237X © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

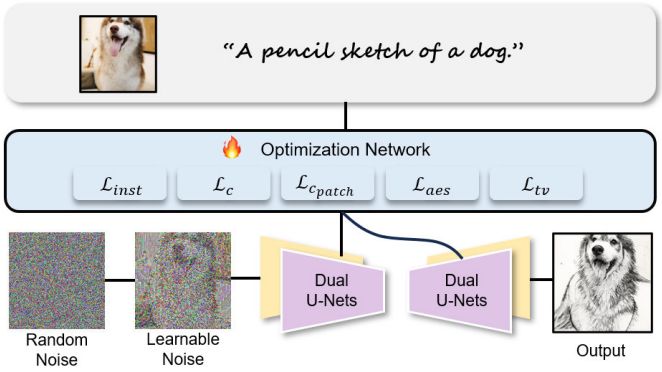


Fig. 1. Illustration of a DiffStyler is showcased, wherein the model generates a stylized image by leveraging a text-based description style and the content image as input.

training phase. This noise gradually transforms the image, making it progressively more featureless and resembling a process of entropy increase. As the noise level increases, the order and regularity within the image weaken, leading to significant changes in the content representation. Therefore, it remains challenging to maintain the global content for arbitrary images based on the diffusion model.

To solve the image content preservation problem of the diffusion model, we propose a novel controllable double-diffusion text-driven image stylization method, the working schematic of which is shown in Fig. 1. Unlike traditional diffusion models for image generation, we make the following three improvements. First, we replace random noise with learnable noise in the free diffusion process of the content image, preserving the main structure of the content image. Second, for each step of the inference process, we adopt dual diffusion architectures so that the stylized results are both explicit in content and abstract in esthetics. Furthermore, we optimize the simulation of the diffusion process in terms of numerical methods, so that the generation process can guarantee quality while speeding up the sampling speed. Experiments show that our diffusion model-based image content retention and style transfer is effective and achieves excellent results on both quantitative metrics and manual evaluation (see Fig. 2). In summary, our main contributions are as follows.

1) We present a new dual diffusion-based text-driven image stylization framework that generates output matching the text prompt of the desired style while preserving the main structure of the input content image.
2) We apply learnable noise based on the content image to overcome the destructive effect of adding random noise to the image content during traditional diffusion.
3) Numerous experimental and qualitative examples show that DiffStyler outperforms baseline methods and achieves outstanding results with desirable content structures and style patterns.

II. Related Work

A. Image Style Transfer

Style transfer aims to migrate the style of a painting to a photograph, maintaining its original content. Initially, Gatys et al. [1] found that hierarchical layers in CNNs can be used to extract image content structures and style texture information and proposed an optimization-based iterative method to generate stylized images. An increasing number of methods have been developed thereafter to advance the quality of stylization. Arbitrary style transfer in real-time [27] is improved by minimizing perceptual loss which is the combination of feature reconstruction loss as well as the style reconstruction loss. More generally, arbitrary style transfer has gained more attention in recent years.

Huang et al. [2] 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 [7], [8], [28] to fuse the content and style features. Deng et al. [23] propose StyTr^2 that contains two different transformer encoders to generate domain-specific sequences for content and style, respectively. Zhang et al. [29] presented contrastive arbitrary style transfer (CAST) to learn style representation directly from image features by analyzing the similarities and differences between multiple styles and taking the style distribution into account. In conclusion, while traditional image style transfer is stable, it requires the provision of additional style images and the results lack artistry and creativity.

B. Text-Driven Image Manipulation

In the existing text-guided image synthesis, the encoders for text embedding work as guide conditions for generative models. OpenAI proposes the high-performance text–image embedding model CLIP [30], based on which several methods manipulate images with textual conditions. StyleCLIP [9] performed attribute manipulation by exploring the learned latent space of StyleGAN [7]. They control the generative process toward a given textual condition by finding the appropriate vector direction. Therefore, StyleGAN-NADA [10] proposed a model modification method using text conditions only and modulated the trained model into a novel domain. Although these methods have been successful in specific domains, they are difficult to apply to arbitrary data.

Based on the above works, Kwon and Ye [11] proposed CLIPstyler that contains a patch-wise text–image matching loss with multiview augmentations for realistic texture transfer. Fu et al. [21] proposed a contrastive language visual artist (CLVA) that learns to extract visual semantics from style instructions and accomplished LDAST by the patch-wise style discriminator. The stylization effect of CLIPstyler [11] and LDAST [21] is dependent on the patch style discriminator, and the quality of the randomly sampled patch will have a critical impact on the results of the transfer.

C. Diffusion Models for Image Synthesis

The diffusion model [31] is a generative model learning to generate images by removing noise from random signals in a stepwise manner, which has received a great deal of attention recently. Sohl-Dickstein et al. [31] were the first to implement image generation using the diffusion model, and with continued research in subsequent approaches [14], [15], [16], [32], [33], [34], [35], the diffusion model is generating

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

Fig. 2. Image stylization results of DiffStyler with diverse text prompts. The input content image for each result is shown as an inset. Using text prompts for stylization, the proposed could generate digital art in various styles.

high-resolution images with unprecedented quality, often surpassing GANs.

Although denoising diffusion probability models (DDPMs) [33] can produce high-quality samples, they require hundreds to thousands of iterations to produce the final samples. Some previous methods have successfully accelerated DDPMs by adjusting variance schedules (e.g., IDDPMs [36]) or denoising equations (e.g., DDIMs [37]). However, these acceleration methods cannot maintain the quality of the samples. Liu et al. [38] proposed the idea that DDPMs should be considered as solving differential equations on manifolds and proposed pseudo-numerical methods for diffusion models (PNDMs) to accelerate the inference process while maintaining the sample quality.

Several large-scale text–image models have recently emerged, such as DALL.E 2 [15], GLIDE [16], and Imagen [35], demonstrating unprecedented image-generation results. Notably, some studies have focused on enhancing the control of the progressive inference process [14], thereby endowing diffusion models with remarkable controllability. Noteworthy advancements in the field of style transfer have been realized through the remarkable text-to-image diffusion models developed by GLIDE [16] and stable diffusion [17], enabling diffusion-based approaches to achieve superior outcomes.

Since the above works are not designed to maintain the structure of the input image, this does not quite satisfy the problem setting of traditional style transfer tasks [1], [5]. If the above works [15], [17] are applied to an image-to-image application with additional steps, the results will differ significantly from the input image. Kim et al. [34] introduced the method known as DiffusionCLIP, which focuses on performing global changes in images. However, it should be noted that DiffusionCLIP is limited in its applicability to specific domains and is not designed to operate on arbitrary images. In contrast to existing approaches, our objective is to develop a novel method that enables the stylization of arbitrary images guided by text prompts through the utilization of the diffusion method. By leveraging the power of diffusion, we aim to provide users with precise and controllable stylization capabilities, expanding beyond the constraints of specific domains.

D. Noise Impact

Given the uniqueness of the effect of noise on the diffusion process, many researchers [39], [40], [41] have conducted in-depth works on input noises. For instance, Video ControlNet [39] proposes that the initial noisy samples used in the denoising process have a significant impact on texture synthesis. Since spatial translation or distortion of the noise can change the output semantics, Video ControlNet introduces a well-designed input noise combination scheme by employing a sliding window approach to maximize the temporal consistency of the generated video. However, discrepancies and inaccuracies in the estimation of the optical flow result in some temporal discontinuities remaining in the generated video.

PFB-Diff [40] utilizes initial random noise and combines progressive feature blending and attention masking mechanisms to generate intermediate noisy images, which are then progressively denoised to obtain the final edited image. One limitation is that the model may struggle to generate desired scenes in background replacement. Another limitation is that the ability to describe desired objects through text remains limited, making personalized editing challenging.

More recently, VideoFusion [41] resolves the per-frame noise into two parts, namely base noise and residual noise, where the base noise is shared by consecutive frames. While VideoFusion sharing the underlying noise between consecutive frames helps to utilize temporal correlation better, it also limits the motion in the generated video and does not apply to video generation with large differences between frames. This allows the image priors of the pretrained model to be efficiently shared by all frames and therefore facilitate the learning of video data.

In contrast to the aforementioned research, our study investigates the influence of noise on content preservation within the domain of style transfer. In addition, an attempt is made to employ learnable noise within a diffusion modeling framework to facilitate a style stylization task.

III. Method

We now formally introduce DiffStyler, a controllable dual diffusion framework for text-driven image stylization. Given a content image $\mathbf{x}_0$ and the target text prompt $\mathcal{T}$, DiffStyler can transfer $\mathbf{x}_0$ into a stylized one with the desired style. As illustrated in Fig. 3, DiffStyler consists of three main stages. First, we input the content image $\mathbf{x}_0$ into the diffusion model and let it perform the T_1 steps of free-guided diffusion to obtain learnable noise. Second, the result $\hat{\mathbf{x}}_T$ at the T steps of free-guided diffusion is used as the input $\mathbf{x}'_T$ to the dual diffusion model for the reverse sampling process. Third, in the reverse sampling process, we perform the relevant optimization. The process is depicted in Algorithms 1 and 2. We investigate the inherent properties of the diffusion model and propose novel content-preserving

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

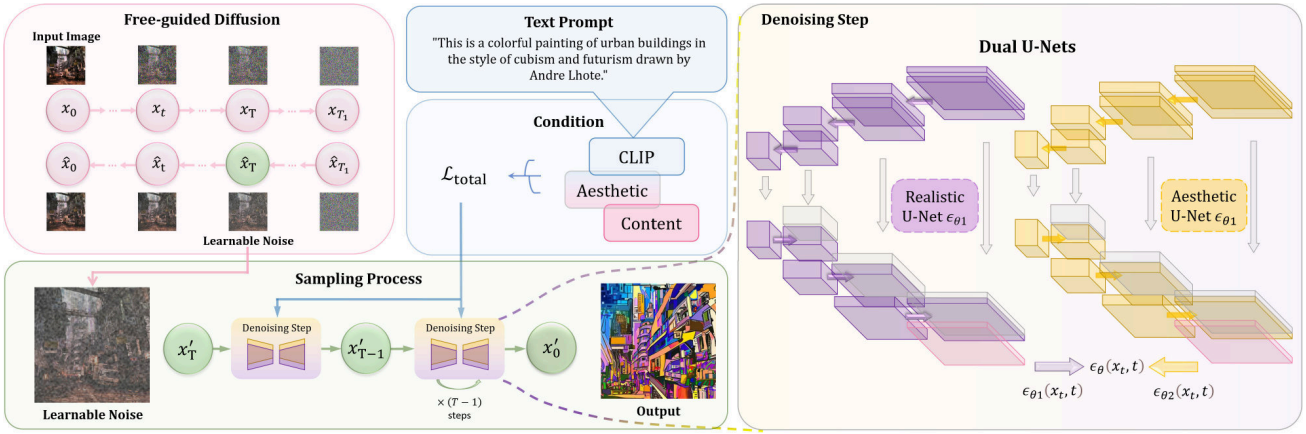


Fig. 3. Pink module represents the free-guided diffusion process, which generates the learnable noise denoted as $\hat{\mathbf{x}}_T$. $\hat{\mathbf{x}}_T$ is set to be equal to $\mathbf{x}'_T$ as the starting map for the sampling process of the green module. In addition, the guided condition of the blue module is incorporated into the denoising step of the sampling process. The dual denoising U-Nets architecture, represented by the yellow and purple mixed-color module, is employed for each step. The purple architecture corresponds to the realistic U-Net, while the yellow architecture represents the esthetic U-Net. After T steps of denoising, the final output $\mathbf{x}'_0$ is obtained.

Algorithm 1 Learnable Noise Generating, Given a Diffusion Model $(\mu_\theta(\mathbf{x}_t), \Sigma_\theta(\mathbf{x}_t))$

Require: The input image $\mathbf{x}_0$, free-guided noising steps T_1, denoising steps T.
Ensure: Learnable noise $\hat{\mathbf{x}}_T$.
  $t = T_1$
  $\mathbf{x}_{T_1} \leftarrow$ sample from $\mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\hat{\mathbf{x}}_{T_1} = \mathbf{x}_{T_1}$
  repeat
    $t - 1 \leftarrow t$
    $\mu_\theta(\hat{\mathbf{x}}_t) \leftarrow \boldsymbol{\epsilon}_\theta(\hat{\mathbf{x}}_t, t \mid \emptyset)$
    $\mu, \Sigma \leftarrow \mu_\theta(\hat{\mathbf{x}}_t), \Sigma_\theta(\hat{\mathbf{x}}_t)$
    $\hat{\mathbf{x}}_{t-1} \sim \mathcal{N}(\mu, \Sigma)$
  until $t = T$

Algorithm 2 Text-Guided Stylized Diffusion Sampling, Given a Diffusion Model $(\mu_\theta(\mathbf{x}_t), \Sigma_\theta(\mathbf{x}_t))$

Require: The input image $\mathbf{x}_0$, text prompt $\mathcal{T}$, gradient scale w, free-guided noising steps T_1, denoising steps T.
Ensure: Stylized result $\mathbf{x}'_0$ guided by $\mathcal{T}$.
  $t = T$
  $\hat{\mathbf{x}}_T =$ learnable noise $(\mathbf{x}'_T)$
  $\mathbf{x}'_T = \hat{\mathbf{x}}_T$
  for t = T, ..., 1 **do**
    $\mu, \Sigma \leftarrow \mu_\theta(\mathbf{x}'_t), \Sigma_\theta(\mathbf{x}'_t)$
    $\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}'_t, t \mid \mathcal{T}) \leftarrow w \cdot \epsilon_{\theta_1}(\mathbf{x}'_{t-1}, t \mid \mathcal{T}) + (1-w) \cdot \epsilon_{\theta_2}(\mathbf{x}'_t, t \mid \mathcal{T})$
    $\mu_\theta(\mathbf{x}'_t) \leftarrow \tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}'_t, t \mid \mathcal{T})$
    $\mathcal{L}_{total} \leftarrow \lambda_d \mathcal{L}_{inst} + \lambda_{c1} \mathcal{L}_c + \lambda_{c2} \mathcal{L}_{c_{patch}} + \lambda_{aes} \mathcal{L}_{aes} + \lambda_{tv} \mathcal{L}_{tv}$
    $\mathbf{x}'_{t-1} \leftarrow$ sample from $\mathcal{N}(\mu_\theta(\mathbf{x}'_t) + \Sigma \mathcal{L}_{total}, \Sigma)$
  end for
  return $\mathbf{x}'_0$

learnable noise and a dual diffusion pathway for style/content decoupling and successfully enable the diffusion model to break through the limitations of the diffusion model in style transfer tasks.

A. Dual Diffusion Models

1) Denoising Diffusion Probabilistic Models: Recently, DDPMs have been shown to generate high-quality images [14], [15], [33]. DDPMs learn the denoising process for parameterized Markov noise images. The isotropic Gaussian noise samples are converted to samples from the training distribution. In the following, we provide a brief overview of DDPMs [14], [33], [36], [37].

At the core of DDPMs lies the forward noising process, which involves adding Gaussian noise with variance $\beta_t \in (0, 1)$ at each time step t to an initial data distribution $\mathbf{x}_0 \sim q(\mathbf{x}_0)$. This creates a sequence of images $\mathbf{x}_1, \ldots, \mathbf{x}_T$ governed by the following equations:

$$q(\mathbf{x}_1, \ldots, \mathbf{x}_T \mid \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$$
$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\left(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right) \quad (1)$$

where $q(\mathbf{x}_1, \ldots, \mathbf{x}_T \mid \mathbf{x}_0)$ denotes the joint distribution of the image sequence given the initial image $\mathbf{x}_0$, and $q(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ represents the conditional distribution of image $\mathbf{x}_t$ given the previous image $\mathbf{x}_{t-1}$. It is important to note that as the number of steps T increases, the final output $\mathbf{x}_T$ tends to approximate an isotropic Gaussian distribution.

A key feature of the forward noising is that each step $\mathbf{x}_t$ may be sampled straight from $\mathbf{x}_0$, eliminating the need to construct intermediary stages

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right)$$
$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \quad (2)$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denotes a Gaussian noise sample, $\alpha_t = 1 - \beta_t$ represents the complementary noise level at time t, and $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$ captures the accumulated noise level up to time t. Importantly, this process can be inverted to generate a fresh sample from the distribution $q(\mathbf{x}_0)$.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

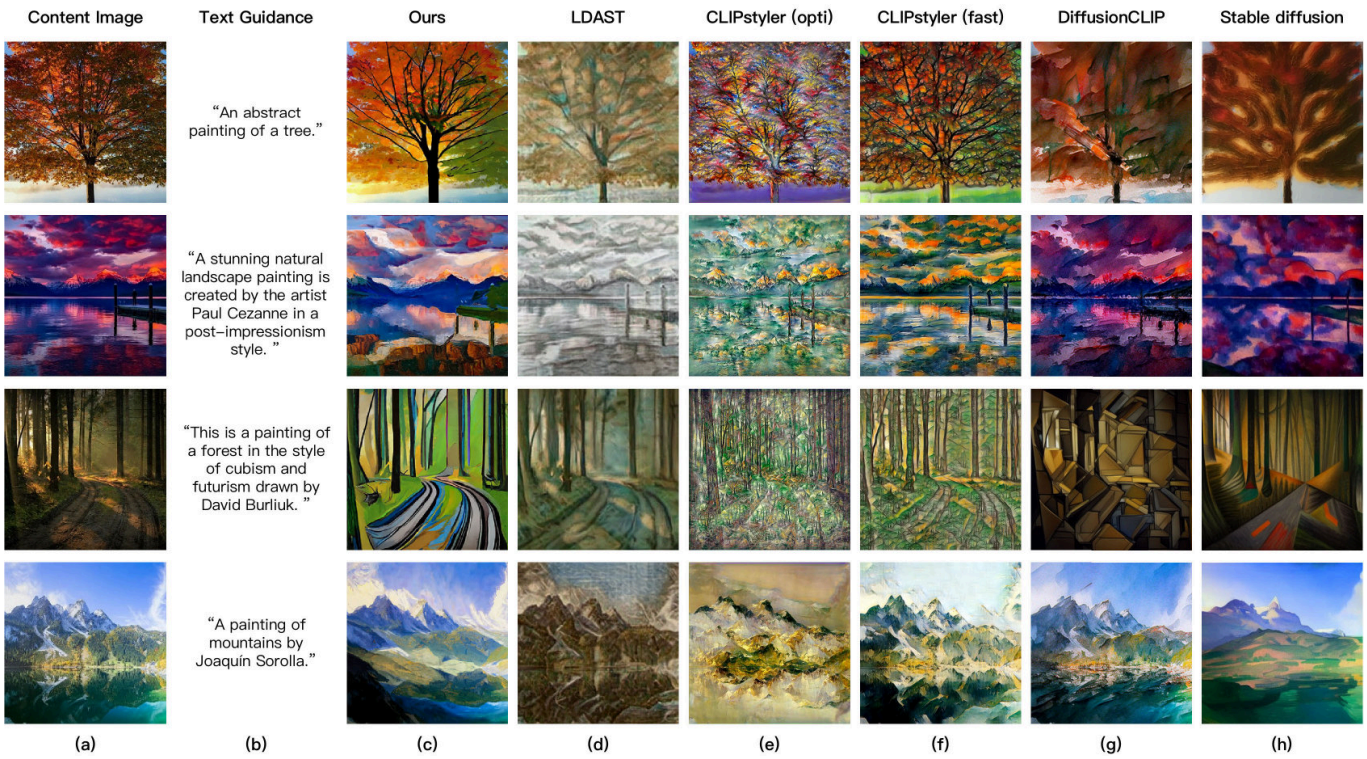


Fig. 4. Qualitative comparison with state-of-the-art image style transfer methods. (a) Content image. (b) Text guidance. (c) Our results. (d) LDAST. (e) CLIPstyler (opti). (f) CLIPstyler (fast). (g) DiffusionCLIP. (h) Stable Diffusion.

The Markovian process is reversed to generate a fresh sample from the distribution $q(\mathbf{x}_0)$. The posteriors, $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$, which were demonstrated to also be Gaussian distributions [31]. They are sampled beginning from a Gaussian noise sample, $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, to produce a reverse sequence. $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ is determined by the unknown data distribution $q(\mathbf{x}_0)$.

To predict the mean and covariance of $\mathbf{x}_{t-1}$ using the input $\mathbf{x}_t$, a deep neural network p_θ is employed. This network enables the sampling of $\mathbf{x}_{t-1}$ from a normal distribution parameterized by these statistics

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)). \tag{3}$$

Since it is very difficult to infer $\mu_\theta(\mathbf{x}_t, t)$ directly, we refer to [33] and first compute the noise prediction $\epsilon_\theta(\mathbf{x}_t, t)$ and then substitute it into Bayes' theorem (refer to Equation 2) to derive

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right). \tag{4}$$

2) Pseudo-Numerical Methods for DDPM: Diffusion models convert Gaussian data to images iteratively. They require many iterations for high-quality samples, slowing large sample generation. Numerical methods have limitations within a limited range. Another problem arises when using numerical methods with the diffusion model equation. The neural network and the equation are well-defined only within a limited range. To address this, a reverse process is used to calculate the derivative of the generated data.

The diffusion model equation is unbounded in most cases, which means that the generation process can produce samples far away from the well-defined area, introducing new errors. The related differential equations of the diffusion model may be derived directly and self-consistently to provide a theoretical relationship between diffusion processes and numerical methods

$$\frac{d\mathbf{x}}{dt} = -\bar{\alpha}'(t)\left(\frac{\mathbf{x}(t)}{2\bar{\alpha}(t)} - \frac{\epsilon_\theta(\mathbf{x}(t), t)}{2\bar{\alpha}(t)\sqrt{1-\bar{\alpha}(t)}}\right). \tag{5}$$

The equation of the numerical method is redefined as

$$\mathbf{x}_{t+\delta} = \mathbf{x}_t + \delta f \Rightarrow \mathbf{x}_{t+\delta} = \phi(\mathbf{x}_t, \epsilon_t, t, t+\delta). \tag{6}$$

Here, f is (5). It has been experimentally demonstrated [38] that the pseudo-linear multistep method is the most efficient method for the diffusion model with similar generation quality.

3) Learnable Noise: The diffusion model [31] is inspired by nonequilibrium thermodynamics. They define a Markov chain of diffusion steps that adds random noise slowly to the data and then learn to reverse the diffusion process by reconstructing the desired data samples from the noise. In previous diffusion work, either a random Gaussian noise image or an input image superimposed with a corresponding amount of random Gaussian noise is usually used as the starting image $\mathbf{x}'_T$ for the sampling process. This operation does not protect the content structure of the input image, making the generated image differ significantly from the input image in terms of content, as shown in the results of stable diffusion in Fig. 4. Despite the remarkable progress of diffusion models for image generation, the destructive nature of the stochastic noising process and the random nature of the reverse denoising process make it difficult to preserve the content of the original

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

image, leading that style transfer by diffusion methods still not well explored. In this context, we investigate and exploit the inherent properties of the diffusion model. Our findings indicate that the final denoising outcome of $\mathbf{x}_T$ is achieved by performing T steps of free diffusion on the input image without any textual guidance. Specifically, we introduce a zero embedding as a guiding condition for the denoising network, represented by

$$\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) = \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t \mid \emptyset). \tag{7}$$

The denoising network $\epsilon_\theta(\mathbf{x}_0)$ represents the parameterized conditional model. Typically, the number of forward and sampling process steps in diffusion methods are equal. However, we discover that increasing the number of free diffusion steps T_1 beyond the number of reverse diffusion steps T better preserves the content structure of the input image. Therefore, we adopt $\hat{\mathbf{x}}_T$ as the initial image $\mathbf{x}'_T$ for the sampling process in our experiments, as shown by the pink arrow in the pipeline in Fig. 3. T_1 and T are set to 150 and 50, respectively. This approach ensures enhanced preservation of the content structure of the input image, rendering it suitable for image stylization tasks. This process is summarized in Algorithms 1 and 2.

4) Basic Architecture of DiffStyler: We found that if the diffusion model is trained only on the natural image dataset, the results will be close to real images and lack artistic appearance. If the diffusion model is trained only on the artistic image training set, the results will be too abstract to preserve the input content. Therefore, we used light denoising networks $\boldsymbol{\epsilon}_{\theta 1}$ and $\boldsymbol{\epsilon}_{\theta 2}$ trained on the Conceptual 12M [42], [43] and WikiArt [44] datasets, respectively. Then, we perform sampling using the following linear combination of the dual channel score estimates:

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t \mid \mathcal{T}) = w\boldsymbol{\epsilon}_{\theta 1}(\mathbf{x}_t, t \mid \mathcal{T}) + (1-w)\boldsymbol{\epsilon}_{\theta 2}(\mathbf{x}_t, t \mid \mathcal{T}) \tag{8}$$

where $0 < w < 1$. We utilize it for each step of the inference process, resulting in stylized outputs that are both explicit in content and abstract in esthetics.

B. Network Optimization

1) Instruction Loss: We leverage a pretrained ViT-B/16 CLIP [30] model to stylize the content image according to the text prompt. The cosine distance between the CLIP embedding of the transferred image $\mathbf{x}_t$ during the diffusion process and the CLIP embedding of the text prompt $\mathcal{T}$ may be used to specify the CLIP-based loss, or $\mathcal{L}_{\text{inst}}$. We define the language guidance function using the cosine distance, which measures the similarity between the embeddings $E_{\mathbf{x}_t}$ and $E_{\mathcal{T}}$. The text guidance function can be defined as

$$\mathcal{L}_{\text{inst}}(\mathbf{x}_t, t, \mathcal{T}) = \mathcal{D}_{\text{CLIP}}\left(E_{\mathbf{x}_t}, E_{\mathcal{T}}\right) \tag{9}$$

where $\mathcal{D}_{\text{CLIP}}$ is the cosine distance of the CLIP embeddings.

2) Content Perceptual Loss: To further enhance the alignment with the content of input images, we use feature maps extracted to compute the content loss and patchwise contrastive content loss [1], [45]

$$\mathcal{L}_c = \frac{1}{N_l}\sum_{i=0}^{N_l}\left\|\phi_i\left(\mathbf{x}'_0\right) - \phi_i(\mathbf{x}_0)\right\|_2 \tag{10}$$

where $\phi_i(\cdot)$ denotes features extracted from the ith layer in a pretrained VGG19 and N_l is the number of layers.

Besides, we utilize contrast learning [45] to match corresponding input–output patches at a specific location. We can leverage the other patches within the input as negatives. Let $\mathbf{x}'_0, \mathbf{x}_0 \in \mathbb{R}^{H_{\text{patch}} \times W_{\text{patch}} \times C}$ be patched from the two signals, in the same spatial location. The patch with resolution $H \times W$ is an RGB image patch ($C = 3$). The encoder F is a simple linear projection of a patch of pixels. The embedding vectors of the network generated and input image patches are defined as $v = F(\mathbf{x}'_0)$ and $v^+ = F(\mathbf{x}_0)$, respectively. The N negatives are mapped to K-dimensional vectors and $\boldsymbol{v}_n^-$ denotes the nth negative. The loss is specified as an $(N+1)$-way classification problem with logits proportional to the similarity of embedded patches. To represent the probability of a positive example being selected over a negative example, the cross-entropy loss is calculated as follows:

$$\ell_{c_{\text{patch}}}\left(v, v^+, v^-\right) = -\log\left[\frac{\exp\left(s\left(v \cdot v^+\right)/\tau\right)}{\exp(s(v \cdot v^+)/\tau) + \sum_{n=1}^{N}\exp\left(s\left(v \cdot v_n^-\right)/\tau\right)}\right] \tag{11}$$

where $\tau = 0.07$ is a parameter to control temperature. $s(v_1, v_2) = v_1^{\mathrm{T}} v_2$ delivers the dot product of the similarity of two encoded patch signals. We leverage the notation $v_l^s \in \mathbb{R}^{D_l}$ to index into the tensor, which is the D_l-dimensional feature vector at the sth spatial location. The collection of feature vectors at all other spatial locations is represented as $\bar{v}_l^s \in \mathbb{R}^{(S_l-1)\times D_l}$, where S_l is the number of spatial locations of the tensor. And the patchwise contrastive content loss is defined as

$$\mathcal{L}_{c_{\text{patch}}}\left(\mathbf{x}'_0, \mathbf{x}_0\right) = \sum_{l=1}^{L}\sum_{s=1}^{S_l}\ell_{c_{\text{patch}}}\left(\hat{v}_l^s, v_l^s, \bar{v}_l^s\right). \tag{12}$$

3) Esthetic Loss: We also employ an esthetic loss to make the model's representation of style more consistent with human preferences. We leverage a model to fit and inference code for CLIP esthetic regressions R trained on Simulacra Esthetic Captions [46], [47]. It is a dataset of over 238 000 synthetic images generated by diffusion models as well as human ratings. We use it as a scorer to evaluate the results generated by DiffStyler and return the weighted esthetic exploring loss for optimization

$$\mathcal{L}_{\text{aes}} = -R(\mathbf{x}_t, \mathcal{T}). \tag{13}$$

$\mathcal{L}_{\text{aes}}$ improves the overall visual quality.

4) Total Variation Loss: Total variation loss ($\mathcal{L}_{tv}$) serves as a regularization technique extensively utilized in image processing and computer vision applications. Its primary objective is to promote the presence of smooth transitions and minimize

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

noise by penalizing sudden alterations in intensity or gradients. The mathematical representation for $\mathcal{L}_{tv}$ is defined as

$$\mathcal{L}_{tv} = \sum_{i,j} \left(|I_{i+1,j} - I_{i,j}| + |I_{i,j+1} - I_{i,j}| \right). \tag{14}$$

In the above equation, the symbol I corresponds to the image under consideration, while $I_{i,j}$ denotes the pixel intensity located at the spatial coordinates (i, j). This loss computation involves the absolute differences between neighboring pixel intensities in both the horizontal and vertical directions. By summing up these differences across the entire image, $\mathcal{L}_{tv}$ encourages smoothness by discouraging abrupt variations in pixel values.

As a consequence, the total loss is formulated as follows:

$$\mathcal{L}_{\text{total}} = \lambda_d \mathcal{L}_{\text{inst}} + \lambda_{c1} \mathcal{L}_c + \lambda_{c2} \mathcal{L}_{c_{\text{patch}}} + \lambda_{\text{aes}} \mathcal{L}_{\text{aes}} + \lambda_{tv} \mathcal{L}_{tv}. \tag{15}$$

We set λ_d, λ_{c1}, λ_{c2}, λ_{aes}, and λ_{tv} to 50, 3.0, 1.0, 10, and 80, respectively. The impacts of loss weights are discussed in Section IV-E.

IV. Experiments

A. Implementation Details

For the CLIP [30] model, we use ViT-B/16 as the vision transformer [48]. For the diffusion model, we use the primary model [43] of a resolution 256 × 256 trained on the Conceptual 12M [42] dataset. In addition, we use a pretrained model [43] on the WikiArt [44] dataset as a secondary diffusion model to enable a superior artistic representation of the entire model. Our approach to the combination of the two pretrained models can be extended to the combination of any diffusion model according to the demands of the application scenario. This results in significant savings in training resources. The output resolution of DiffStyler is 256 × 256. The U-Net [49] architecture we utilize is based on Wide-ResNet [50]. To ensure the quality of the results and to maintain consistency of the parameters, the diffusion step and the time step used for the experiments in our work are set to 50.

B. Computing Consumptions

DiffStyler takes 18 s to generate a 256 × 256 image on a single NVIDIA GeForce RTX 3090. It takes 14 s with a single channel. This is a moderate amount of time compared to related work. DiffusionCLIP [34] and Stable diffusion [17] are two other diffusion-based methods that take significantly longer to reason about compared to our model. This difference is because they require longer inference steps to obtain relatively desirable results. Although there is still a gap between our current speed compared to the nondenoising method [21], given the continuous progress in accelerating the denoising inference process, it can be expected that the efficiency of the current inference process will be further improved in the foreseeable future.

The inference process for one image consumes about 4 GB (6 and 16 GB for DiffusionCLIP [34] and Stable Diffusion [17], respectively), which is much smaller than the general diffusion model [26], [35]. The WikiArt model and the CC12M model, trained using 4 NVIDIA GeForce RTX 3090s, required approximately 40 min and 50 h of training time, respectively.

C. Qualitative Evaluation

1) Comparison With Text-Driven Stylization: We compare DiffStyler with SOTA text-driven image stylization methods including LDAST [21], CLIPstyler [11], DiffusionCLIP [34], and Stable Diffusion [17]. There are two versions of CLIPstyler. CLIPstyler (opti) requires real-time optimization on each content and each text. CLIPstyler (fast) requires real-time optimization on each text. We can observe that LDAST has relatively similar style performance results in different text-driven cases and is less artistic. CLIPstyler (opti) texture patches used to convey style are substantially corrupting the content of the input image, for example, (in the third row) the dog's eyes are shifted to other locations in the result. Both CLIPstyler (opti) and CLIPstyler (fast) are more focused on details and lack macrolines, color blocks, and textures. DiffusionCLIP and Stable Diffusion do not express style as prominently (e.g., the first, second, and fourth rows) and result in severe content loss (e.g., the third row) under some textual conditions.

2) Comparison With Style Transfer: We indirectly compare our results with existing style transfer methods, by comparing the proposed DiffStyler with the "retrieval-stylization" baselines (see Fig. 5). The text retrieved art images are first used as style images [56] and then stylized using the state-of-the-art available artistic style transfer methods, including Gatys et al. [1], AdaIN [2], AdaAttN [52], MCC-Net [53], ArtFlow+AdaIN [22], StyTr^2 [23], CAST [29], CAP-VSTNet [54], and UCAST [55].

The results of earlier baseline models [1], [2], [22], [52] demonstrate that their portrayal of the intended texture style was limited, focusing only on color variations. And the content information is influenced by the content of the style image. Recent advances in traditional style transfer methods [23], [29], [53], [54], [55] show promise in capturing richer, more intricate textural nuances while better maintaining the structural integrity of the source content. Nevertheless, these state-of-the-art methodologies appear to struggle when the defining brushstrokes or stylistic elements within the reference artistic image are less prominently discernible, as exemplified by the inferior second-row outputs relative to other approaches.

In contrast, our method yields results with more intricate brushstrokes that closely resemble real art paintings. The major advantage of the diffusion approach lies in its capability to handle complex and nonstationary stylistic patterns. Unlike some general deep-learning stylization methods that struggle with intricate or rapidly changing artistic styles, diffusion-based methods can effectively adapt to such variations. Diffusion methods are effective in preserving fine details and structural information and flexible in achieving desired artistic effects compared to general deep-learning methods. These advantages make diffusion-based methods well-suited for producing high-quality, faithfully stylized images that capture the essence of various artistic styles.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

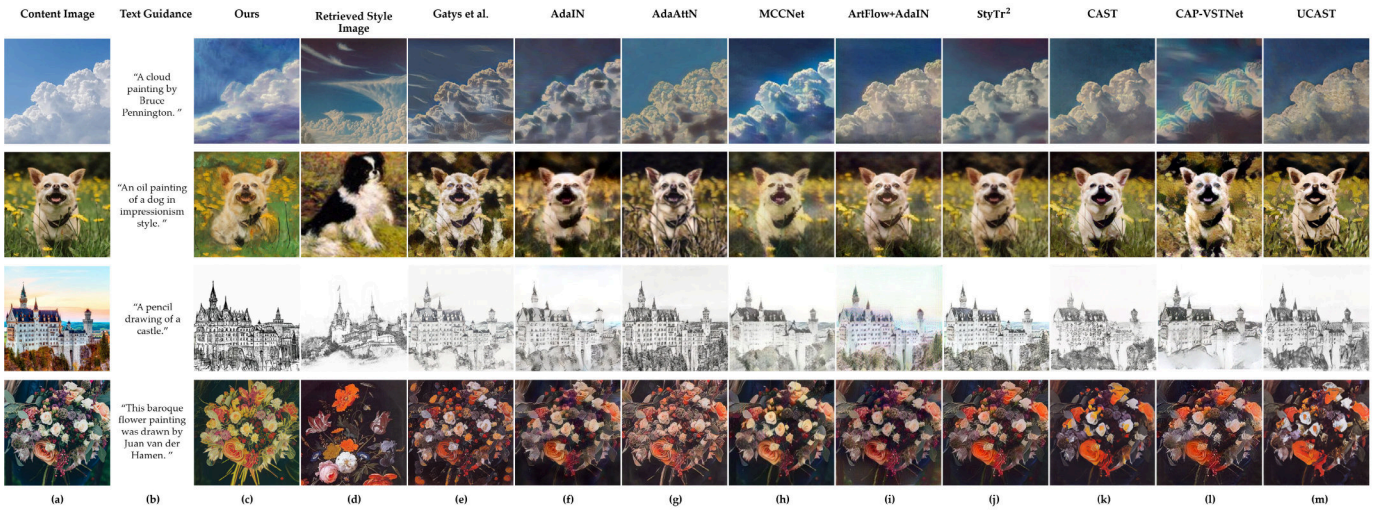


Fig. 5. Qualitative comparison with image style transfer methods. Our approach uses the text guidance directly as the style condition, while others need to first retrieve a style reference from an image database based on the text input before generating the style transfer results. (a) Content image. (b) Text guidance. (c) Our results. (d) Retrieved style image. (e) Gatys et al. (f) AdaIN. (g) AdaAttN. (h) MCCNet. (i) ArtFlow+AdaIN. (j) StyTr^2. (k) CAST. (l) CAP-VSTNet. (m) UCAST.

TABLE I

QUANTITATIVE COMPARISONS FOR TEXT-DRIVEN IMAGE STYLE TRANSFER. WE COMPUTE THE AVERAGE LPIPS AND CLIP SCORE. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** WHILE THE SECOND BEST RESULTS ARE MARKED WITH AN UNDERLINE

	Ours	LDAST [21]	CLIPstyle (opti) [11]	CLIPstyle (fast) [11]	DiffusionCLIP [34]	Stable diffusion [17]
Inference time (s/image) ↓	18	**1**	20	45	40	20
CLIP score [30] ↑	**0.2869**	0.2235	0.2422	0.2515	0.2389	0.2773
LPIPS [51] ↓	0.505	0.744	0.573	**0.472**	0.619	0.637

3) Content Modifications: DiffStyler's dual diffusion is concerned with decoupling the representation of content and style. As a result, we can create more realistic, fine-grained structures in a variety of styles. As shown in Fig. 4, our approach maintains better overall content compared to other approaches (e.g., the first and third rows of Fig. 4). In addition, to further verify the effectiveness of our method in content preservation, we also compare it with traditional style transfer methods. As can be seen in Fig. 5, DIffStyler can preserve both the detailed content and has a very harmonious overall effect. Balancing content and style preservation can be challenging for image style transfer. A significant advantage of DiffStyler is that there are no obvious artificial traces of content preservation of other traditional style transfer methods. Finally, we experimentally found that the content modification is related to the chosen style and is specified in Section IV-F.

D. Quantitative Evaluation

For quantitative evaluation, we randomly selected 20 content images and defined text prompts containing 25 styles and artistic movement descriptions, generating 500 stylized images for each method.

1) CLIP Score: To measure the correspondence between text prompts and stylized textures, we computed the CLIP score, which is the cosine similarity between the target texts derived from the CLIP encoder. Since we employ the ViT-B/16 CLIP model throughout the inference phase, we compute the CLIP score using the ViT-B/32 CLIP [30] model for fairness.

The results (the first row of Table I) show that our method achieves the highest score of 0.2869. Stable diffusion [17] follows closely with a score of 0.2773, marking the second-best result. This indicates that our approach demonstrates superior capability in expressing textual requirements through stylized image outputs.

2) LPIPS Metric: The learned perceptual image patch similarity (LPIPS) [51] metric is widely used to measure the difference between two images based on LPIPS, which aligns more closely with human perception compared to traditional methods. The lower the LPIPS value, the higher the similarity between the images.

Analyzing the second row of Table I, it can be observed that both the CLIPstyler (fast) and the DiffStyler methods outperform the other approaches in terms of content retention, as indicated by their lower LPIPS values. The proposed methods effectively preserve the content of the input images while applying the desired style. This performance is desirable, as it ensures that the stylized images maintain the key features and characteristics of the original content.

3) Content Loss: To perform a quantitative evaluation of our approach, we leverage content loss [57] alongside establishing the state-of-the-art stylization techniques employed for test ratings. Table II presents the reported scores for both our proposed DiffStyler method and the baseline models. Notably, our approach showcases several advantages in terms of content preservation, even when compared to traditional stylization methods. The employed learnable noise methods

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

TABLE II

QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART TRADITIONAL STYLIZATION METHODS. THE BEST RESULTS ARE SHOWN IN **BOLD**, WHILE THE SECOND-BEST RESULTS ARE MARKED WITH AN UNDERLINE

	Ours	Gatys et al. [1]	AdaIN [2]	AdaAttN [52]	MCCNet [53]	ArtFlow+AdaIN [22]	StyTr^2 [23]	CAST [29]	CAP-VSTNet [54]	UCAST [55]
content loss ↓	<u>0.156</u>	0.188	0.176	0.162	0.159	0.172	**0.148**	**0.148**	0.175	0.158

TABLE III

USER STUDY RESULTS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, WHILE THE SECOND BEST RESULTS ARE MARKED WITH AN UNDERLINE

	Ours	LDAST [21]	CLIPstyler (opti) [11]	CLIPstyler (fast) [11]	DiffusionCLIP [34]	Stable Diffusion [17]	MGAD [18]
Quality	**4.41**	3.77	3.75	4.08	3.82	<u>4.13</u>	4.10
Content	<u>3.97</u>	3.94	3.48	**4.06**	3.65	3.83	3.73
Consistency	**4.39**	3.73	3.84	3.95	3.71	<u>4.12</u>	4.08

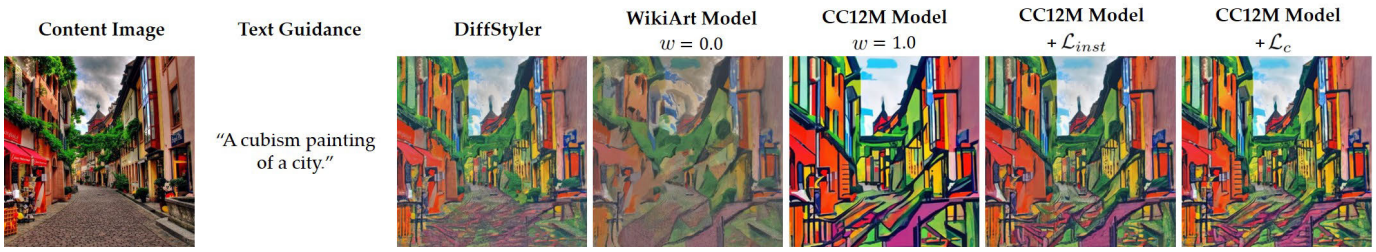


Fig. 6. Ablation study of single and dual diffusion architecture.

successfully retain the essential content of the input images while effectively applying the desired style.

4) User Study: Since the evaluation of artistic aspects is relatively subjective, we conducted a user perception assessment to further compare our methods. We compared DiffStyler with several state-of-the-art text-guided style transfer methods (i.e., including LDAST [21], CLIPstyler [11], DiffusionCLIP [34], and Stable diffusion [17]) and multimodal-driven image generation method MGAD [18]. All baselines we used to perform inference were publicly available implementations with default configurations. For each participant, 28 input–output pairs were randomly selected. Participants were requested to rate the following metrics on a 1–5 scale based on the stylized results: 1) quality of stylized results; 2) content preservation; and 3) consistency of stylized results with text prompts. Finally, we collected 6720 scored results from 80 participants. Table III indicates that DiffStyler scores are high for all three evaluation metrics, which indicates that our findings excel at the level of text-guided stylization.

E. Ablation Study

1) Dual Diffusion Guidance Scales: The guidance scale w of the diffusion model determines how strongly each of the two-path diffusion guides the final stylization outputs. As shown in Fig. 6, when the CC12M model is not utilized and only the WikiArt model is employed (as shown in column 4), the stylization effect closely aligns with the text guidance. However, this approach results in significant content deformation while expressing the desired style. On the other hand, when solely relying on the CC12M model without the WikiArt model (as shown in column 5), the generated image retains clearer content and exhibits a more pronounced style expression. Nevertheless, the generated image, when solely relying on the CC12M model without the WikiArt model, deviates from the style explicitly specified in the textual guidelines. Additionally, it displays issues such as oversaturated colors, unnatural textures, and noticeable artifacts. Furthermore, columns 6 and 7 of Fig. 6 highlight that the impact of the loss on image generation outcomes cannot be equated with the impact of the U-Net model alone. These results demonstrate the importance of employing the dual diffusion model architecture.

2) Loss Terms: We ablate the different loss terms in our objective by qualitatively comparing our results when training with our full objective (15) and with a specific loss removed. The results are shown in Fig. 7. We set up four control groups for ablation experiments to illustrate the effectiveness of loss. λ_d is used as a control variable for $\mathcal{L}_{\text{inst}}$, and the generated results are more consistent with the text as λ_d increases. λ_{c1} is the control variable of $\mathcal{L}_c$, which makes the overall features of the generated image more consistent with the content map. λ_{c2} is the control variable of $\mathcal{L}_{c_{\text{patch}}}$ for better performance of the detailed features. As shown in Table IV, without $\mathcal{L}_c$ and $\mathcal{L}_{c_{\text{patch}}}$, the LPIPS increased by 0.088 and 0.016. λ_{aes} makes the results more consistent with human esthetic preferences. $\mathcal{L}_{tv}$, which is common in image generation work, can remove artifacts and improve smoothing as λ_{tv} increases. Thus, when we use all the proposed loss functions, we can obtain better-stylized results in the perceptual domain.

3) Diffusion Steps: The number of steps determines the quality of the generation and the length of the generation time. We found experimentally (see Fig. 8) that the number of inference steps T is set to 50, and we can obtain good results. When the number of inference steps is larger than 50,

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

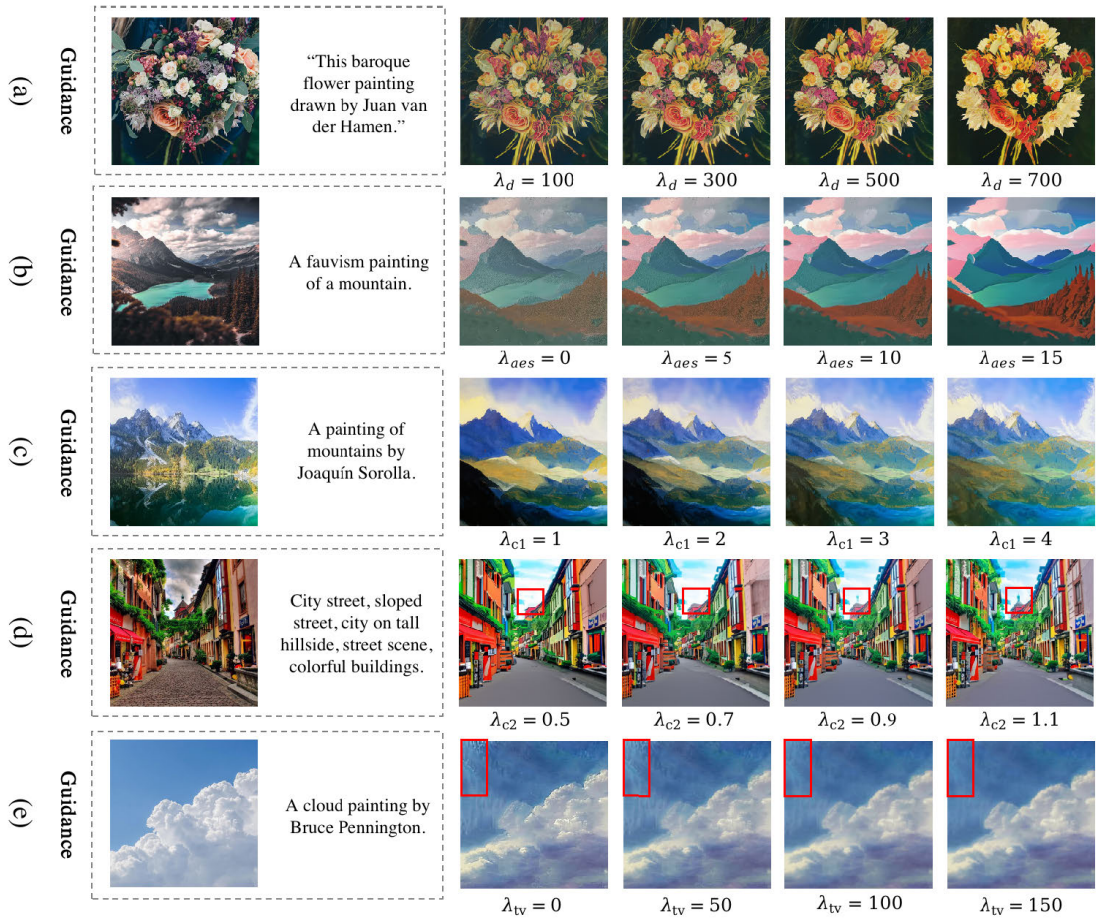


Fig. 7. Results of ablation studies on loss functions. (a) Our full model, (b) without $\mathcal{L}_c$, (c) without $\mathcal{L}_{\text{c}_{\text{patch}}}$, (d) without λ_{tv}, and (e) without $\mathcal{L}_{\text{aes}}$.

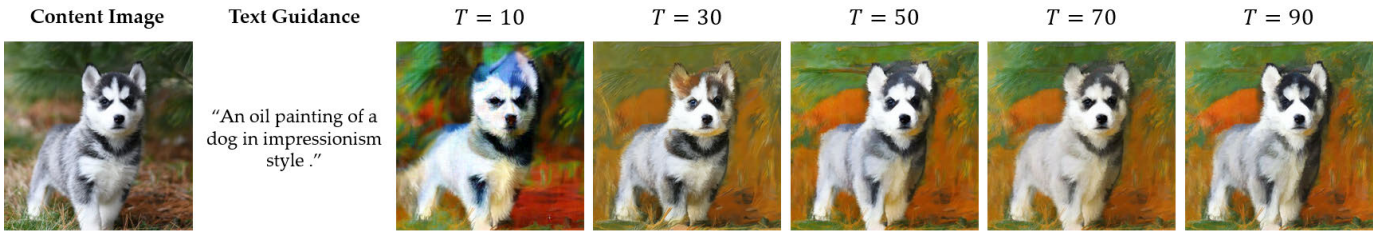


Fig. 8. Results of different diffusion inference steps.

TABLE IV
ABLATION STUDY ON CONTENT PRESERVATION

Learnable noise	$\mathcal{L}_{\text{c}}$	$\mathcal{L}_{\text{c}_{patch}}$	LPIPS↓
✗	✓	✓	0.674
✓	✗	✓	0.593
✓	✓	✗	0.521
✓	✓	✓	0.505

the result changes are no longer significant and converge to the upper mass limit. To maintain the structure of the input content, we find it beneficial when the number of free diffusion steps T_1 of learnable noise is more than the number of reverse sampling steps T. Ablation study results (Fig. 9) show that at $T = 50$. As shown in Table IV, LPIPS increased by 0.169 in the absence of learnable noise. The content of the stylized results is better maintained as T_1 increases gradually, and the image quality no longer changes significantly after a certain number of diffusion steps. To ensure the quality of the results and the efficiency of the generation process, the number of free diffusion steps T_1 used in the experiments of this work is 150 and the number of inference process steps T is 50.

4) Effect of the CLIP Guidance Scale: CLIP guidance scale λ_d determines how consistently the results match the text prompts. We compared the generated results by taking different λ_d values to verify the impact of text guidance. As shown in Fig. 10, the generated results are more consistent with the input content image when using a smaller λ_d. As λ_d is gradually increased, the generated results are brought closer to the semantic content of the prompt and

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

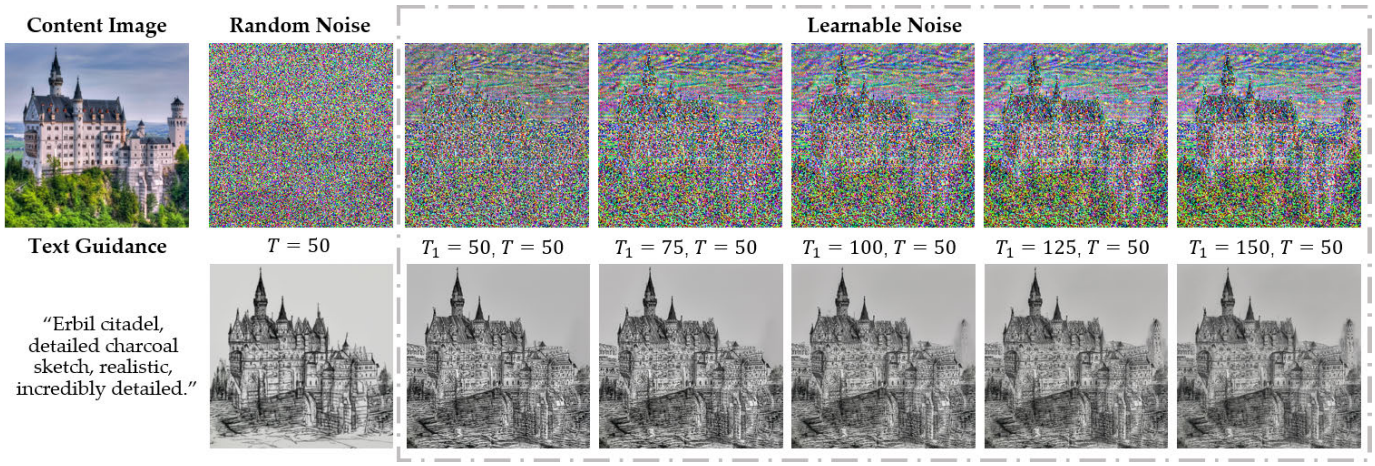


Fig. 9. Introduction of noise induces alterations in image features, thereby influencing the reconstruction outcomes of the image content. The visualization graph provides a comparative analysis of the impact of random noise and learnable noise on the resulting content of the produced images.

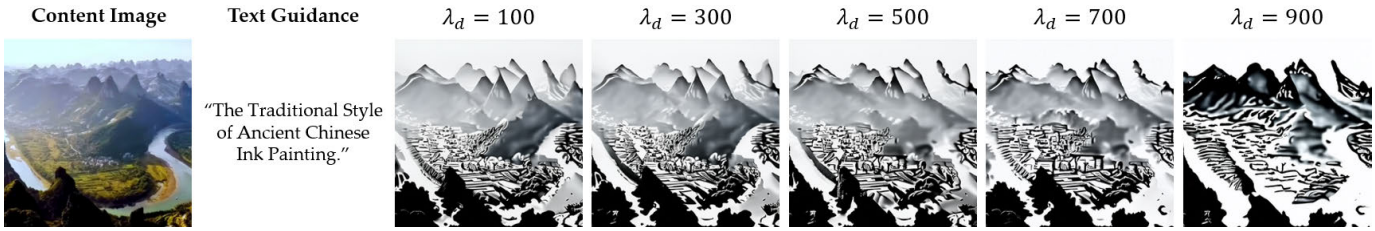


Fig. 10. Effect of the CLIP guidance scale λ_d on the generation results.

Fig. 11. Ablation study of the mixed dataset results.

the user's requirements. However, when the value of λ_d is too large, the discrepancy between the stylized result and the content of the input image increases, which will cause the stylization to fail. Therefore, with an appropriate CLIP guidance scale value, the result can be made to match the text prompt content while keeping the content of the input image.

5) Mixed Dataset Results: As shown in Fig. 11, we conducted ablation experiments on a mixed dataset. The inclusion of both the CC12M dataset, primarily composed of nature images, and the WikiArt dataset, consisting of art images, introduced a substantial scale disparity between the two datasets. Consequently, combining these datasets resulted in an imbalance in the expression of artistic features, wherein the art features were inadequately represented, and the potential of the art image dataset remained underutilized. This challenge was further exacerbated by the limited availability of online art images, hindering the comprehensive representation of diverse art features, as demonstrated by the aforementioned findings in our research. To tackle this issue, we propose a two-tier architectural model that aims to overcome the aforementioned limitations and improve the incorporation of art features in the stylization process.

F. Discussion

Further experiments and comparisons revealed that the modification of the content is related to the chosen style. As Fig. 12 shows, the content of the input image in the traditional styling task remains influenced by the style image, and the content remains influenced by the input text in the text-guided styling task. Compared with the SOTA image/text-guided style transfer approaches [11], [21], [29], [54], [55], the dual diffusion pathway in DiffStyler focuses on the decoupling of content and style representation. As a result, we can generate more natural and fine-grained structures in different styles that are significantly better than other styles (e.g., castle windows and outlines). For a long time, balancing content and style preservation has been challenging for image style transfer. Nevertheless, with text input, the user can further improve the results by changing the text prompt.

G. Failure Cases

While our approach demonstrates high-quality results for text-guided image stylization, it is important to acknowledge its limitations. As illustrated in Fig. 13, we present several instances where our model fails to produce satisfactory results. In the first row, there is a partial disappearance of the cat's fur; in the second row, the distinction between "river," "land," and "mountain" becomes less apparent; and in the third row, the windows of the house vanish. We have observed that the

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

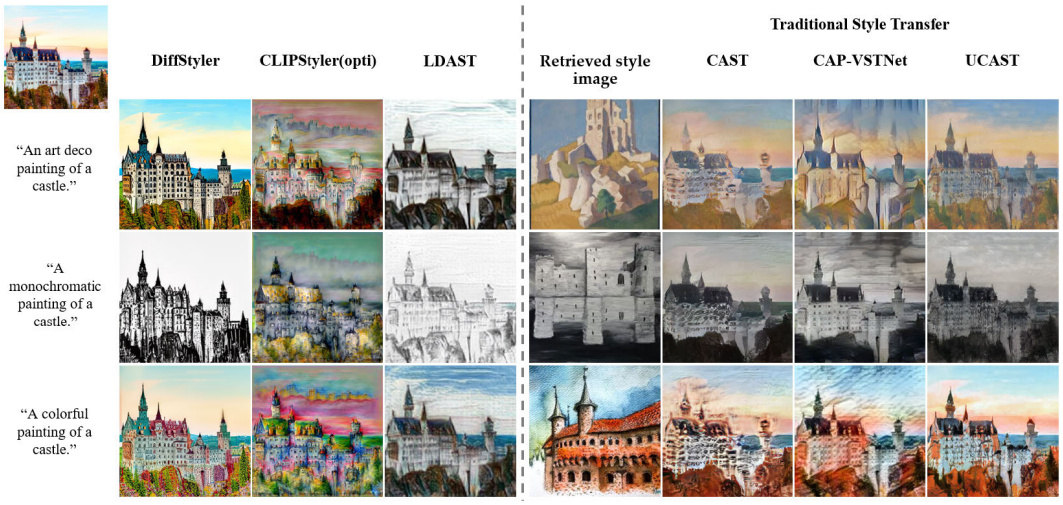


Fig. 12. Comparison experiments for discussion.

Fig. 13. Failure cases of DiffStyler.

generated results primarily emphasize the elements mentioned in the text, potentially causing other elements to be overlooked and subsequently disappear.

In conclusion, these failure cases highlight the potential drawback of our approach, which relies on textual guidance. This reliance may lead to a limited focus on elements that are not explicitly mentioned or implicitly referenced. Consequently, important visual details or contextual relationships may be omitted, distorted, or inadequately represented. To address these limitations, it is crucial to enhance the model's capacity to capture and incorporate relevant contextual information beyond explicit textual guidance. This improvement would enable a more comprehensive and faithful stylization process.

V. Conclusion and Future Work

In this work, we propose a novel controllable dual diffusion framework, namely DiffStyler, for text-driven image stylization. Our method contains a dual diffusion processing architecture to achieve a tradeoff between abstraction and realism of the stylization results. To eliminate the destructive effect of noise on the content of the input image, we use the results of the learnable diffusion process based on the content image to replace the random noise. As the novel baseline for single text-driven style transfer using a diffusion model, the DiffStyler model has a highly distinguished ability to represent artistic style and provides new insights into the challenging problem of style transfer. At present, the sample-time speed of our method is not as fast as some GAN-based approaches. Performing diffusion processes in latent space to speed up the computation will be a possible future research direction. Although we present high-quality results under a single text condition, our method is not without limitations. In particular, the input content may not be fully preserved in the results (see Section IV-G). We plan to investigate more fine-grained control over stylization results in the future.

In addition, DiffStyler has not been able to implement video styling at this time, so we consider it very promising to extend the approach further into the video data domain. We envision that possible future implementations are as follows: one approach is that noise can be used as crucial information for generating each frame in a video to achieve spatiotemporal consistency in video stylization. Another approach is that more realistic and high-quality video reconstruction can be achieved by introducing noise constraints in the video reconstruction process. The potential for utilizing our approach to enhance the visual esthetics and artistic expression of moving images is an exciting direction for future exploration.

References

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[2] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[3] Z. Ma et al., "Dual-affinity style embedding network for semantic-aligned image style transfer," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 7404–7417, Oct. 2023.

[4] S. Jiang, J. Li, and Y. Fu, "Deep learning for fashion style generation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4538–4550, Sep. 2022.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

[5] X. Kong et al., "Exploring the temporal consistency of arbitrary style transfer: A channelwise perspective," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 2023, doi: 10.1109/TNNLS.2022.3230084.

[6] C. Liang, M. Zhu, N. Wang, H. Yang, and X. Gao, "PMS-GAN: Parallel multistage GANs for face image translation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 2023, doi: 10.1109/TNNLS.2022.3233025.

[7] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

[8] H. Wang, Y. Li, Y. Wang, H. Hu, and M.-H. Yang, "Collaborative distillation for ultra-resolution universal style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1857–1866.

[9] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "StyleCLIP: Text-driven manipulation of StyleGAN imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2065–2074.

[10] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or, "StyleGAN-NADA: CLIP-guided domain adaptation of image generators," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 141:1–141:13, 2022.

[11] G. Kwon and J. C. Ye, "CLIPstyler: Image style transfer with a single text condition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18041–18050.

[12] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[13] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.

[14] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 8780–8794.

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

[16] A. Nichol et al., "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," 2021, *arXiv:2112.10741*.

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.

[18] N. Huang, F. Tang, W. Dong, and C. Xu, "Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 1085–1094.

[19] X. Wu, "Creative painting with latent diffusion models," 2022, *arXiv:2209.14697*.

[20] R. Gal et al., "An image is worth one word: Personalizing text-to-image generation using textual inversion," 2022, *arXiv:2208.01618*.

[21] T.-J. Fu, X. E. Wang, and W. Y. Wang, "Language-driven artistic style transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 717–734.

[22] J. An, S. Huang, Y. Song, D. Dou, W. Liu, and J. Luo, "ArtFlow: Unbiased image style transfer via reversible neural flows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 862–871.

[23] Y. Deng et al., "StyTr^2: Image style transfer with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11316–11326.

[24] Y.-S. Liao and C.-R. Huang, "Semantic context-aware image style transfer," *IEEE Trans. Image Process.*, vol. 31, pp. 1911–1923, 2022.

[25] D. Diffusion. *Disco Diffusion*. Accessed: Jun. 25, 2022. [Online]. Available: https://github.com/alembics/disco-diffusion

[26] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22500–22510.

[27] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 694–711.

[28] T. Lin et al., "Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5137–5146.

[29] Y. Zhang et al., "Domain enhanced arbitrary image style transfer via contrastive learning," in *Proc. Special Interest Group Comput. Graph. Interact. Techn. Conf. Proc.*, Aug. 2022, pp. 12:1–12:8.

[30] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.

[31] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 2256–2265.

[32] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2019, pp. 11895–11907.

[33] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, pp. 6840–6851, 2020.

[34] G. Kim, T. Kwon, and J. C. Ye, "DiffusionCLIP: Text-guided diffusion models for robust image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2416–2425.

[35] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 35, 2022, pp. 36479–36494.

[36] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8162–8171.

[37] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[38] L. Liu, Y. Ren, Z. Lin, and Z. Zhao, "Pseudo numerical methods for diffusion models on manifolds," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[39] E. Chu, S.-Y. Lin, and J.-C. Chen, "Video ControlNet: Towards temporally consistent synthetic-to-real video translation using conditional image diffusion models," 2023, *arXiv:2305.19193*.

[40] W. Huang, S. Tu, and L. Xu, "PFB-Diff: Progressive feature blending diffusion for text-driven image editing," 2023, *arXiv:2306.16894*.

[41] Z. Luo et al., "Retraction notice: VideoFusion: Decomposed diffusion models for high-quality video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10209–10218.

[42] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3557–3567.

[43] C. Katherine. *V-Diffusion-PyTorch*. Accessed: Nov. 21, 2022. [Online]. Available: https://github.com/crowsonkb/v-diffusion-pytorch

[44] W. Volunteer Team. *Wikiart Dataset*. Accessed: Dec. 10, 2022. [Online]. Available: https://www.wikiart.org/

[45] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 319–345.

[46] C. Katherine. *Simulacra Aesthetic Models*. Accessed: Oct. 22, 2022. [Online]. Available: https://github.com/crowsonkb/simulacra-aesthetic-models

[47] J. D. Pressman. *Simulacra Aesthetic Captions*. Accessed: Oct. 22, 2022. [Online]. Available: https://github.com/JD-P/simulacra-aesthetic-captions

[48] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[49] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[50] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016, *arXiv:1605.07146*.

[51] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[52] S. Liu et al., "AdaAttN: Revisit attention mechanism in arbitrary neural style transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6629–6638.

[53] Y. Deng, F. Tang, W. Dong, H. Huang, C. Ma, and C. Xu, "Arbitrary video style transfer via multi-channel correlation," in *Proc. 35th AAAI Conf. Artif. Intell. (AAAI)*, 2021, vol. 35, no. 2, pp. 1210–1217.

[54] L. Wen, C. Gao, and C. Zou, "CAP-VSTNet: Content affinity preserved versatile style transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18300–18309.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.

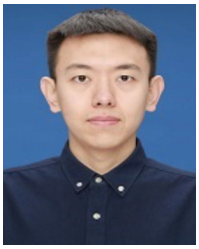
[55] Y. Zhang et al., "A unified arbitrary style transfer framework via adaptive contrastive learning," *ACM Trans. Graph.*, vol. 42, no. 5, pp. 169:1–169:16, Oct. 2023.
[56] B. Romain et al. *Clip Retrieval System*. Accessed: May 29, 2023. [Online]. Available: https://github.com/rom1504/ clip-retrieval
[57] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017.

Nisha Huang received the B.Sc. degree in aerospace information engineering from Beihang University, Beijing, China, in 2021. She is currently pursuing the M.E. degree with the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, and the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing.

Her research interests include multimedia analysis, computer vision, and machine learning.

Yuxin Zhang (Graduate Student Member, IEEE) received the B.Sc. degree in automation from Tsinghua University, Beijing, China, in 2020. She is currently pursuing the Ph.D. degree with the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, and the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing.

Her research interests include computer vision, computer graphics, and machine learning.



Fan Tang received the B.Sc. degree in computer science from North China Electric Power University, Beijing, China, in 2013, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2019.

He is an Assistant Professor with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer graphics, computer vision, and machine learning.

Chongyang Ma received the B.S. degree from the Fundamental Science Class (Mathematics and Physics), Tsinghua University, Beijing, China, in 2007, and the Ph.D. degree in computer science from the Institute for Advanced Study, Tsinghua University, in 2012.

He is currently a Research Lead with Kuaishou Technology, Beijing. His research interests include computer graphics and computer vision.

Haibin Huang received the B.Sc. and M.Sc. degrees in mathematics from Zhejiang University, Hangzhou, China, in 2009 and 2011, respectively, the Ph.D. degree in computer science from UMass Amherst, Amherst, MA, USA, in 2017.

He is currently a Senior Staff Research Scientist with Kuaishou Technology, Beijing, China.

Weiming Dong (Member, IEEE) received the B.Sc. and M.Sc. degrees from Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from the University of Lorraine, Nancy, France, in 2007.

He is a Professor at the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences, Beijing. His research interests include image synthesis, image recognition, and computational creativity.

Changsheng Xu (Fellow, IEEE) is a Professor at the State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences.

Dr. Xu received the Best Associate Editor Award of *ACM Transactions on Multimedia Computing, Communications and Applications* in 2012 and the Best Editorial Member Award of *ACM/Springer Multimedia Systems Journal* in 2008. He has served as an Associate Editor, Guest Editor, General Chair, Program Chair, Area/Track Chair, Special Session Organizer, Session Chair, and a Transactions on Professional Communication (TPC) Member for over 20 IEEE and ACM prestigious multimedia journals, conferences, and workshops. He is the International Association for Pattern Recognition (IAPR) Fellow and the ACM Distinguished Scientist.

Authorized licensed use limited to: INSTITUTE OF COMPUTING TECHNOLOGY CAS. Downloaded on October 19,2026 at 12:40:41 UTC from IEEE Xplore. Restrictions apply.