Inversion-Based Creativity Transfer with Diffusion Models

Yuxin Zhang\textsuperscript{1,2}, Nisha Huang\textsuperscript{1,2}, Fan Tang\textsuperscript{3}, Haibin Huang\textsuperscript{4}

Chongyang Ma\textsuperscript{4}, Weiming Dong\textsuperscript{1,2}, Changsheng Xu\textsuperscript{1,2}

\textsuperscript{1}NLPR, Institute of Automation, Chinese Academy of Sciences  \textsuperscript{2}School of Artificial Intelligence, UCAS  \textsuperscript{3}Institute of Computing Technology, Chinese Academy of Sciences  \textsuperscript{4}Kuaishou Technology

Figure 1. Creativity transfer results using our method. Given only a single input painting image, our method can accurately transfer the creative attributes such as semantic elements, material, object shape, brushstrokes and colors of the references to a natural image with a very simple learned textual description.

Abstract

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

1. Introduction

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

Generating an artistic image from one or multiple examples has attracted many interests in recent years. A typical task is style transfer [1, 16, 30, 48, 52], which can create a new artistic image from an input natural image and a painting image, by combining the content of the natural image and the style of the painting image. However, particular creative attributes such as object shape and semantic elements are difficult to be transferred (see Figures 2(b) and 2(e)).
Text guided stylization [13, 15, 26, 34] produces an artistic image from a natural image and a text prompt, but usually the text prompt for target style can only be a rough description of material (e.g., “oil”, “watercolor”, “sketch”), art movement (e.g. “Impressionism”, “Modernism”, see Figure 2(a)), artist (e.g., “Vincent van Gogh”, “Tony Toscani”, see Figure 2(d)) or a famous artwork (e.g., “Starry Night”, “The Scream”). Diffusion-based methods [8, 22, 24, 32, 47] generate high-quality and diverse artistic images based on a text prompt, with or without image examples. In addition to the input image, a detailed auxiliary textual input is required to guide the generation process if we want to reproduce some vivid contents and styles, which may be still difficult to reproduce the creative idea of a specific painting in the result.

In this paper, we propose creativity transfer, a novel example-guided artistic image generation task related to style transfer and text-to-image synthesis, to alleviate all the above problems. Given only a single input painting image, our method can learn and transfer its creativity to a natural image with a very simple text prompt (see Figures 1 and 2(f)). The resulting image exhibit very similar creative attributes of the original painting, including material, brushstrokes, colors, object shape and semantic elements, without losing diversity. Furthermore, we can also control the content of the resulting image by giving a text description (see Figure 2(c)).

To achieve this merit, we need to obtain the representation of image creativity, which refers to the set of attributes that appear in the high-level textual description of the image. We define the textual descriptions as “new words” that do not exist in the normal language and get the embeddings via inversion method. We benefit from the recent success of diffusion models [38, 45] and inversion [2, 14]. We adapt diffusion models in our work as a backbone to be inverted and as a generator in image-to-image and text-to-image synthesis. Specifically, we propose an efficient and accurate textual inversion based on the attention mechanism, which can quickly learn key features from an image. We use CLIP [35] image embedding to obtain high-quality initial points, and learn key information in the image through multi-layer cross-attention. Taking a painting image as a reference, the attention-based inversion module is fed with its CLIP image embedding and then gives its textual embedding. The diffusion models conditioning on the textual embedding can produce new images with the learned creativity of the reference.

To demonstrate the effectiveness of our method, we conduct comprehensive experiments, applying our method to numerous images of various artists and styles. All of the experiments show that our method produces impressive results, generating artistic images that both imitate well the creative attributes to a high degree, and achieve content consistent with the input natural images or textual descriptions. We demonstrate much improved visual quality and artistic consistency as compared to state-of-the-art approaches. These outcomes demonstrate the generality, precision and adaptability of our method.

We summarize our main contributions as follows:

- We introduce the task of artistic creativity transfer. Given a single painting image, the goal is to generate new artistic images with high fidelity to its creative attributes, by using natural images or text descriptions to control the contents.
- We propose an attention-based single image textual inversion method, which can quickly and accurately learn the overall semantics and artistic technique of an image, so as to capture the complete creativity of the painting.
- We perform comparative experiments to show that our creative learning approach can achieve state-of-the-art performance and novel visual effects.

2. Related Work

Image style transfer Image style transfer has been widely studied as a typical mechanism of example guided
artistic image generation. Traditional style transfer methods use low-level hand-crafted features to match the patches between content image and style image [46, 50]. In recent years pre-trained deep convolutional neural networks are used to extract the statistical distribution of features which can capture style patterns effectively [17, 18, 25]. Arbitrary style transfer methods use unified models to handle arbitrary inputs by building feed-forward architectures [7, 23, 27–29, 33, 44, 48, 51]. Liu et al. [30] learn spatial attention score from both shallow and deep features by an adaptive attention normalization module (AdaAttN). An et al. [1] alleviate content leak by reversible neural flows and an unbiased feature transfer module (ArtFlow). Chen et al. [3] apply internal-external scheme to learn feature statistics (mean and standard deviation) as style priors (IEST). Zhang et al. [52] learn style representation directly from image features via contrastive learning to achieve domain enhanced arbitrary style transfer (CAST). Besides CNN, visual transformer has also been used for style transfer tasks. Wu et al. [48] perform content-guided global style composition by a transformer-driven style composition module (StyleFormer). Deng et al. [6] propose a transformer-based method (StyTr²) to avoid the biased content representation in style transfer by taking long-range dependencies of input images into account. Image style transfer methods mainly focus on learning and transferring colors and brushstrokes, but have difficulties in imitating other artistic creative attributes such as object shape and semantic elements.

**Text-to-image synthesis** Text guided synthesis methods can also be used to generate artistic images [10, 36, 37, 40, 49]. CLIPDraw [12] synthesizes artistic images from text by using CLIP encoder [35] to maximize similarity between the textual description and generated drawing. VQGAN-CLIP [5] uses CLIP-guided VQGAN [11] to generate artistic images of various styles from text prompts. Rombach et al. [38] train diffusion models [21, 42] in the latent space to reduce complexity and generate high quality artistic images from texts. Those models only use text guidance to generate an image, without fine-grained content or style control. Some methods add image prompt to increase controllability to the content of the generate image. CLIPstyler [26] transfers an input image to a desired style with a text description by using CLIP loss and PatchCLIP loss. StyleGAN-NADA [15] use CLIP to adaptively train the generator, which can transfer a photo to artistic domain by text description of the target style. Huang et al. [22] propose a diffusion-based artistic image generation approach by utilizing multimodal prompts as guidance to control the classifier free diffusion model. Hertz et al. [19] change an image to artistic style by using the text prompt with style description and injecting the source attention maps. Those methods are still difficult to generate images with complex or special artistic characteristics which cannot be described by normal texts. StyleCLIPDraw [41] jointly optimizes textual description and style image for artistic image generation. Liu et al. [31] extract style description from CLIP model by a contrastive training strategy, which enables the network to perform style transfer between content image and a textual style description. These methods utilize the aligned image and text embedding of CLIP to achieve style transfer via narrowing the distance between the generated image and the style image, while we obtain the image representation straight from the artistic image.

**Inversion of diffusion models** Inversion of diffusion models is to find a noise map and a conditioning vector corresponding to a generated image. It is a potential way for improving the quality of example-guided artistic image generation. However, naively adding noise to an image and then denoising it may yield an image with significantly different content. Choi et al. [4] perform inversion by using noised low-pass filter data from the target image as the basis for the denoising process. Dhariwal et al. [9] invert the deterministic DDIM [43] sampling process in closed form to obtain a latent noise map that will produce a given real image. Ramesh [36] develop a text-conditional image generator based on the diffusion models and the inverted CLIP. The above methods are difficult to generate new instances of a given example while maintaining fidelity.

Gal et al. [14] presents a textual inversion method to find a new pseudo-word to describe visual concept of a specific object or artistic style in the embedding space of a fixed text-to-image model. They use optimization-based methods to directly optimize the embedding of the concept, which is inefficient. Ruiz et al. [39] implant a subject into the output domain of the text-to-image diffusion model so that it can be synthesized in novel views with a unique identifier. Their inversion method is based on the fine-tuning of diffusion models, which demands high computational resources. Both methods learn concepts from pictures through textual inversion, while they need multiple (3-5) images to depict the concept. Moreover, the concept they aim to learn is usually an object. Our method can learn the corresponding textual embedding from a single image and use it as a condition to guide the generation of artistic images without fine-tuning the generative model.

### 3. Method

#### 3.1. Preliminaries

Diffusion probabilistic models (DPMs) [20] are probabilistic models designed to learn a data distribution by gradually denoising a normally distributed variable, which corresponds to learning the reverse process of a fixed Markov Chain of length $T$. For image synthesis, models can be in-
interpreted as an equally weighted sequence of denoising autoencoders \( \epsilon_\theta(x_t, t); t = 1, ..., T \), which are trained to predict a denoised variant of their input \( x_t \), where \( x_t \) is a noisy version of the input \( x \). For Stochastic DPMs, given the mean estimator \( \mu_T \) and \( x_T \sim \mathcal{N}(0, I) \), the image \( x := x_0 \) is generated through \( x_{t-1} \sim \mathcal{N}(\mu_T(x_t, t), \text{diag}(\sigma^2_t)) \). Deterministic DPMs generate images with the ODE formulation. Given the mean estimator \( \mu_T \), deterministic DPMs generate \( x := x_0 \) via \( z := x_t \sim \mathcal{N}(\mu_T(x_t, t), t = T, ..., 1) \). Latent diffusion models (LDMs) \([38]\) first utilize an autoencoding model which learns a space that is perceptually equivalent to the image space, to compute a latent code. An encoder \( E \) maps an image \( x \) into a spatial latent code \( z = E(x) \). A decoder \( D \) learns to map such latent codes back to images, such that \( D(E(x)) \approx x \). A diffusion model is trained in the learned latent space to generate target latent codes. The conditioned LDM loss is formulated as:

\[
L_{LDM} := \mathbb{E}_{E(x), t \sim \mathcal{N}(0,1), t} \left[ \| e - \epsilon_\theta(z_t, t) \|_2^2 \right] . \tag{1}
\]

The underlying UNet backbones of LDMs are augmented with the cross-attention mechanism, which is effective for learning attention-based models of various input modalities (e.g., class labels, segmentation masks, and the jointly trained text-embedding). A model that converts a conditioning input, \( y \), into a conditioning vector is called \( \tau_\theta(y) \). The conditioned LDM loss is then determined as:

\[
L_{LDM} := \mathbb{E}_{E(x), \tau, e \sim \mathcal{N}(0,1), t} \left[ \| e - \epsilon_\theta(z_t, t, \tau(y)) \|_2^2 \right] . \tag{2}
\]

We use Stable Diffusion Models (SDMs) \([45]\) as the backbone. Following the structure of LDMs, SDMs get into the latent space via an autoencoder and perform sampling on a low-dimensional space.

### 3.2. Problem Formation of Creativity Transfer

We believe that art speaks for itself and artistic creativity within every painting deserves to be appreciated and learned. Creativity transfer is a novel task that aims to discover the hidden stories behind each artwork. Artistic creativity may not be able to be described by existing words, but we can learn novel “words” via the inversion approaches. We define the creativity in our task as the attributes in the corresponding high-level textual descriptions of artistic images, including semantics (subjects, scene and decorative elements, etc.) and artistic techniques (i.e., material, colors, brushstrokes and shapes, etc.). The formal definition of the creativity transfer task is formulated as a semantic representation of creativity, which is directly learned from painting images through inversion or other representation learning methods. The key elements in creativity transfer are: (i) a guided artistic image \( y \) containing the creativity to be learned; (ii) a high-level textual representation \( v \) containing the creativity of the guided artistic image; (iii) an algorithm or network \( F \) that learns high-level textual information from images; (iv) a text-to-image generative method or model \( G \) to be inverted; (v) synthesis tasks for demonstrating the effectiveness of the method. The creativity transfer process can be expressed as: \( v = F(y, G) \). We define the metrics for evaluating this representation as accuracy and editability. Accuracy is reflected in the image-to-image transfer task, whether it is possible to obtain results consistent with the visual characteristics of the reference image when the content layout exists. The task can be expressed as \( I = G_{\text{style}}(v, x) \), where \( x \) denotes the input content image. Editability is reflected in the example-guided text-to-image synthesis task, whether new artistic images with different semantics from the guided image can be synthesized. The task can be expressed as \( I = G_{\text{txt2img}}(v, \text{noise}) \), where \( I \) denotes the synthesized image and \( \text{noise} \) denotes random noise.

In this work, we use inversion as the basis of our creativity transfer framework and SDMs as the generative backbone. Note that our framework is not restricted to a specific generative model. As shown in Figure 3, our method involves pixel space, latent space, and textual space. During training, image \( x \) is equal to image \( y \). The image embedding of image \( x \) is obtained by the CLIP image encoder and then sent to the attention-based inversion module.
multi-layer cross-attention, the key information of the image embedding is learned. The inversion module gives text embedding \( v \), which is converted into the standard format of caption conditioning SDMs. Conditioned on the input textual information, the generative model obtains a series of latent codes \( z_t \) through sequence denoise process from to the random noise \( z_T \) and finally gives the latent code \( z \) corresponding to the art image. The inversion module is optimized by the simple loss of LDMs computed on the “latent noise” of forward process and the reverse process (See Sec. 3.3). In the inference process, \( x \) is the content image, and \( y \) is the reference image. The textual embedding \( v \) of the reference image \( y \) guides the generative model to generate a new artistic image according to the content image \( x \).

### 3.3. Textual Inversion

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

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

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

Thanks to CLIP’s aligned latent space of image embedding and text embedding, it provides powerful guidance for our optimization process. We propose a learning method based on multi-layer cross attention. The input artistic image is first sent into the CLIP image encoder and gives image embeddings. By performing multi-layer attention on these image embeddings, the key information of the image can be quickly obtained. The CLIP image encoder \( \tau_{\theta} \) projects \( y \) to an image embedding \( \tau_{\theta}(y) \). The multi-layer cross attention starts with \( v_0 = \tau_{\theta}(y) \). Then each layer is implementing \( \text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^{T}}{\sqrt{d}} \right) \cdot V \) with:

\[
Q_i = W_Q^{(i)} \cdot v_i, \quad K = W_K^{(i)} \cdot \tau_{\theta}(y), \quad V = W_V^{(i)} \cdot \tau_{\theta}(y),
\]

\[
v_{i+1} = \text{Attention}(Q_i, K, V).
\]

To avoid overfitting, we apply a dropout strategy in each cross-attention layer which set to 0.05.

Our optimization goal can finally be defined as:

\[
\hat{v} = \text{arg min}_{v} \mathbb{E}_{z,x,y,t} \left[ \| \epsilon - \epsilon_{\theta}(z_t, t, \text{MultiAttn}(\tau_{\theta}(y))) \|^2_2 \right],
\]

where \( z \sim E(x), \epsilon \sim \mathcal{N}(0, 1) \). \( \tau_{\theta} \) and \( \epsilon_{\theta} \) are fixed during training. In this way, \( \hat{v} \) can be optimized to the target area efficiently.

### 3.4. Stochastic Inversion

We observe that in addition to the text description, the random noise controlled by the random seed is also important for the representation of the image. [19] demonstrates that the changes of random seed leads to obvious changes of visual differences, while a fixed random seed and cross attention map results in minimal changes. We define pre-trained text-to-image diffusion model-based image representation into two parts: holistic representation and detail representation. The holistic representation refers to the text conditions, and the detail representation is controlled by the random noise. To get a fine and accurate representation of the painting image, we need to reduce the effects of randomness. In other words, we aim to obtain the certain detailed representation of each artistic image, that is, the corresponding “random” noise. We define the process from an image to noise maps as an inversion problem, and design an inversion method, which we called stochastic inversion. Specifically, for each image \( z \), the stochastic inversion module takes the image latent code \( z = E(x) \) as input. Set \( z_t \), the noisy version of \( z \), as computable parameters, then \( \epsilon_t \) is obtained by:

\[
\hat{\epsilon}_t = (z_{t-1} - \mu_T(z_t, t)) \sigma_t.
\]

### 4. Experiments

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

**Implementation details** We retain the original hyper-parameter choices of SDMs. The training process takes about 20 minutes each image on one NVIDIA GeForce RTX3090 with a batch size of 1. The base learning rate was set to 0.001. The base learning rate is scaled by the number of GPUs and the batch size, for an effective rate of 0.04. The synthesis process takes the same time as SDM, which depends on the steps.
4.1. Comparison with Style Transfer Methods

As shown in Figure 4, we compare our method with the state-of-the-arts image style transfer methods, including ArtFlow [1], AdaAttN [30], StyleFormer [48], IEST [3], StyTr$^2$ [6] and CAST [52] to show the effectiveness of creativity transfer. From the results, we can see apparent advantages of our creativity transfer method on transferring the semantics and artistic techniques of the reference images to the content images over traditional style transfer methods. For example, our method can better transfer the shapes of important objects, such as the facial forms and eyes (the 1st to 5th rows), the mountain (the 7th row) and the sun (the 8th row). Our method can capture some special semantics of the reference images and reproduce the visual effects in the results, such as the stars on the background (the 1st row), the flower headwear (the 3rd row) and
we show the optimization process of [14] and ours. Our method can quickly optimize to the target text embedding in about 1000 iterations, while [14] usually takes 10 times the iterations of ours due to its simple optimization-based scheme. In Figure 6, we demonstrate our superior generality and editability by giving additional semantic descriptions that do not appear in the reference image. It can be easily observed that our method is more robust to those additional descriptions and is able to generate results that match both the textual description and the reference image. However, [14] loses adaptability to these texts and is not able to depict the specific artistic visual effect.

**Comparison with SDMs** We compare with the state-of-the-art text-to-image generative model SDMs [45]. SDMs can generate high-quality images from text descriptions. However, it is difficult to describe the style of a specific painting with only text as a condition, so satisfactory results cannot be obtained. As shown in Figure 7, our method better captures the unique creative attributes of the reference images.

### 4.3. Discussions and Limitations

Although our method can transfer typical colors, when there is a significant difference between the colors of the content image and the reference image, our method may fail to transfer the color in a one-to-one correspondence semantically. For example, the green hair of the content images in the 1st row of Figure 4 is not transferred into brown. As shown in Figure 9, we employ an additional tone transfer module [23] to align the color of content and reference images. However, we observe that different users have different preferences on whether the colors of the content image should be retained. We believe that the colors of a photograph is crucial, so we defer to the tone of the original picture.
5. Conclusion

We introduce a novel example-guided artistic image generation task called creativity transfer, which refers to learning the high-level textual descriptions of a single painting image and then guides the text-to-image generative model to create images of specific artistic appearance. We propose an attention-based textual inversion method to invert a painting into the corresponding textual embeddings, which benefits from the aligned text and image feature spaces of CLIP. Extensive experimental results demonstrate that our method achieves superior image-to-image and text-to-image generation results compared with state-of-the-art approaches. Our approach is intended to pave the way for upcoming unique artistic image synthesis tasks.
References

[1] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. ArtFlow: Unbiased image style transfer via reversible neural flows. In IEEE/CVF Conferences on Computer Vision and Pattern Recognition (CVPR), pages 862–871, 2021. 1, 3, 6

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

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

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

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

[6] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. StyTr²: Image style transfer with transformers. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11326–11336, 2022. 2, 3, 6

[7] Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. Arbitrary style transfer via multi-adaptation network. In ACM International Conference on Multimedia, pages 2719–2727, 2020. 3

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

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

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

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

[12] Kevin Frans, Lisa B Soros, and Olaf Witkowski. CLIPDraw: Exploring text-to-drawing synthesis through language-image encoders. arXiv preprint arXiv:2106.14843, 2021. 3

[13] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven artistic style transfer. In European Conference on Computer Vision (ECCV), pages 717–734, 2022. 2

[14] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618, 2022. 2, 3, 7

[15] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. ACM Transactions on Graphics, 41(4):141:1–141:13, 2022. 2, 3

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

[17] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2414–2423, 2016. 3

[18] Leon A. Gatys, Alexander S. Ecker, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3730–3738, 2017. 3

[19] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626, 2022. 3, 5

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

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

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

[23] Xin Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In IEEE International Conference on Computer Vision (ICCV), pages 1501–1510, 2017. 3, 7

[24] Bahjat Kawar, Shiran Zada, Oran Lang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Style-based real image editing with diffusion models. arXiv preprint arXiv:2210.09276, 2022. 2

[25] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10043–10052, 2019. 3

[26] Gihyun Kwon and Jong Chul Ye. CLIPstyler: Image style transfer with a single text condition. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 18062–18071, 2022. 2, 3

[27] Xueting Li, Sifei Liu, Jian Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3804–3812, 2019. 3
diffusion models with deep language understanding. \textit{arXiv preprint arXiv:2205.11487}, 2022. 3

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

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

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In \textit{International Conference on Learning Representations (ICLR)}, 2021. 3

[44] Jan Svozboda, Asha Anoosheh, Christian Osendorfer, and Jonathan Masci. Two-stage peer-regularized feature recombination for arbitrary image style transfer. In \textit{IEEE/CVF Conference on Computer Vision and Pattern Recognition}, pages 13816–13825, 2020. 3

[45] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. \url{https://github.com/huggingface/diffusers}, 2022. 2, 4, 7

[46] Wei Zhang, Wenping Wang, Huaiyang Wang, and Jianguang Sun. Efficient example-based painting and synthesis of 2D directional texture. \textit{IEEE Transactions on Visualization and Computer Graphics}, 10(3):266–277, 2004. 3

[47] Xianchao Wu. Creative painting with latent diffusion models. \textit{arXiv preprint arXiv:2209.14697}, 2022. 3

[48] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. StyleFormer: Real-time arbitrary style transfer via parametric style composition. In \textit{IEEE/CVF International Conference on Computer Vision (ICCV)}, pages 14598–14607, 2021. 1, 3, 6

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

[50] Wei Zhang, Chen Cao, Shifeng Chen, Jianzhuang Liu, and Xiaoou Tang. Style transfer via image component analysis. \textit{IEEE Transactions on Multimedia}, 10(3):266–277, 2004. 3

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

[52] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. \textit{arXiv preprint arXiv:2208.12242}, 2022. 3

[53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. \textit{arXiv preprint arXiv:2205.11487}, 2022. 3

[54] [55] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. Style-CLIP: Text-driven manipulation of stylegan imagery. In \textit{IEEE/CVF International Conference on Computer Vision (ICCV)}, pages 2085–2094, 2021. 2

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In \textit{International Conference on Machine Learning (ICML)}, pages 8748–8763, 2021. 2, 3, 4

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

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

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

[60] [57] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. Style-CLIP: Text-driven manipulation of stylegan imagery. In \textit{IEEE/CVF International Conference on Computer Vision (ICCV)}, pages 2085–2094, 2021. 2