FreeStyler: A Free-Form Stylization Method via Multimodal Vector Quantization

WuQin Liu[1,2][0009−0002−1787−1593], MinXuan Lin[3][0009−0006−5130−5754], HaiBin Huang[3][0000−0002−7787−6428], ChongYang Ma[3][0000−0002−8243−9513], and WeiMing Dong[2][0000−0001−6502−145X]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 101408, China
[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[3] Kuaishou Technology, Beijing 100085, China

Abstract. Image stylization refers to the process of transforming an input image into a new one, while retaining its original content but in different styles. However, most existing works only support single-modal guidance, which is not ideal for real-world applications. To tackle this limitation, we propose FreeStyler, a flexible framework for image stylization that is capable of handling various input scenarios. Our approach goes beyond the traditional approach of relying on content and style images to generate a stylized image and supports situations where these references are absent. Specifically, in such cases, FreeStyler allows performing the stylization through text or audio information. The core of FreeStyler is a vector quantized style transfer framework that encodes content and style information into a shared discrete latent feature space, followed by a stylization transformer for style fusion and an image decoder for stylized image reconstruction. To enable free-form stylization, we introduce a novel pseudo-paired token predictor that can estimate tokens from varying input forms without the need for additional text or audio data. Specifically, we leverage Contrastive Language-Image Pre-training (CLIP) as prior knowledge to align discrete representations across different modalities and train the framework using an image and pseudo caption pair provided by Bootstrapping Language-Image Pre-training (BLIP). Through qualitative and quantitative experiments, our method has demonstrated superior performance compared to state-of-the-art stylization methods.

Keywords: Vector quantization · multi-modal stylization · image stylization · contrastive learning.

1 Introduction

In this paper, we study the problem of image stylization, which provides a convenient way for amateurs to create vivid artwork without requiring professional skills. One aspect of image stylization work is style transfer, which aims to transform the semantic texture of a style image into a given content image. Initially, Gatys et al. [11] employed

Fig. 1: Multimodal guided stylization results generated by our method. In the first row, we showcase images, text, or audio that serve as style information. The first column, on the other hand, presents content information. This approach enables users to guide the stylization process flexibly by selecting input modalities according to their specific content and style preferences.

a pre-trained VGG network to accomplish style texture transfer by matching the Gram matrices of the content and style images. Subsequent works [15, 37, 21, 42] have built upon and improved this approach. Although these approaches can generate visually appealing artworks by transferring styles of arbitrary artworks into real-world photos, they necessitate the availability of both a reference style image and a content image. However, we argue that this requirement may be impractical in many real-life scenarios where content or style images are unavailable. Consider, for instance, situations where users aim to transfer a specific texture style that only exists in their imagination, without a corresponding content figure. Alternatively, they may lack artistic pictures to stylize a particular photo. Furthermore, users may wish to provide content information in the form of textual descriptions or sounds produced by objects, or convey style information through texture descriptions or sound-to-color synesthesia, or even utilize both modalities simultaneously. These scenarios pose significant challenges to style transfer methods and present noteworthy complexities in image stylization tasks.

To address these challenges, we propose FreeStyler, a general and unified framework for image stylization that explicitly decouples content and style information and supports free-form modality guidance. The key component of FreeStyler is a vector quantization module that bridges the application of style transfer across multi-modalities. Our approach adopts a two-stage learning strategy. In the first stage, we develop a vector quantized style transfer network that employs a shared weight encoder and codebook to project information into a discrete and joint latent space. This is followed by a stylization transformer and decoder module to generate stylization results. In the second stage, we train a pseudo-paired token predictor to reconstruct tokens from various input forms. We leverage the strong prior knowledge of CLIP to align the space using an image and

pseudo caption pair generated by BLIP. Figure 1 shows nine combinations of image, text, and audio information used as content and style inputs. Additionally, our model facilitates a range of applications, including mask-guided stylized inpainting and text-driven stylized image editing. In this article, we highlight the results of text-to-image and image-to-image stylization. For more examples, please refer to the supplementary materials.

To summarize, our main contributions are as follows:

- We propose FreeStyler, the first unified image stylization framework that supports free-form multi-modal input as content or style guidance. FreeStyler is designed to cater to a wide range of stylization application scenarios, offering users with enhanced convenience and efficiency.
- To enable free-form multimodal control in image stylization, we present a vector quantization-based approach proposed that establishes a discrete and joint latent space for content and style. This is achieved through a combination of contrastive and adversarial learning techniques. Additionally, we proposed pseudo-paired token predictor that uses a CLIP-based condition transformer to align features across multiple modalities and train the model using an image and pseudo caption pair generated by BLIP.
- Qualitative and quantitative experiments demonstrate that our method can achieve better results in both reference-guided image-to-image style transfer and text-guided image stylization. Moreover, we show the editability of our approach through several novel applications.

2 Related Work

Style transfer. Inspired by Gatys et al.[11], who defined the distance between two images in the feature space of the VGG network as a measure of style using Gram matrices, several works[15, 37, 21] proposed training end-to-end models by incorporating reasonable content and style constraints without optimization. After that, the arbitrary style transfer (AST) problem has received increasing attention . Li et al. [22, 23] employed the whitening and coloring transform (WCT) method to migrate the distribution of style features, further Yoo et al. [42] proposed a wavelet transform-based WCT to improve the performance in realistic style transfer task. Huang et al.[14] introduced Adaptive Instance Normalization (AdaIN), which replaces the mean and standard deviation of content features with those of style features. With the development of attention mechanism, Park et al. [29] adopted a style-attentional network (SANet) that effectively and flexibly decorates local style patterns based on the semantic spatial distribution of content images. Chen et al. [2] proposed an internal-external and contrastive learning style transfer (IEST) algorithm incorporating two contrastive losses and Zhang et al. [45] proposed CAST approach to address the problem of local distortion and incomplete styles guided by second-order statistics. Furthermore, The QuantArt method by Huang et al. [13] utilizes vector quantization to discretize the latent representation of generated artworks, allowing for flexible control over content preservation, style similarity, and visual fidelity. Later, Liu et al.[28] proposed adaptive attention normalization (AdaAttN) module to enhance visual quality and extend it for video style transfer. Deng

et al. [5] employed a transformer-based style transfer framework instead of convolution to focus on the relation of global features. Chen et al. [2] proposed an internal-external and contrastive learning style transfer (IEST) algorithm incorporating two contrastive losses and Zhang et al. [45] proposed CAST approach to address the problem of local distortion and incomplete styles guided by second-order statistics. Furthermore, The QuantArt method by Huang et al. [13] utilizes vector quantization to discretize the latent representation of generated artworks, allowing for flexible control over content preservation, style similarity, and visual fidelity.

In practice, users do not have access to reference style images but still want to mimic a particular painter's texture. Some work [27, 25] treated the works of the same painters as a domain and control the results by attribute labels. Kwon et al. [17] proposed CLIPstyler which allows users to enter text to change the style of the content image. Meanwhile, Lee et al. [19] introduced LISA, which uses audio to locate and edit images. However, existing frameworks do not support trimodal input of text, images, and audio, respectively, in the content and style aspects. Therefore, we propose a generic framework to overcome this limitation.

Multimodal-to-image generation. The field of multimodal image generation encompasses two main areas: text-to-image generation and audio-to-image generation. The former has been advancing rapidly with the introduction of large-scale pre-trained text-image embedding models like CLIP [31] and the development of diffusion models. However, the latter has received relatively less attention in comparison.

Early approaches to text-guided image generation involved training a convolutional generator that directly predicts pixels from a given text embedding. Later, transformer-based generators [7, 33, 9]that map text embeddings to discrete representations of images (VQGAN [9] or VQ-VAE [38, 34]) achieve significantly improvement on visual quality. And then, diffusion models have been shown to be superior in image generation [6] and gradually take over the main position in digital art generation [12, 8, 35, 36, 32]. However, diffusion models have difficulty in generating an image of the specified content, which is consistent with the style of the painting based on a particular painting. Although Kwon et al. [18] attempted to decouple the content and style of an image, it is still a far cry from the performance of traditional style transfer. In terms of audio-guided image generation, attempts have been made to combine wav2clip [40] and VQGAN to generate audio-guided images. However, this area still requires further research and development. Therefore, our approach combines transformer and CLIP to enable users to input arbitrary, images, text and audio as content or style information. Inspired by [39], we train a standard GPT transformer module using pseudo captions to align space without the need for paired text-image data.

Vector-quantized image representation. The vector-quantized generative models, including VQVAE [34, 38] and VQGAN [9], were originally designed to achieve compact and efficient image modeling. VQVAE utilizes a vector-quantized autoencoder architecture to represent images using a discrete set of tokens. Building upon VQVAE, VQGAN further enhances the model with an adversarial learning scheme. By discretizing continuous image features, these models offer several advantages, such as enabling smoother

post-processing operations. This discretization also empowers the utilization of alternative modalities or expressions to guide the process of image generation, ultimately resulting in more precise and personalized outcomes. In line with these advantages, our approach leverages the power of discretization to expand the applications of style transfer. By providing a more convenient and user-friendly framework, our method aims to facilitate style transfer tasks and empower users with greater control and flexibility over the generated results.

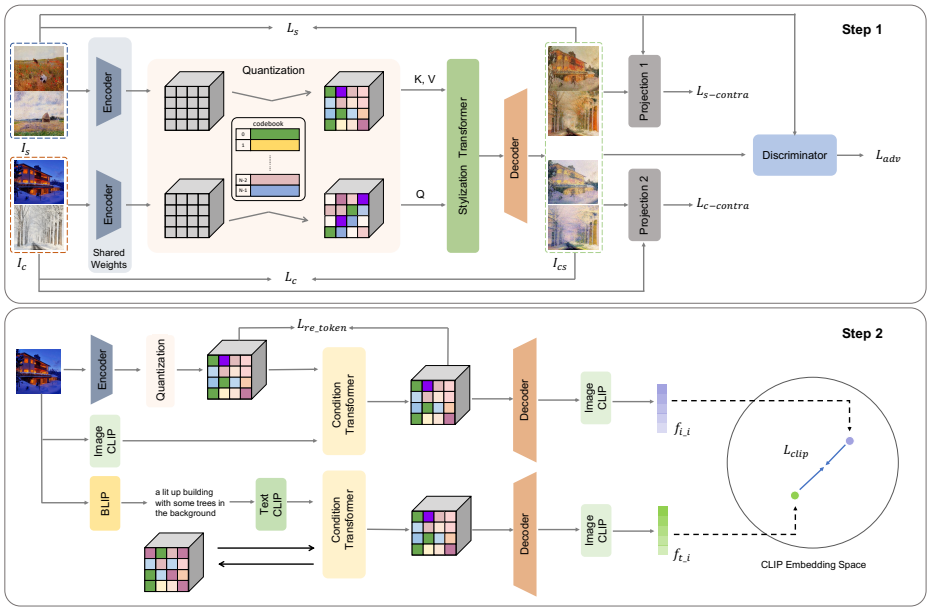

Fig. 2: An overview of our method. It includes a VQGAN model to discretize features, a stylization transformer for style fusion, as well as a condition transformer as the multimodal input guidance module. There are two steps in the training stage: (1) The first step is to train a vector quantization framework, which consists of a pretrained VQGAN encoder and a codebook to quantize the image features. The stylized image is then decoded and sent to two projection modules, which calculate the contrastive loss for content and style respectively. (2) In the second step, we utilize BLIP to provide the pseudo caption for each image and freeze all the modules except the condition transformer to learn the reconstruction of tokens under the CLIP constraint, allowing the user to freely choose the input modality in the inference step.

3 Method

Our goal is to achieve image, text, or audio guided stylization in arbitrary domains for both content and style. To achieve this, our proposed method follows two main steps:

(1) training a vector quantized style transfer framework by contrastive and adversarial learning. (2) training a multi-modal guided token predictor in a pseudo-paired approach. Figure 2 illustrates the overall framework of our method.

3.1 Vector Quantization Framework

In our approach, we employ a vector quantization strategy to encode the features of the input image. This strategy ensures a consistent representation between the image and text modalities and makes it easier to construct and manipulate the content and style space in their respective branches. Since all information is discretely represented, we utilize a stylization transformer as an intermediate module to fuse the content and style features. In the context of image style transfer, contrastive learning has been proved to be effective [2]. In this technique, stylized images with the same style should exhibit closer relationships in the style feature space compared to those with different styles. Similarly, stylized images based on the same content should have higher content similarity compared to those based on different content images. To enhance the decoupling of content and style in the feature space, we introduce contrastive loss, which encourages the desired separation and distinctiveness of content and style representations.

Vector quantization representation. First of all, we denote the content input image as $I_c \in \mathbb{R}^{H\times W\times 3}$ and the style input image as $I_s \in \mathbb{R}^{H\times W\times 3}$. A shared-weights encoder E is applied to encode both content and style. Then, we construct a perceptually rich codebook $\mathcal{Z} = \{z_k\}_{k=1}^{K} \subset \mathbb{R}^{n_z}$ by adopting a strategy similar to VQGAN [9], where n_z is the dimension of code. In the subsequent training, we freeze the encoder E and the codebook $\mathcal{Z}$ and only optimize the stylization transformer and the decoder G. We use the weights of a pre-trained VQGAN as an initialization for the encoder E and the codebook $\mathcal{Z}$ to reduce computational cost and to provide a good prior. Therefore, given an input content image I_c and a style image I_s, they are mapped into spatial embeddings by the encoder and are then discretized to get z_{qc} and z_{qs} according to the codebook.

It is worth emphasizing that the primary objective of VQ-GAN during the training stage is to reconstruct the original image while preserving its details. To ensure the best reconstruction effect for both real and artistic domains, we carefully selected the pre-trained model from the gallery[4] based on its lowest reconstruction loss as our initial loading checkpoint. This decision guarantees that image details are maintained during the discretization process. In the initial training phase, we keep the encoder and codebook fixed and focus on fine-tuning the remaining modules. We employ a domain discriminator during this phase, which is different from the VQ-GAN discriminator. This domain discriminator specifically enhances the artistic quality of the generated images. It is noteworthy that existing methods similar to VQGAN-CLIP [3] utilize checkpoints trained on ImageNet. Nevertheless, the artistic images generated using our approach still produce satisfactory results. Consequently, using the same codebook does not degrade our generation performance.

[4] https://github.com/CompVis/taming-transformers

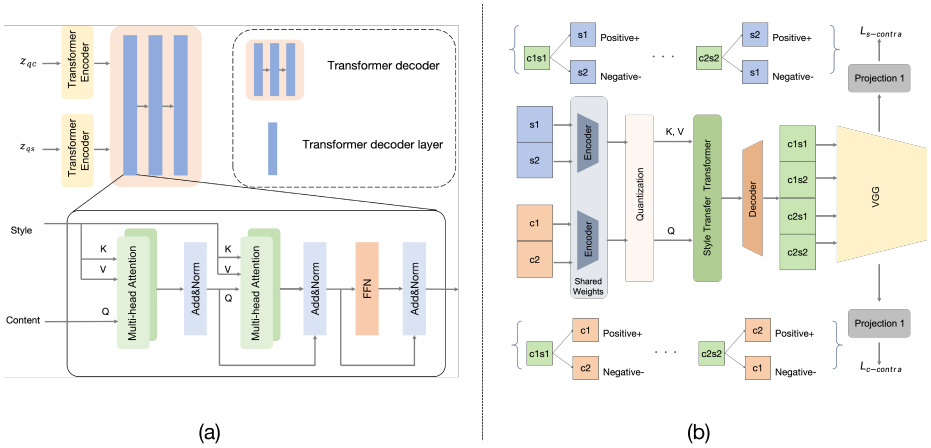

Fig. 3: Two important modules in the framework. (a) The Stylization transformer module proposed for feature style transfer. (b) The contrastive loss used to learn the stylization-to-stylization relations.

Stylization transformer. Discretized image features can be tokenized due to their language-like properties, making transformers a natural choice for feature fusion modules. Recent methods [5, 43] have demonstrated that transformer-based style transfer frameworks are better at extracting and preserving global information than CNN-based frameworks. In this work, we utilize a vanilla cross-attention transformer as the feature fusion module backbone, injecting user-specified style information to modulate content signals.

As illustrated in Fig 3(a), we utilize two distinct transformer encoders to generate domain-specific sequences for both content and style. Subsequently, a multi-layer transformer decoder is employed to stylize the content sequence by incorporating the style sequence. This process involves taking the discretized features of the content image z_{qc} as the query (Q), while the discretized features of the style image z_{qs} are considered as the key (K) and value (V). Both of them are sent to the stylization transformer in different branches and appended the corresponded position embedding. Finally, we obtain the stylized image I_{cs} using the decoder G.

Contrastive learning. In recent years, contrastive learning has proved its superiority in style transfer. For example, Chen et al. [2] adopt contrastive learning to enhance the relations of multi-style-single-content and single-style-multi-content results neglected by traditional style transfer methods and In order to improve the generation quality of stylized images, we also leverage contrastive learning to improve the generation quality of stylized images.

As shown in Fig 3(b), in each training batch, two different content images I_{c_1}, I_{c_2}, and two different style images I_{s_1}, I_{s_2} are processed. We generate all possible combinations of content-style pairs and denote the corresponding results as $I_{c_1 s_1}$, $I_{c_1 s_2}$, $I_{c_2 s_1}$ and $I_{c_2 s_2}$. We utilize a pre-trained VGG19 network as a base projection prior and compute content and style similarity by feeding the features extracted from that into

either a content projection network h_c or a style projection network h_s, comprising a two-layer MLP. The content projection is denoted as $l_c = h_c(\phi_{relu4\_1}(*))$ and the style projection as $l_s = h_s(\phi_{relu3\_1}(*))$, where ϕ_i denotes the i-th layer of features extracted from the pre-trained VGG19 model. To simplify the notation, we set the features of projection space as $C_{c_i s_j} = l_c(I_{c_i s_j})$, and $S_{c_i s_j} = l_s(I_{c_i s_j})$. Our approach involves first pushing the stylized image closer to the original content image and further away from other content images in the content projection space, constructing a meaningful and comparable distance space under relationship constraints from opposing directions. The content contrastive loss can be defined as:

$$\mathcal{L}_{c\_contra} = -\log\left(\frac{\exp\left(\frac{{C_{c1s1}}^T C_{c1}}{\tau}\right)}{\exp\left(\frac{{C_{c1s1}}^T C_{c1}}{\tau}\right) + \exp\left(\frac{{C_{c1s1}}^T C_{c2}}{\tau}\right)}\right). \tag{1}$$

where τ is a temperature parameter. At the same time, in the style projection space, we employ the same strategy for style features. The style contrastive loss can be denoted as:

$$\mathcal{L}_{s\_contra} = -\log\left(\frac{\exp\left(\frac{{S_{c1s1}}^T S_{s1}}{\tau}\right)}{\exp\left(\frac{{S_{c1s1}}^T S_{s1}}{\tau}\right) + \exp\left(\frac{{S_{c1s1}}^T S_{s2}}{\tau}\right)}\right). \tag{2}$$

Finally, we obtain the full contrastive loss $\mathcal{L}_{contra}$ as below:

$$\mathcal{L}_{contra} = \mathcal{L}_{c\_contra} + \mathcal{L}_{s\_contra}. \tag{3}$$

Network training. To ensure the generated results preserve the structure of content branch and texture features of style branch, we construct two perceptual losses. Similar to [11, 15, 1], we use the pre-trained VGG19 model to extract the features of the image and calculate the content loss and style loss. The content loss and style loss is defined as:

$$\mathcal{L}_c = \frac{1}{N_l} \sum_{i=0}^{N_l} \left\| \phi_i(I_{cs}) - \phi_i(I_c) \right\|_2 . \tag{4}$$

where N_l denotes the total number of layers. In our experiments, we use features from the layers of relu4_1 and relu5_1 with equal weights.

Meanwhile, the style loss is defined as:

$$\begin{aligned} \mathcal{L}_s = & \frac{1}{N_l} \sum_{i=0}^{N_l} \left\| \mu(\phi_i(I_{cs})) - \mu(\phi_i(I_s)) \right\|_2 \\ & + \frac{1}{N_l} \sum_{i=0}^{N_l} \left\| \sigma(\phi_i(I_{cs})) - \sigma(\phi_i(I_s)) \right\|_2 . \end{aligned} \tag{5}$$

where μ and σ represent the mean and variance of the extracted features, respectively. In our experiments, we use features from the layers of relu1_1, relu2_1, relu3_1, relu4_1, and relu5_1 with equal weights.

Similar to [24, 29, 46], we use identity loss terms to make network learning more accurate and rich in content and style information, so that the generated results can maintain more content structure and style features. Therefore, we ensure the similarity in the aspects of pixel space and feature space. The identity loss in the pixel space can be defined as:

$$\mathcal{L}_{identity1} = \|I_{cc} - I_c\|_2 + \|I_{ss} - I_s\|_2 . \tag{6}$$

And the identity loss in the feature space is written as:

$$\begin{aligned} \mathcal{L}_{identity2} = & \frac{1}{N_l} \sum_{i=0}^{N_l} \|\phi_i(I_{cc}) - \phi_i(I_c)\|_2 \\ & + \frac{1}{N_l} \sum_{i=0}^{N_l} \|\phi_i(I_{ss}) - \phi_i(I_s)\|_2 . \end{aligned} \tag{7}$$

where I_{cc} is the output image generated when both the content image and the style image are I_c, and I_{ss} is generated similarly using I_s.

In addition, we use Generative Adversarial Network (GAN) [11, 44] to align generated images with the distribution of input art images to learn the human perception of style information, which consists of a generator G and a competing discriminator D. We express the adversarial loss as:

$$\mathcal{L}_{adv} = \mathbb{E}[\log(D(I_s))] + \mathbb{E}[\log(1 - D(I_{cs}))]. \tag{8}$$

So, the overall optimization objective of the model is demonstrated as:

$$\begin{aligned} \mathcal{L}_{step\_1} = & \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_s + \lambda_3 \mathcal{L}_{identity1} + \lambda_4 \mathcal{L}_{identity2} \\ & + \lambda_5 \mathcal{L}_{adv} + \lambda_6 \mathcal{L}_{contra}. \end{aligned} \tag{9}$$

In our experiments, we set $\lambda_1 = 8$, $\lambda_2 = 8$, $\lambda_3 = 70$, $\lambda_4 = 1$, $\lambda_5 = 1$ and $\lambda_6 = 2$ to balance different loss terms.

3.2 Pseudo-Paired Token Predictor

To support multimodal control in content and style branches, we adopt the CLIP encoder as a prior to align different modalities during pseudo-paired training. Specifically, we use a condition transformer to predict the token of the image based on the high-level semantic information extracted by CLIP in an autoregressive manner, as inspired by [39]. Given an input image I, we obtain its embedding e_c using the CLIP image encoder and obtain the token $s = \mathcal{Z}(E(I))$ of its discretization feature through the encoder in the step 1. By using the condition transformer, we restore the original image token, which we denote as $s^{'}$. Using high-level semantic information from CLIP as a guide, we obtain low-level spatial semantic information built from the codebook. Therefore, to achieve our goal in this step, we aim to introduce a token reconstruction loss to enhance the consistency between the original and reconstructed image tokens.

The token reconstruction loss $\mathcal{L}_{re\_token}$ is expressed based on the cross-entropy loss $\mathcal{L}_{CE}$ as below:

$$\mathcal{L}_{re\_token} = \mathcal{L}_{CE}(s, s^{'}). \tag{10}$$

To enhance the performance of image generation, we utilize BLIP to generate a pseudo-caption for the input image I. We then generate the corresponding image $I_t^{'}$ using the network and calculate its similarity with the reconstructed image $I^{'} = G(s^{'})$ in the CLIP space. Thus, we defined the similarity loss as follows:

$$\mathcal{L}_{clip} = \left\| E(I_t^{'}) - E(I^{'}) \right\|_2. \tag{11}$$

Thus, our overall loss in step 2 is demonstrated as:

$$\mathcal{L}_{step\_2} = \lambda_7 \mathcal{L}_{re\_token} + \lambda_8 \mathcal{L}_{clip}. \tag{12}$$

where $\lambda_7 = 1$ and $\lambda_8 = 1000$.

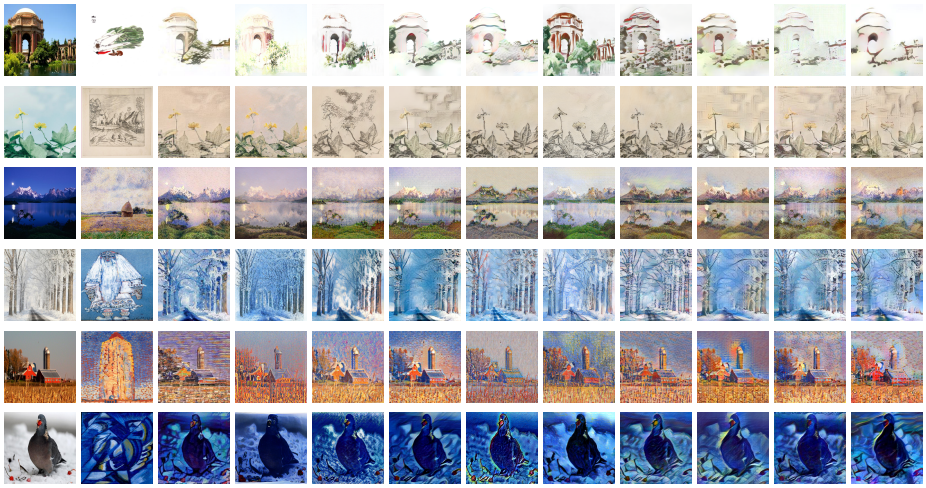

Fig. 4: Qualitative comparisons of image style transfer results with several state-of-the-art methods.

4 Experiments

4.1 Implementation Details

In the first stage of our training process, we use MS-COCO [26] as the content dataset and WikiArt [30] as the style dataset. During the training process, all images are first

enlarged to 512x512 resolution and then randomly cropped to 256x256 size. We adopt the Adam [16] optimizer and the warm-up adjustment strategy. Meanwhile, the learning rate is set as 5e-4. Then, we train our model on 8 V100 machines with a batch size of 2 and set the number of full training rounds as 40 epochs. In the second stage, we use the MS-COCO dataset as input and train the condition transformer about 200 epochs, the learning rate is also set as 5e-4. We start $\mathcal{L}_{clip}$ after 50 epochs. In this article, we set a standard GPT-2 model of 24 layers as the backbone of the condition transformer.

Baselines. For image style transfer, we compare our approach to AdaIN [14], ArtFlow [1], MCCNet [4], AdaAttN [28], IEST [2], StyleFormer [41], StyTr^2 [5], CAST [45] and QuantArt [13]. All baselines are trained using publicly available implementations with default configurations. For text-guided style transfer, we compare with three state-of-the-art text-guided stylization methods, including CLIPstyler [17], VQGAN-CLIP [3] and LDAST [10].

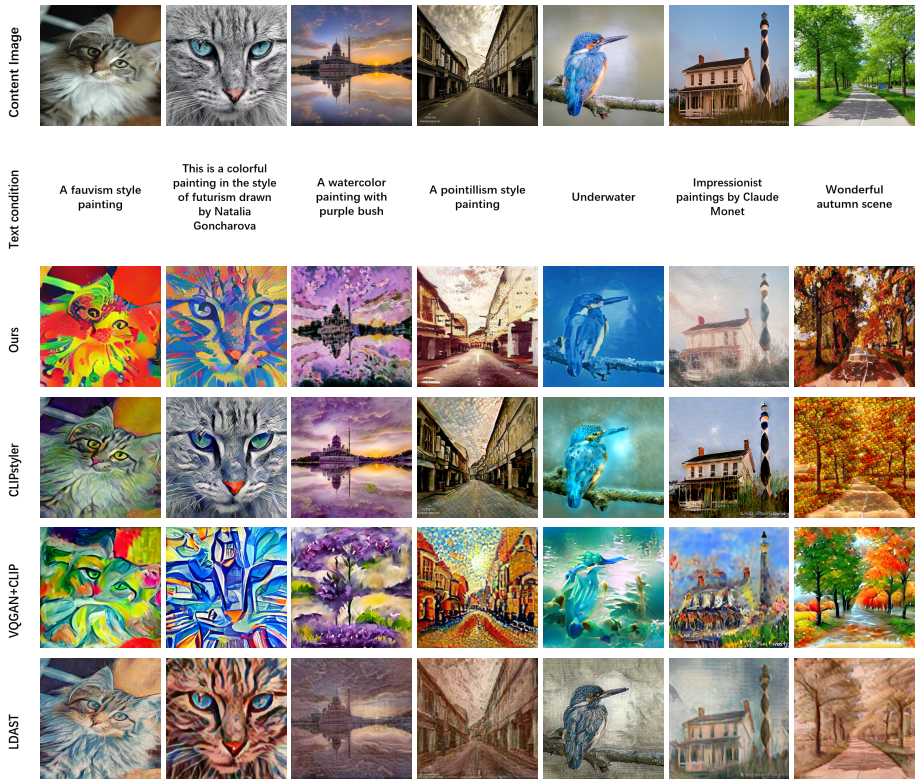

Fig. 5: Qualitative comparisons of text-guided image stylization results with several state-of-the-art methods.

4.2 Qualitative Results

Image Style Transfer We first show the visual results of qualitative comparison of our method with state-of-the-art methods in Figure 4. The comparison shows the superiority of Freestyler in terms of visual quality and all methods generate stylization images under the same reference image guidance. The results of AdaIN [14] do not preserve the style well and appear crack artifacts, resulting in poor quality of the generated results (e.g., the 3^{rd}, $5^{th}$and 6^{th} rows) and ArtFlow [1] may result in insufficient style or inaccuracy (e.g., the $1^{st}$and 6^{th} rows). MCCNet [4] can well preserve the content, but there is often a problem of overflow around objects, namely halo artifacts (e.g., the 3^{rd}, $6^{th}$and 7^{th} rows). AdaAttN [28] cannot transfer some colors to the generated image, so there will be a style gap between results and reference images (e.g., the 1^{st}, $3^{rd}$and 5^{th} rows). IEST [2] can generate high-quality stylization, but it has obvious color distortion problems (e.g., the 1^{st}, $3^{rd}$and 7^{th} rows). StyleFormer [41] loses a lot of detail information and generates a number of redundant colors (e.g., the 3^{rd}, $5^{th}$and 6^{th} rows). StyTr^2 [5] can generate results with high fidelity, but the learning of style is still difficult, and many colors and textures cannot be transferred (e.g., the 5^{th}, $6^{th}$and 7^{th} rows). CAST [45] has a great improvement on both style consistency and content fidelity, however, for some special styles, it is hard to balance the relationship between the style and content, and some detailed textures are distorted(e.g., the 2^{nd}, $5^{th}$and 7^{th} rows). QuantArt [13] focuses more on preserving the details of the content, thus retaining the style of many elements such as colors and textures (e.g., the 1^{st}, $4^{th}$and 5^{th} rows).

In contrast, our method achieves high quality in both content fidelity and style learning, even with styles containing blank space or large color blocks. By using vector quantization, our decoder only needs to learn a limited number of mapping relationships, improving the quality of stylization images.

Text-Guided Image Stylization Figure 5 compares our text-guided stylization results with existing baselines that use the same content input. Our method outperforms the baselines in terms of overall quality. CLIPstyler [17] requires model finetuning for each style, which is inconvenient and time-consuming. Additionally, it is challenging for it to learn styles with color blocks (e.g., 1st and 2nd column). VQGAN-CLIP [3] is also an optimization-based method, but its biggest drawback is that the content fidelity gradually decreases as the number of optimization steps increases. In this study, we set the number of optimization steps as 10 to strike a balance, but the content information still cannot be retained as well as our method (e.g., 7th and 8th column). LDAST [10] aims to learn the visual attributes or emotional effects of styles, but it lacks the ability to understand a specific painting style. We observe poor performance in various cases with descriptions from open domains.

In contrast, our method excels in transferring both artistic styles and intentions (e.g., underwater and autumn) effectively to the content image, resulting in visually appealing outcomes. Therefore, our method proficiently disentangles the content and style in separate branches, allowing free-form inputs to guide the structure and texture accordingly. This enables the alignment of text and image features in the discrete space, thereby generating stylized images closely aligned with the provided descriptions.

Table 1: Statistical and quantitative comparison of inference time, content and style loss value with state-of-the-art methods. The user study results show the average percentage where other methods perform better than ours in terms of overall quality, content preservation, and style criteria. The best and second-best results are indicated in bold and underlined, respectively.

Method		Ours	QuantArt [13]	CAST [45]	StyTr^2 [5]	StyleFormer [41]	IEST [2]	AdaAttN [28]	MCCNet [4]	ArtFlow [1]	AdaIN [14]
Inference time(ms/image)		24	32	<u>11</u>	530	14	370	71	13	65	**7**
Content loss($\mathcal{L}_c$)		**2.238**	2.332	2.440	2.297	2.808	2.402	2.645	2.334	<u>2.283</u>	2.524
style loss($\mathcal{L}_s$)		2.098	2.520	2.488	**1.425**	2.303	2.662	1.958	1.687	2.169	<u>1.560</u>
SSIM		<u>0.598</u>	0.547	**0.603**	0.538	0.558	0.529	0.465	0.348	0.308	0.319
User Study	Content	—	40.12%	31.88%	27.50%	44.38%	33.75%	34.38%	27.50%	51.25%	10.00%
	Style	—	21.86%	45.00%	43.75%	31.88%	17.50%	19.38%	40.00%	25.63%	18.13%
	Overall	—	33.78%	33.75%	30.00%	41.88%	23.75%	21.25%	30.00%	26.88%	12.50%

4.3 Quantitative Results

Image Style Transfer We present quantitative evaluation results in Table 1, including style and content differences between the generated results and input images, and a user survey assessing the quality of our approach. The first row shows the average inference time of several style transfer methods, and the second and third rows display content and style loss of different methods, respectively. The fourth row shows the SSIM loss of different methods. We calculate content differences according to Eq. 4 and style differences according to Eq. 5 for each method. We generate 900 stylized images by randomly selecting 30 style images and 30 content images. The discrete feature representation derived from vector quantization enables us to achieve optimal performance in measuring content loss, albeit at the expense of a slight loss in style information. However, this does not have any significant impact on visual quality or style similarity, as perceived by humans. In quantitative evaluation, we obtain relatively low scores, striking a suitable balance between content fidelity and style similarity, considering the trade-off between the two.

User study. We compare our method with nine state-of-the-art style transfer methods to evaluate which method is more favored by humans. We invite 36 users to take part in our survey, and for each participant, 50 questionnaires covering a pair of content and style images are randomly presented, which includes the results of our method and one of the other methods. Each user is asked to answer: (1) which result has better visual quality overall (2) which stylization result preserves the content structure better, and (3) which stylization result transfers style patterns more consistently. We show statistical results in Table 1 and our method outperforms other methods in overall quality and style consistency, achieving the second best result in content preservation.

Text-Guided Image Stylization

User study. We compare our method with the three state-of-the-art stylization methods which support text-guided stylization. The questionnaire contains 29 questions in total, and each question includes a result randomly from the four methods. The participants need to rate from the following three aspects: (1) the content fidelity degree of the

Table 2: The user study of text-guided image stylization in terms of overall quality, content preservation, and style consistency, rating from 1-5. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Ours	CLIPstyler [17]	VQGAN-CLIP [3]	LDAST [10]
Content	<u>4.301</u>	**4.354**	2.656	3.801
Style	**3.966**	<u>3.573</u>	3.531	2.800
Overall	**3.989**	<u>3.801</u>	2.908	3.033

stylized result and the content image, (2) the similarity between the stylized result and the style provided by text, and (3) the quality of the generation overall. Each problem will be rated from 1 to 5. We collect 24 valid questionnaires, and we present the results of the user survey in Table 2. The scores demonstrate that our method achieves better results in style and overall quality than other methods. Compared to the CLIPstyler [17], a slight drop occurs in content due to the lack of sufficient style strength. Generally, our method beats other methods in most comparisons.

4.4 Ablation Study

Fig. 6: Results of ablation study in step 1.

Step 1 To verify the effectiveness of our framework in step 1, we conduct ablation experiments by removing the adversarial loss and contrastive loss, respectively. As depicted in Figure 6, we train a network without a discriminator, resulting in obvious checkerboard effects and blurred images. It is difficult to learn the texture of the original style, so the generated results are far from the brushstrokes of the reference style image. In another experiment, we remove the entire contrastive loss, including content and style contrastive loss. It is apparent that the generated images in the first row suffer from poor content fidelity and exhibit gaps in color when compared with the style reference images. Without the contrastive constraints, it is difficult to learn the texture

Table 3: Quantitative results of ablation study in step 1.

$\mathcal{L}_{adv}$	$\mathcal{L}_{contra}$	content loss($\mathcal{L}_c$)↓	style loss($\mathcal{L}_s$)↓	SSIM↑
✗	✓	2.561	2.443	0.501
✓	✗	2.336	2.502	0.523
✓	✓	2.238	2.098	0.598

information of the style image. However, by incorporating the full set of losses, our method can better retain the content information of the original content image while faithfully transferring the texture and color of the style image. As shown in Table 3, without $\mathcal{L}_{adv}$ and $\mathcal{L}_{contra}$, the $\mathcal{L}_c$ increased by 0.323 and 0.225, while the $\mathcal{L}_s$ increased by 0.345 and 0.402. Additionally, the SSIM decreased by 0.097 and 0.075.

step 2 Moreover, to verify the effectiveness of our framework in step 2, we conduct an ablation study by removing the clip loss and token reconstruction loss, respectively. As depicted in Figure 7, it is apparent that certain words are missing in the stylization results when the clip loss is excluded. For instance, the term "river" in the first line and "autumn" in the second line are not accurately reproduced.

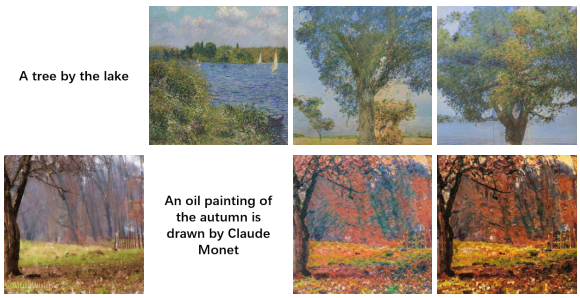

Fig. 7: Results of ablation study in step 2.

4.5 Applications

Audio-guided image stylization. In addition to image and text-guided stylization, our approach can also leverage the aligned image, text, and audio clip space [20] to encode audio. Figure 8 shows the synesthetic effect of audio as style guidance. For instance, the sound of fire leads to a predominantly red image, while the sound of wind produces a hazy effect. Our method can also reflect the mood of the music into the style of the image. Please refer to the supplementary material for more examples.

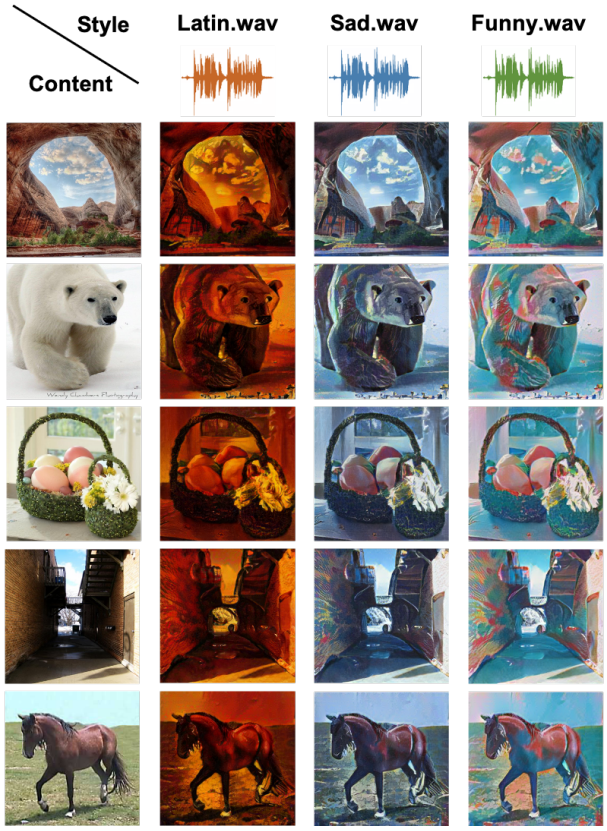

Fig. 8: Results of music-guided image stylization. The mood of different music show various visual effects.

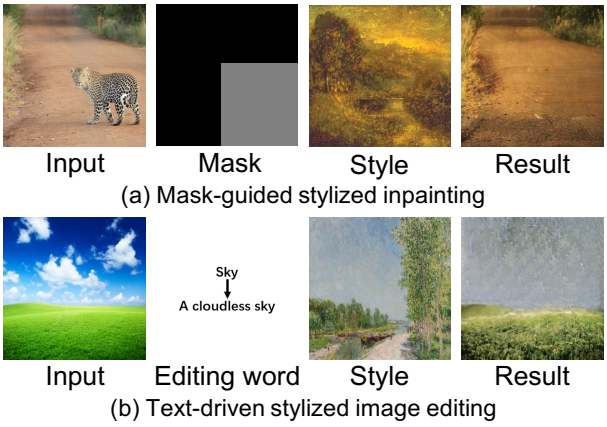

Fig. 9: Results of different novel applications.

Mask-guided stylized inpainting. As shown in Figure 9(a), we can redraw some uninteresting areas by removing objects of the input image by mask and achieving stylization with free-form guidance simultaneously.

Text-driven stylized image editing. As shown in Figure 9(b), we can achieve semantic editing by assigning new text caption to specific area and stylization with free-form guidance at the same time.

5 Limitations and future work

Although our method achieves high-quality results on stylization task for multi-modalities input as content or style, there are still several technical issues that need to be addressed. First of all, due to the limited number of tokens after quantization, it doesn't hold up very well for some painting styles or detailed strokes. However,we think it can be handled by extending the number of tokens. Secondly, our method will produce some failure cases for explicit text guided stylization generation due to some minor misalignment of content and style features in CLIP space. In the future, we will try to optimize our model to solve the problem by applying better encoder. At the same time, taking advantage of the quantization of features, there are many image editing operations that can be realized, e.g., inpainting or outpainting.

6 Conclusion

In this paper, we propose a unified image stylization framework which explicitly decouples the content and style branch, and supports various input forms, allowing image, text and audio as content and style information, respectively. The framework is built on a vector quantized style transfer approach, with a stylization transformer to fuse content and style information. Additionally, we incorporate a CLIP-based conditional transformer to align the discrete representation of image, text, and audio modality. Pesudo-paired token predictor is introduced and trained by image and pesudo caption pair generated by BLIP to estimate the representation from multi-modalities. We believe that our framework, which relies on discrete representations of input modalities, has the potential to be applied to other image processing tasks, and is worth exploring in future research.

References

1. An, J., Huang, S., Song, Y., Dou, D., Liu, W., Luo, J.: Artflow: Unbiased image style transfer via reversible neural flows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 862–871 (2021)
2. Chen, H., Wang, Z., Zhang, H., Zuo, Z., Li, A., Xing, W., Lu, D., et al.: Artistic style transfer with internal-external learning and contrastive learning. Advances in Neural Information Processing Systems **34**, 26561–26573 (2021)

3. Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., Raff, E.: VQGAN-CLIP: Open domain image generation and editing with natural language guidance. In: European Conference on Computer Vision (ECCV). pp. 88–105 (2022)
4. Deng, Y., Tang, F., Dong, W., Huang, H., Ma, C., Xu, C.: Arbitrary video style transfer via multi-channel correlation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1210–1217 (2021)
5. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: StyTr^2: Image style transfer with transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11326–11336 (2022)
6. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
7. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. Advances in Neural Information Processing Systems **34**, 19822–19835 (2021)
8. Ding, M., Zheng, W., Hong, W., Tang, J.: Cogview2: Faster and better text-to-image generation via hierarchical transformers. arXiv preprint arXiv:2204.14217 (2022)
9. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
10. Fu, T.J., Wang, X.E., Wang, W.Y.: Language-driven artistic style transfer. In: European Conference on Computer Vision. pp. 717–734. Springer (2022)
11. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
12. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706 (2022)
13. Huang, S., An, J., Wei, D., Luo, J., Pfister, H.: Quantart: Quantizing image style transfer towards high visual fidelity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5947–5956 (2023)
14. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)
15. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Kwon, G., Ye, J.C.: CLIPstyler: Image style transfer with a single text condition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18062–18071 (2022)
18. Kwon, G., Ye, J.C.: Diffusion-based image translation using disentangled style and content representation. arXiv preprint arXiv:2209.15264 (2022)
19. Lee, S.H., Kim, C., Byeon, W., Yoon, S.H., Kim, J., Kim, S.: Lisa: Localized image stylization with audio via implicit neural representation. arXiv preprint arXiv:2211.11381 (2022)
20. Lee, S.H., Roh, W., Byeon, W., Yoon, S.H., Kim, C., Kim, J., Kim, S.: Sound-guided semantic image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3377–3386 (2022)
21. Li, C., Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks. In: European conference on computer vision. pp. 702–716. Springer (2016)
22. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. Advances in neural information processing systems **30** (2017)

23. Li, Y., Liu, M.Y., Li, X., Yang, M.H., Kautz, J.: A closed-form solution to photorealistic image stylization. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 453–468 (2018)
24. Lin, J., Pang, Y., Xia, Y., Chen, Z., Luo, J.: Tuigan: Learning versatile image-to-image translation with two unpaired images. In: European Conference on Computer Vision. pp. 18–35. Springer (2020)
25. Lin, M., Tang, F., Dong, W., Li, X., Xu, C., Ma, C.: Distribution aligned multimodal and multi-domain image stylization. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **17**(3), 1–17 (2021)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
27. Liu, A.H., Liu, Y.C., Yeh, Y.Y., Wang, Y.C.F.: A unified feature disentangler for multi-domain image translation and manipulation. Advances in neural information processing systems **31** (2018)
28. Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., Ding, E.: Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6649–6658 (2021)
29. Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5880–5888 (2019)
30. Phillips, F., Mackintosh, B.: Wiki art gallery, inc.: A case for critical thinking. Issues in Accounting Education **26**(3), 593–608 (2011)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
32. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
33. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)
34. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
36. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487 (2022)
37. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6924–6932 (2017)
38. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
39. Wang, Z., Liu, W., He, Q., Wu, X., Yi, Z.: Clip-gen: Language-free training of a text-to-image generator with clip. arXiv preprint arXiv:2203.00386 (2022)
40. Wu, H.H., Seetharaman, P., Kumar, K., Bello, J.P.: Wav2clip: Learning robust audio representations from clip. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4563–4567. IEEE (2022)

41. Wu, X., Hu, Z., Sheng, L., Xu, D.: Styleformer: Real-time arbitrary style transfer via parametric style composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14618–14627 (2021)
42. Yoo, J., Uh, Y., Chun, S., Kang, B., Ha, J.W.: Photorealistic style transfer via wavelet transforms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9036–9045 (2019)
43. Zhang, C., Yang, J., Wang, L., Dai, Z.: S2wat: Image style transfer via hierarchical vision transformer using strips window attention. arXiv preprint arXiv:2210.12381 (2022)
44. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International conference on machine learning. pp. 7354–7363. PMLR (2019)
45. Zhang, Y., Tang, F., Dong, W., Huang, H., Ma, C., Lee, T.Y., Xu, C.: Domain enhanced arbitrary image style transfer via contrastive learning. arXiv preprint arXiv:2205.09542 (2022)
46. Zhao, Y., Wu, R., Dong, H.: Unpaired image-to-image translation using adversarial consistency loss. In: European Conference on Computer Vision. pp. 800–815. Springer (2020)