Bridging Text and Image for Artist Style Transfer via Contrastive Learning

Zhi-Song Liu, Li-Wen Wang, Jun Xiao, Vicky Kalogeiton

Bridging Text and Image for Artist Style Transfer via Contrastive Learning

Zhi-Song Liu[1], Li-Wen Wang[2], Jun Xiao[2], and Vicky Kalogeiton[3]

[1] Lappeenranta-Lahti University of Technology LUT, Finland
zhisong.liu@lut.fi
[2] The Hong Kong Polytechnic University, Hong Kong
[3] LIX, Ecole Polytechnique, IP Paris

Abstract. Image style transfer has attracted widespread attention in the past few years. Despite its remarkable results, it requires additional style images available as references, making it less flexible and inconvenient. Using text is the most natural way to describe the style. More importantly, text can describe implicit abstract styles, like styles of specific artists or art movements. In this paper, we propose a Contrastive Learning for Artistic Style Transfer (CLAST) that leverages advanced image-text encoders to control arbitrary style transfer. We introduce a supervised contrastive training strategy to effectively extract style descriptions from the image-text model (i.e., CLIP), which aligns stylization with the text description. To this end, we also propose a novel and efficient adaLN based state space models that explore style-content fusion. Finally, we achieve a text-driven image style transfer. Extensive experiments demonstrate that our approach outperforms the state-of-the-art methods in artistic style transfer. More importantly, it does not require online fine-tuning and can render a 512×512 image in 0.03s.

Keywords: Style transfer · multimodal learning · vision and language · text guidance · domain transfer · contrastive learning

1 Introduction

Image style transfer is a popular topic that aims to apply the desired painting style to an input content image. The transfer model requires the information of *"what content"* in the input image and *"which painting style"* to be used [29,50]. Conventional style transfer methods require a content image accompanied by a style image to provide the content and style information [4,11,22,23,34,41,51,67]. However, people have specific aesthetic needs. Usually, finding a single style image that perfectly matches one's requirements is inconvenient or infeasible. Text or language is a natural interface to describe the preferred style. Instead of using a style image, using text to describe style preference is easier to obtain and more adjustable. Furthermore, achieving perceptually pleasing artist-aware stylization typically requires learning from collections of art, as one reference image is not representative enough.

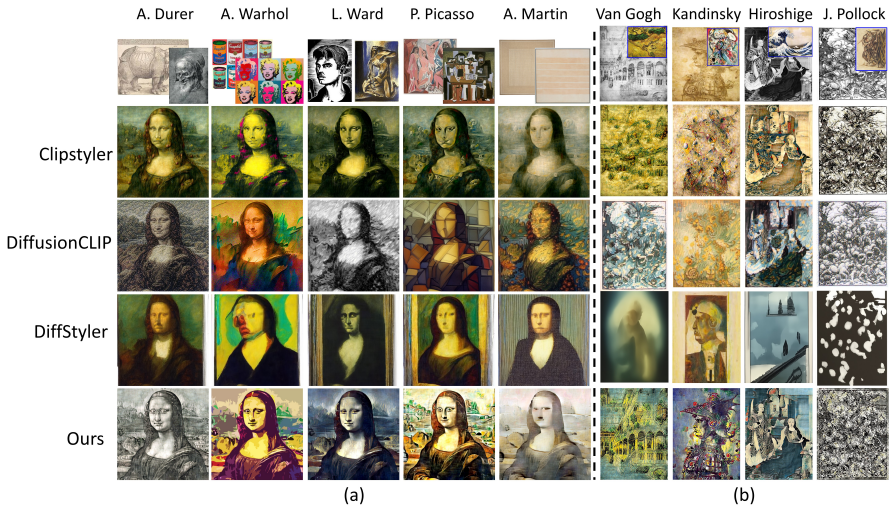

Fig. 1: Artist-aware style transfer. For (a), given five artists (top row) and the text prompt *"a Mona Lisa painting in Pablo Picasso style"*, we replace the artist's name in the prompt with each of the five and get stylization results. We show results for Clipstyler [29], DiffusionCLIP [25], DiffStyler [21], and our proposed CLAST . Our method mimics the most representative styles from each artist, while Clipstyler [29] shows little style changes, DiffStyler changes facial features, and DiffusionCLIP does not learn representative styles. For (b), we show the stylization of four sketch paintings (top row). We can see that ours can transfer the styles without affecting the contents.

In this work, we learn arbitrary artist-aware image style transfer, which transfers the painting styles of any artist to the target image using just texts. Most studies on universal style transfer [31, 41, 50] limit their applications to using reference images as style indicators that are less creative or flexible. Text-driven style transfer has been studied [16, 29] and has shown promising results using a simple text prompt. However, these approaches require either costly data collection and labeling or online optimization for every content and style (e.g. DiffusionCLIP [25] and Clipstyler [29] in Figure 1). Instead, our proposed Text-driven artistic aware Style Transfer model CLAST overcomes these two problems and achieves better and more efficient stylization. Figure 1 illustrates this. It shows the artist-aware stylization on *Mona Lisa* by comparing text-driven methods (Clipstyler [29], DiffusionCLIP [25], DiffStyler [21] and Ours) on five artists. We input text prompts such as *"a Mona Lisa painting in Pablo Picasso style"* and change the artist's name accordingly. CLAST faithfully mimics the painters' style and preserves the contents reliably. Figure 1 (b) shows four examples of transferring four different sketches to the target styles, we can see that ours can better preserve the contours and lines of the original sketches, while others fail to achieve this.

To obtain artist awareness, CLAST explicitly explores the latent space: it maximizes the global distance amongst different artworks of different artists, while it minimizes the distance amongst artworks of the same artists. Specifically, artists have their style visualized in many paintings. The name of the artist is the only "label" that can connect artists to their painting style. CLAST not only is driven by different artists' names, but it also learns to group the painters' works to extract the most representative features. Specifically, given artists' names, CLAST projects features from different artists onto the CLIP space for classification. Moreover, we explore text-image correlations by empirically analyzing the co-linearity between artists[4] and paintings in the CLIP space demonstrate the effectiveness of text-driven style transfer. To achieve real-time inference, we adopt the adaLN based State Space Model (adaLN-SSM) to realize the style fusion as a linear sequential regression. It can greatly reduce the computation time and improve the style transformation. Our contributions can be summarized as follows: (1) To achieve text-driven image style transfer, we propose to embed the task-agnostic CLIP image-text model into our proposed CLAST . This enables CLAST to obtain style preference from *text descriptions*, making the image style transfer more interactive. (2) We propose a adaLN-based state space model (adaLN-SSM) to explore style-content fusion, which can efficiently model both local and global feature correlations. The stylized image not only can visualize the statistical similarity to the text description, but also preserve the original contents. (3) We suggest using a supervised contrastive training strategy (Sections 3.2) to learn art collection awareness. It can align corresponding artistic texts and images offline so the model can apply stylization in real time. (4) We conduct extensive experiments on text-driven style transfer. We show that CLAST outperforms the state-of-the-art methods both quantitatively and qualitatively.

2 Related Work

Arbitrary style transfer. It can be split into two groups: (1) style-aware optimization [18, 27, 32, 56] and (2) universal style transfer [23, 31, 51, 54]. Inspired by the attention mechanism [57], a few works use it to explore statistical correlations. SANet [41] matches the content and style statistics via cross-attention. AdaAttN [34] further explores the second-order attention to preserve more content information without losing style patterns. Artflow [4] and VAEST [36] investigate the normalization flow [30] and VAE [26] to fuse style and content images. The latter focuses on zero- or first-order statistics for real-time style transfer. AdaIN [23] proposes Adaptive instance normalization to shift the deep content

[4] We use artists' names rather than art movements (e.g. impressionism or cubism) as each artist has a relatively consistent painting style while one art movement can include different painters. For instance, Paul Cezanne and Claude Monet are both impressionists. Cezanne's paintings are more colorful and imaginative while Monet's paintings are clear and natural [1].

feature to the style space. ReReVST [59] follows Avatar-net [54] to further develop inter-channel feature adjustment for both image and video style transfer. ArtNet [51] proposes Zero-channel Pruning to reduce model complexity. More recently, there are some developments in transformer-based style transfer [11, 58] that use the vision transformer [12] for stylization. [68] modify the attention modules to pay attention to the fine-grained styles. Most recently, [22] proposes vector quantization to achieve latent space feature fusion for stylization. [60, 66] propose to utilize the power of a pre-trained diffusion model for image-based style transfer.

Artistic style transfer. Arbitrary-style transfer methods suffer from the fact that they learn style statistics from one reference image, which is not representative enough of the desired pattern. Artistic style transfer learns robust statistics from many art works so that it can transfer the most distinctive styles to the content images. A straightforward approach is to collect a number of desirable images to train a specific model for style transfer [13]. For example, AST [50] proposes an artist-aware style transfer to achieve art stylization. In other words, they collect specific artworks as style references, e.g., Van Gogh, to train a network for specific style transfer. DualAST [8] follows this idea and proposes a more flexible network to balance both artwork style and artist style via Style-Control Block. CAST [67] learns the style similarities and differences between multiple styles directly from image features. MAST [61] proposes a per-pixel process via style kernel, which can dynamically adapt to the contents for artistic style transfer.

Text-driven image style transfer. Style transfer is a subjective topic, that is, different people may have different preferences for stylization. Using style images as references may not be as sufficiently good as texts can describe styles in a more abstract and aesthetic manner. The success of CLIP [46], VQVAE [40] and multimodality [5,6,28] show that text and image can be related via a shared projection space. Some pioneer works on image editing [15,17,35,43,47,49,52,53] and video understanding [14, 38, 39] show that language can be used for user-guided applications. [53] constructs open-ended vocabularies to flexibly recompose the visual content in the latent spaces of GANs. Most recently, [16] uses the CLIP as the condition for style transfer that increases the cross-correlation between the output and text description for text-guided style transfer. Clipstyler [29] further developed this idea by using both global and patch CLIP losses to generate high-resolution stylized images. On a different trend, there are several approaches [3,7,48,65] that use text and/or image prompts for image synthesis and creative painting. For instance, [10,20,42,65] use text prompts to create a novel painting by semantic editing, like face and fashion design. Their main limitation is that they synthesize the new image much closer to the text description and the contents get distorted and do not match the original content images. Most recently, diffusion models have gained great attention for text-driven style transfer. DiffusionCLIP [25] and InST [66] fine-tune the pre-trained diffusion model for image manipulation, and Zecon [63] uses zero-shot learning to transfer patch-based patterns to the content images. StylerDALLE [62] proposes to combine

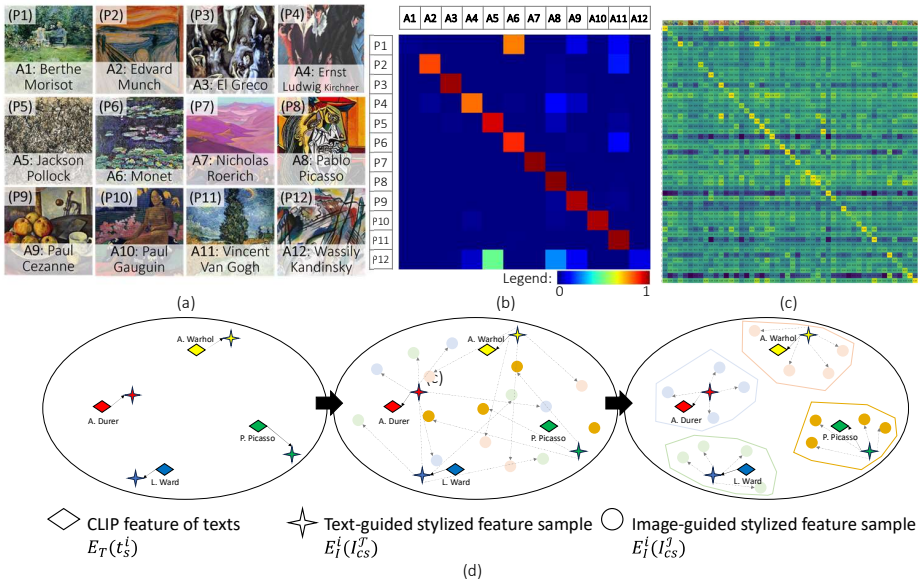

Fig. 2: Correlation between artists and paintings.(a) A set of Artist-Painting paired samples from WikiArt dataset [2]. The artists (Abbr.: **A**) are represented using text, and their paintings (Abbr.: **P**) are in the format of color images. Different artists and their paints are given different index numbers for clear visualization. (b) Feature relationship between artists and paintings. Features of the artists and paintings are extracted by the CLIP [46] with language and visual portions. The horizontal and vertical axes are the artists and paintings respectively. The significantly larger values on the diagonal elements suggest that the features from the CLIP model are aware of the high-level painting style of a different artist. (c) shows the painting-artist correlations on the whole WikiArt (d) shows how contrastive learning is used for style transfer.

Dall-E [47] and CLIP to achieve arbitrary text-driven stylization. However, it requires hours of online optimization for one style, which makes it less applicable to practice. For this, TxST [37] proposes to learn the artistic styles through Contrastive learning offline so that arbitrary stylization can be achieved online. Diffstyler [21] utilizes the diffusion model to take the text as the condition for cross-model guidance.

3 Approach

3.1 Text for Style Transfer

Painting style, or *painting language*, represents the painting tastes of artists. It is a high-level abstract representation of the images. Previous style transfer methods [23,41,51] use a referenced image to describe the desired painting style. The core idea of our approach is to describe painting style using texts, because

of its convenience and flexibility. The benefits of the text model are twofold: (1) high-level style representation that avoids ambiguity caused by the image content; and (2) flexible adjustment with no effort for searching proper references. Recently developed text models, like CLIP [46], motivate us to ask *"Can text models be aware of different style representations?"*

We believe that text and image can work interchangeably in the CLIP space for style transfer. In other words, text and image are co-linear in the CLIP space and hence, they can both be used as style indicators. From the aspect of the style description, this co-linearity also exists between artists and their paintings, making it possible to realize artist-aware style transfer. To verify this, we have made an analysis between images and style queries using a pre-trained CLIP model [46]. Initially, we collected a set of Artist-Painting pairs from WikiArt dataset [2], as shown in Figure 2(a). To investigate the correlation between artists and paintings, the names of the artist are first tokenized and encoded to extract the feature $E_T \in \mathbb{R}^{512}$. Next, we encoded the paintings for the image feature $E_I \in \mathbb{R}^{512}$. The correlation was then calculated by the dot product between the text $E_T(t_s)$ and image features $E_I(I_s)$. For each painting I_s^i, we used the softmax function to find the probability score s for different artists t_s^j, as follows:

$$s\left(i,j\right) = \frac{\exp\left(S\left(E_I(I_s^i) \cdot E_T(t_s^j)\right)\right)}{\sum_{k=0}^{k=C} \exp\left(S\left(E_I(I_s^i) \cdot E_T(t_s^k)\right)\right)} \tag{1}$$

where S is the cosine function, C is the number of artists. A larger score $s\left(i,j\right)$ means more likely the painting I_s^i is drawn by the artist t_s^j. The relationship among scores is shown in Figure 2(b), where we observe the significantly larger values of the diagonal. Intuitively, we observe a strong correlation between the artists and their paintings in the CLIP feature space.[5] The same painting-artist correlations are also observed from the complete WikiArt (Figure 2(c)).

3.2 Proposed text-driven image style transfer (CLAST)

Figure 2(d) highlights the supervised contrastive learning process about how we can use CLIP to achieve text-driven style transfer. The CLIP loss helps us find the co-linearity between style texts and output images (the solid lines between diamonds and stars), while the contrastive similarity loss minimizes the feature distance between text-guided results $E_I^i(I_{cs}^{\mathcal{T}})$ and image-guided results $E_I^i(I_{cs}^{\mathcal{I}})$ (dashed lines between stars and circles). The names of artists (t_s) also work as labels to supervise the contrastive training process. Therefore, we guide the model to minimize the intra-class distance between different art collections

[5] There are two outliers in the figure, i.e., "**P**1-**A**6" and "**P**12-**A**5". They indicate that the painting **P**1 by *Berthe Morisot* has the highest score to the painting of artist *Monet*(**A**6). The reason is that both artists lived in the same country (France), and the artworks were created during the same artistic period (both were born in the 1840s) which we call the Impressionism movement. For another outlier "**P**12-**A**5", the painting from *Wassily Kandinsky* has similar styles to *Jackson Pollock*. Both paintings **P**5 and **P**12 in Figure 2(a) use geometric symbols in complex and abstract forms that are difficult to distinguish.

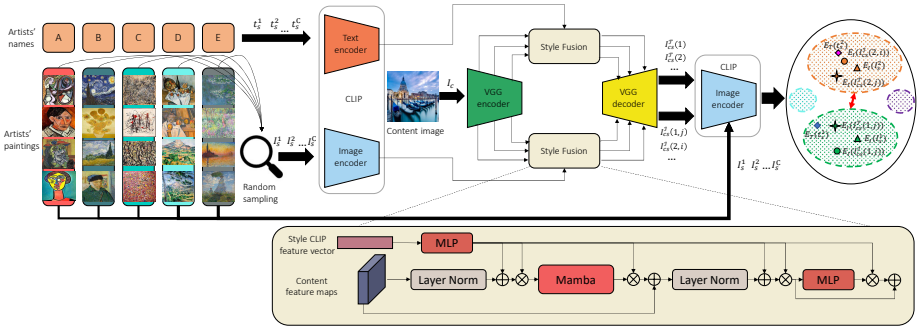

Fig. 3: Training process of the proposed CLAST . We show the overall training processes of our proposed CLAST . Given content images, artists' names and paintings, we process them by CLIP to cluster stylized results generated by texts or images. The Style Fusion module utilizes the Adaptive layer norm and Mamba structure to transfer textual or visual styles to the content images.

painted by the same painter, as well as maximize the inter-class distance between different painters. Based on this, in this section, we introduce the proposed CLAST for artist-aware image style transfer, which is illustrated in Figure 3.

Given content images I_c and desirable styles, artist's name t_s and corresponding painting I_s, CLAST outputs the stylized image I_{cs}. We denote text-guided stylized images as $I_{cs}^{\mathcal{T}}$ and image-guided stylized images as $I_{cs}^{\mathcal{I}}$. In the training stage, we first use CLIP to project images to the latent space. To preserve style consistency, the network computes the contrastive similarity to minimize the inter-distance of paired stylized images. We propose **Supervised Contrative loss (SupCon loss)** to ensure that given one artist's name, the generated $I_{cs}^{\mathcal{T}}$, $I_{cs}^{\mathcal{I}}$ and target paintings I_s can be clustered together with uniformly distinct feature representation. At test time, either text (t_s) and image (I_s) can be input as style indicators to generate consistent stylization.

Architecture. CLAST consists of four parts: Image encoder, Image decoder, CLIP (text and image encoders), and Style fusion (adaLN based state space model). The structure of the image encoder follows VGG-19 [55], discarding the fully connected layers. The image decoder is symmetric to the image encoder, with gradual upsampling feature maps towards final stylized images. We can use CLIP to encode the paired texts ($t_s^i, i = 1, 2, .., C$, where C is the number of artists) and images ($I_s^i, i = 1, 2, .., N$, where N is the number of paintings) to obtain corresponding text features $E_{\mathcal{T}}(t_s^i)$ and image features $E_{\mathcal{I}}(I_s^i)$, respectively. The VGG encoder encodes the content image, and then the Style Fusion module fuses the style and content features. Finally, the VGG decoder outputs the stylized results. Below, we introduce the proposed Style Fusion module in detail.

• **Style Fusion.** Recently, DiT [44] has shown that the Adaptive Layer Norm (adaLN) works well on the conditional diffusion model. The idea is to regress the conditional vector to obtain dimension-wise scaling and shift parameters and use them to adjust the target distribution. It resonates with the AdaIN [23] process that we can use adaptive normalization to shift the content features to the style domain for style manipulation. Following this idea, we incorporate adaLN into Style Fusion. As can be seen from Figure 3, the CLIP feature vector z is regressed by MLP to output scale and shift parameters that can represent feature distributions. The content feature x is renormalized by the learned style features as,

$$y1 = (LN(x + \alpha_1 \cdot SSM(LN(x) \cdot \mu_1 + \sigma_1)) + \alpha_2) + \sigma_2, \quad y = y1 + \alpha_2 \cdot MLP(y1) \quad (2)$$

Equation (2) illustrates that style fusion is a learnable normalization process. The target feature vector (generated from style texts or images) generates scale and shift vectors that can transform the content feature maps to the target style domain. *Mamba* implements state space modelling [19] process, which learns global feature correlations as an efficient sequence transformation.

3.3 Loss functions

• **Directional CLIP loss.** To guide the content image following the semantic of the target text (artist's name), we use the directional CLIP loss [17, 29] to align the CLIP-space direction between the text-image pairs of the target t_s and output $I_{cs}^{\mathcal{T}}$. It is defined as:

$$\Delta T = E_T(t_s) - E_T(t_o), \quad \Delta I = E_I(I_{cs}^{\mathcal{T}}) - E_I(I_c), \quad L_{clip} = 1 - \frac{\Delta I \cdot \Delta T}{|\Delta I||\Delta T|} \quad (3)$$

where t_o is the text description for the content image. We define it as *Photo* following the design in [17], which indicates no stylization is applied. Compared to original CLIP loss [46], equation (3) can stabilize the optimization process and produce results with better quality.

• **Supervised Contrative loss (SupCon loss).** To encourage image features to be correlated with the target style and uncorrelated with other styles, we propose to use Supervised Contrastive similarity loss (SupCon loss) [9] to maximize the style similarity between different paintings of the same painters. Given one content image I_c, and a set of n target artists (name, painting, label) $\{t_s^i, I_s^i, y_i\}_{i=1}^n$, $2n$ training pairs can be created by randomly selecting either artist's name or his/her painting for style transfer, $\{\hat{t}_s^i, \hat{I}_s^i, \hat{y}_i\}_{i=1}^{2n}$. Mathematically, we have,

$$L_{supcon}(I_{cs}^{\mathcal{T}}, I_{cs}^{\mathcal{T}}) = -\sum_{i=1}^{2N} \frac{1}{2|N_i| - 1} \sum_{j \in N(y_i), j \neq i} log \frac{exp(z_i \cdot z_j / \tau)}{\sum_{k \in I, k \neq i} exp(z_i \cdot z_k / \tau)} \quad (4)$$

where τ is the temperature factor, $z_k = Norm(Proj(E_I(I_{cs}^{\mathcal{T}}(k))))$, in which $Norm(\cdot)$ is the l2 norm operation, $Proj(\cdot)$ is a Linear layer of size 128, $E_I(\cdot)$ is the CLIP image encoder that encodes either stylized results $I_{cs}^{\mathcal{T}}$ edited by paintings or results $I_{cs}^{\mathcal{T}}$ edited by names. $N_i = \{j \in I : \hat{y}_j = \hat{y}_i\}$ contains a set

of indices of samples with label y_i. For one specific artist, the style similarities should exist in the paintings and stylized images. Hence, we can incorporate two more pairs as $L_{supcon}(I_s, I_{cs}^{\mathcal{T}})$ and $L_{supcon}(I_s, I_{cs}^{\mathcal{I}})$. The final SupCon loss is,

$$L_{supcon} = L_{supcon}(I_{cs}^{\mathcal{I}}, I_{cs}^{\mathcal{T}}) + L_{supcon}(I_s, I_{cs}^{\mathcal{T}}) + L_{supcon}(I_s, I_{cs}^{\mathcal{I}}) \tag{5}$$

• **Content and style feature loss.** Following existing style transfer methods [24, 36, 41], we employ content and style feature losses by using the pretrained VGG network to minimize the distance in the feature space as:

$$L_{sty} = \sum_{i}^{4} \|W_s^i(I_{cs}^{\mathcal{I}}) \times W_s^i(I_{cs}^{\mathcal{I}})^T - W_s^i(I_s) \times W_s^i(I_s)^T\|_1^1 \;\; L_{con} = \sum_{i}^{2} \|W_c^i(I_{cs}^{\mathcal{I}}) - W_c^i(I_c)\|_1^1 \tag{6}$$

As in [24], we use VGG-19 to extract W_s features (*relu1_2*, *relu2_2*, *relu3_4*, *relu4_1*) to compute the style loss. We also extract W_c features (*relu2_2*, *relu3_4*) to compute the content loss.

• **Total loss.** We train the network using the losses from Eqs. ((3)) to ((6)). We also utilize LPIPS loss [64] (L_{lpips}) to supervise perceptual similarity. Hence, we define the final loss as $L = \lambda_{clip}L_{clip} + \lambda_{supcon}L_{supcon} + \lambda_{sty}L_{sty} + \lambda_{con}L_{con} + \lambda_{lpips}L_{lpips}$, where $\lambda_{clip}, \lambda_{clip_f}, \lambda_{supcon}, \lambda_{sty}, \lambda_{con}$, and λ_{lpips} are coefficients to balance these loss components.

4 Experiments

4.1 Implementing Details

• **Datasets.** We use the images from MS-COCO [33] (about 118k images) for the image reconstruction task in the first training stage. In the second training stage, we train CLAST with MS-COCO [33] as our content image set and WikiArt [2] (about 81k images) as the style image set. In the training phase, we load the images with the size of 512×512 and randomly crop them as training patches of size 256×256. As data augmentation, we randomly flip the content and style images. For inference, our CLAST can handle the images with any resolution. In this Section, we use the images with 512×512 resolution for a fair comparison.

• **Parameter setting.** We train CLAST using Adam optimizer with the learning rate of 1×10^{-4}. The batch size is set to 30 and CLAST is trained for 100k iterations (about 8 hours) on a PC with one NVIDIA V100 GPU using PyTorch deep learning platform. The weighting factors in the total loss are defined empirically as: $\lambda_{clip} = 1, \lambda_{lpips} = 1, \lambda_{supcon} = 2, \lambda_{sty} = 50, \lambda_{con} = 0.02$.

• **Metrics and evaluation.** We use CLIP loss (text/image-image cosine similarity) [29] to measure the semantic similarity between the target texts/images and stylized images. For artist awareness, we follow AST [50] to compute the style transfer deception rate, which is calculated as the fraction of generated images that are classified by the VGG-16 network as the artworks of an artist for which the stylization was produced. A higher value means closer to the artist's style. We also show image-driven experiments in the supplementary to demonstrate the versatility of our CLAST , which takes either text or image for stylization.

4.2 Text-driven Artistic-aware Style Transfer

Feature extracted by the CLIP model has high-semantic discriminate power that can be used for similarity measurement. Therefore, we compute CLIP scores. First, we extract the feature embedding from the content, style, and transfer images by CLIP, then we compute content s_{cont} and style s_{style} scores, as defined by Equations (6), for which the higher score means better performance (↑ in the table). SSIM and VGG content loss are also used to compute content differences.

$$s_{cont}\left(I_c, I_{cs}\right) = \frac{E_I(I_c) \cdot E_I(I_{cs})}{\|E_I(I_c)\| \times \|E_I(I_{cs})\|}, \quad s_{style}\left(t_s, I_{cs}\right) = \frac{E_T(t_s) \cdot E_I(I_{cs})}{\|E_T(t_s)\| \times \|E_I(I_{cs})\|} \tag{6}$$

To show the efficiency of our proposed approach, we compare our approach with AST [50], MGAD [20], LDAST [16], CLIPstyler [29] (CLIPstyler(fast) and CLIPstyler(opti)), DALL-E-2 [47], VQGAN-CLIP [48], Stable Diffusion [49][6], TxST [37], DiffusionCLIP [25] and DiffStyler [21] With default diffusion settings (e.g., steps).

Method	Clip Scores Content↑	Style↑	SSIM↑	Deception Rate↑	Running time (s)
AST [50]	0.538	0.269	0.235	0.664	1.3
CLIPstyler(fast) [29]	0.736	0.254	0.334	0.469	0.9
CLIPstyler(opti) [29]	0.624	0.306	0.407	0.441	220
LDAST [16]	0.669	0.207	0.255	0.435	1.6
MGAD [20]	0.397	0.203	0.338	0.339	604
VQGAN-CLIP [48]	0.557	0.230	0.217	0.682	240
DALL-E-2 [47]	0.665	0.228	0.358	0.425	34
Stable Diffusion [49]	0.542	0.332	0.445	0.702	37
TxST [37]	0.678	0.313	0.487	0.769	0.7
DiffusionCLIP [25]	0.602	0.443	0.435	0.669	530
DiffStyler [21]	0.540	0.334	0.301	0.603	6.5
Ours	0.667	0.402	0.491	0.747	0.03

Table 1: Text-driven style transfer comparison. We use Clip style scores and the deception rate to measure the style similarity. Clip content score is used to measure content preservation.(Red: best and Blue: 2^{nd} best).

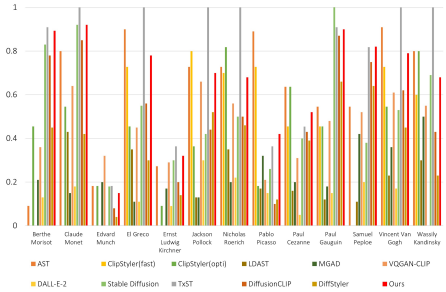

Fig. 4: Deception rate of text-driven style transfer methods. We show the deception rate of each artist from a WikiArt subset. Our method (red columns) achieves better performance than others.

Quantitative Comparison. Table 1 reports the results on 512 × 512 image stylization. We observe that CLIPstyler(fast) leads to the best content similarity score (0.736); however, the transferred images have poor artistic style performance. The results of AST have the second-best deception rate, but the similarity to the content image is only 0.538. CLIPstyler(opti), a slow but optimal version compared to CLIPstyler(fast), reaches good CLIP scores but achieves the worst deception rate. This is expected since CLIPstyler(opti) requires dedicated training for each individual content and style image. DiffusionCLIP [25] leads to the

[6] `https://beta.dreamstudio.ai/dream`, we utilized the official model that employs a content image and style prompt as input to apply the Stable unCLIP inference.

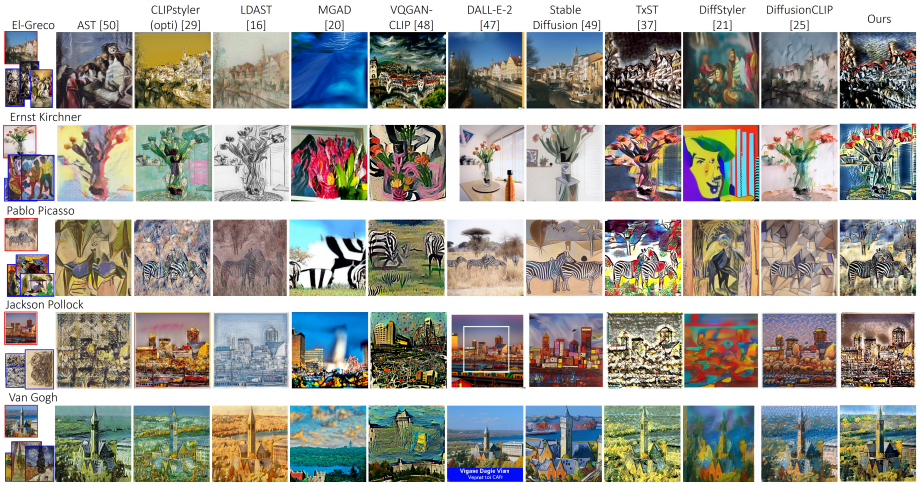

Fig. 5: Compare among text-dirven style transfers. We show five examples (content images in red boxes) using eight different methods. We use the names of painters as text prompts for style transfer. For reference, we also list representative paintings in blue boxes for comparison. The styles of artists are very abstract and subjective. We highly recommend readers check each artist online for better comparison.

best CLIP style score but with poor performance on content scores. TxST [37] performs best on the deception rate but the CLIP style score is low. Overall, our CLAST provides a good balance between style and content, reaching the best performance on inference time (over 100× speedup) and second-best CLIP score and deception rate, which demonstrates its effectiveness. It is important to note that CLIPstyler(opti) MGAD, VQGAN-CLIP and DiffusionCLIP need very long training time for each artist; in contrast, CLAST does not need retraining, and therefore it is much faster than others. Figure 4 shows the class-specific accuracy of the deception rate. We observe that using our CLAST (red bars) achieves better performance for all artists.

User study. For user study, we invite users from different backgrounds, like art, design, literature, and science, to ensure the study is as fair as possible. The questionnaire compares the painting styles of eight artists from the WikiArt subset. For each artist, we collect the results from AST [50], CLIPstyler(opti) [29], DALL-E-2 [47], VQGAN-CLIP [48], Stable diffusion [49] and the proposed CLAST by using three different content images. In each question, the users were given three results from the same content image using the six methods. The users are requested to rank the style similarity to the artist, where 6 denotes the most similar, and 1 means the most different. We average the scores from all users as the Mean Opinion Score (MOS). The results are shown in Figure 6. We observe that CLAST achieves the highest average MOS,

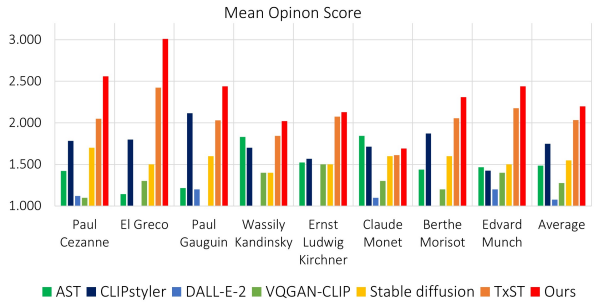

Fig. 6: User study on text-driven style transfer. We use the names of eight artists in WikiArt as text input for style transfer. We invite users to rank different approaches. The higher the Mean Opinion Score (MOS), the higher the style similarity to the target artists.

Method	Aes Score
AST [50]	4.976
CLIPstyler [29]	5.415
DALL-E-2 [47]	4.052
VQGAN-CLIP [48]	4.360
Stable Diffusion [49]	5.268
DiffusionCLIP [25]	4.368
DiffStyler [21]	3.723
TxST [37]	5.436
Our	5.667

Table 2: Comparison on Aesthetic score. The pretrained Aesthetic regression model [45] is used to rank different methods to see which one is most preferable to humans.

0.1∼0.5 improvements compared to Clipstyler, Stable diffusion and TxST, which suggests that our results have the most similar painting style to the target artists *at a perceptual level*. Quantitatively, we also utilize aesthetic scores [45] to rank all text-driven results to see which one is mostly consistent with human preferences. It was trained on over 238000 AI synthetic images with human rating scores. It can demonstrate that ours is more visually pleasant to humans.

Qualitative Comparison. Figure 5 shows the visual comparison of different methods for artist-aware style transfer. Note that AST does not require text input. For others, we use artists' names as texts to guide stylization. For reference, we also show three representative paintings in blue boxes. Our findings are summarized as: (1) our results show similar or better content preservation compared to AST and CLIPstyler. (2) CLAST can faithfully mimic the signature styles of specific artists, such as the color tone and temperature patterns from *El-Greco, Van Gogh* and the distorted curves in *El-Greco* and *Van Gogh*. (3) VQGAN-CLIP, DiffStyler generates bizarre images resembling random generation. The official DALL-E-2 cannot perform style transfer but changes contents to some extent. Stable diffusion and DiffusionCLIP achieve visually pleasing results in *P. Picasso* but it cannot learn the most representative style for other artists.

4.3 Ablation Study

• **Loss terms.** Here we ablate the losses CLAST is trained on and report the results in Table 3. *Baseline* is the model that uses content L_{con} and style L_{style} losses. We observe that using either Directional CLIP loss (L_{CLIP}, 2nd row) or unsupervised contrastive loss (L_{unsup}, 3rd row) improves the CLIP style scores compared to the Baseline. This implies that these two losses can guide the stylization close to the target text description. By adding LPIPS loss (L_{lpips}), we can see from row 5 that it improves the VGG content loss and SSIM approximately

Loss Terms	VGG ↓ content	CLIP↑ style score	SSIM↑
Baseline ($L_{\text{con}} + L_{\text{style}}$)	80.12	0.213	0.402
Baseline+L_{clip}	86.40	0.355	0.383
Baseline+L_{unsup}	82.05	0.218	0.400
Baseline+$L_{\text{clip}} + L_{\text{unsup}}$	82.18	0.369	0.394
Baseline+$L_{\text{clip}} + L_{\text{unsup}} + L_{\text{lpips}}$	78.26	0.354	0.481
Baseline+$L_{\text{clip}} + L_{\text{supcon}} + L_{\text{lpips}}$	78.56	0.402	0.479

Table 3: Comparison on loss terms. To show the effect of different loss terms, we compare the content preservation by VGG content loss and SSIM, and CLIP style score to measure style similarity.

Style Fusion module	Training ↓ Time (hours)	Inference ↓ time (seconds)	Number of parameters
Attention+AdaIN	8.7 (1×)	0.62 (1×)	2362112
Attention+adaLN	8.5 (1×)	0.62 (1×)	2393421
Linear_Attn+AdaIN	6.5 (1.3×)	026 (1.2×)	1959408
Linear_Attn+adaLN	5.7 (1.5×)	0.32 (1.2×)	1989237
Mamba+AdaIN	6.1 (1.4×)	0.03 (20×)	1399893
Mamba+adaLN	4.9 (1.8×)	0.03 (20×)	1423104

Table 4: Comparison on style fusion modules. The training time starts from the same initialization, to reach the target CLIP score of 0.4. Both training and inference are tested on one V100 GPU.

by 3.9 and 0.09, respectively. The last row shows that our proposed supervised contrastive loss (L_{supcon} can improve the CLIP style score by 0.05 with slight decreases in the VGG content loss and SSIM.

• **Style Fusion Speedup.** We also compare the design of the style fusion module that can learn the cross-modal correlations between texts and images. From Table 4, we compare ours (last row) with others, including the vanilla Attention [41] module and AdaIN [23] structure. We can see that using our proposed style fusion module can speed up the training and testing by 1.8× and 20× with 59% fewer parameters.

4.4 Video style transfer and failure analysis

To demonstrate the style consistency of our method, one video sequence is applied to different style transfer approaches. To measure the motion changes, we use the open package Pyflow [7] and visualize the motion magnitude and direction. Figure 7 demonstrates the results. We observe that using our method can obtain similar optical flows as the original video, which (1) firstly reveals its consistency in stylization, i.e., frames from the same video can remain the same artist's style and (2) secondly, highlights its content preservation, i.e., the motion changes do not affect much the stylization quality. On the other hand, other methods do not learn the correct style (*Albrecht Durer* is famous for his woodcut print, for which other approaches fail to grasp this style) and they also fail to provide consistent stylization.

Overall, our proposed CLAST can perform well for text-guided style transfer. However, (1) for artistic style transfer, it performs poorly for artists with few examples in WikiArt [50]. In Figure 8 (a), WikiArt only contains 20 paintings of *Walter Osbon*, a painter of the post-impressionism movement. We observe that CLAST does not predict the most distinct style for stylization. (2) When the content image contains rich information, CLAST performs poorly for content preservation. In Figure 8 (b), though CLAST successfully adds fog to the image and blurs some contents, it also changes the color tone and loses some details. (3)

[7] `https://github.com/pathak22/pyflow`

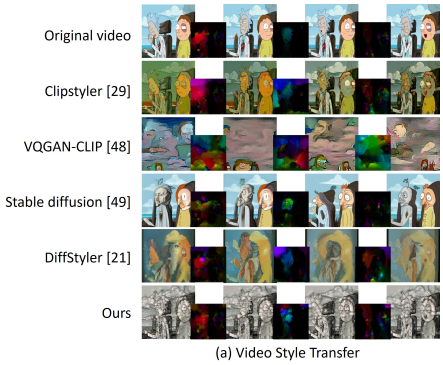

Fig. 7: Video-based text-driven style transfer. Four neighborhood frames and their optical flows are shown. We use *Albrecht Durer* as the text prompt to apply five different style transfer.

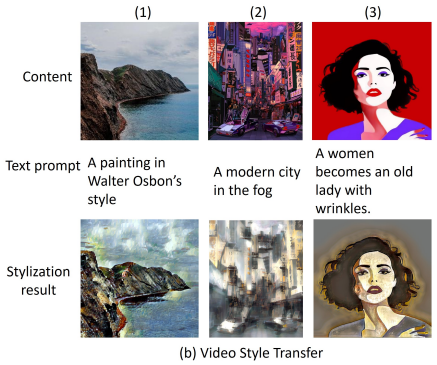

Fig. 8: Failure cases of text-driven style transfer. Three failure cases are shown here, including unbalanced training, insensitive to general texts, and text prompt misunderstanding.

CLAST may ignore texts with detailed descriptions. In Figure 8 (c), it misinterprets the "old lady" as an old picture and does not add wrinkles to the woman's face. All failure cases can be resolved by following the same pipeline: (i) training the model on a large-scale dataset, (ii) increasing the length of text prompts with more details, and (iii) computing losses for multiscale content preservation; this is, however, out of the scope of this work and we leave it for future work.

5 Conclusion

In this paper, we proposed a text-driven approach for artist-aware style transfer, coined CLAST. To extract style descriptions effectively from the CLIP-based image-text space, CLAST leverages a supervised contrastive similarity training strategy and a new adaLN-SSM based module for style fusion. This approach explores the inherent relationship between texts and images by clustering artistic styles into different groups, eliminating the need for extensive data collection and online training. Extensive results show that CLAST achieves perceptually pleasing arbitrary stylization in real time, revealing its ability to extract critical representations from the CLIP space and produce aesthetics close to the artists' works. CLAST also points to a new direction for text-driven style transfer. Future work includes combining images, texts, and other cues to deliver a more flexible user-guided style transfer.

Acknowledgements

This work was supported by ANR-22-CE23-0007, Hi!Paris collaborative project.

References

1. Claude monet and paul cezanne. https://www.claude-monet.com/monet-and-cezanne.jsp (2010)
2. K. nichol. painter by numbers, wikiart. https://www.kaggle.com/c/painter-by-numbers (2016)
3. Adverb: The bigsleep: Biggan+clip. https://www.kaggle.com/c/painter-by-numbers (2022)
4. An, J., Huang, S., Song, Y., Dou, D., Liu, W., Luo, J.: Artflow: Unbiased image style transfer via reversible neural flows. In: CVPR (2021)
5. Audebert, N., Herold, C., Slimani, K., Vidal, C.: Multimodal deep networks for text and image-based document classification. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (2019)
6. Bianchi, F., Attanasio, G., Pisoni, R., Terragni, S., Sarti, G., Lakshmi, S.: Contrastive language-image pre-training for the italian language. arXiv preprint arXiv:2108.08688 (2021)
7. Chen, H., Zhao, L., Wang, Z., Ming, Z.H., Zuo, Z., Li, A., Xing, W., Lu, D.: Artistic style transfer with internal-external learning and contrastive learning. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021), https://openreview.net/forum?id=hm0icunzGW
8. Chen, H., Zhao, L., Wang, Z., Zhang, H., Zuo, Z., Li, A., Xing, W., Lu, D.: Dualast: Dual style-learning networks for artistic style transfer. In: CVPR (2021)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations (2020)
10. Cong, G., Li, L., Liu, Z., Tu, Y., Qin, W., Zhang, S., Yan, C., Wang, W., Jiang, B.: Ls-gan: Iterative language-based image manipulation via long and short term consistency reasoning. p. 4496–4504. MM '22, New York, NY, USA (2022)
11. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: Stytr2: Image style transfer with transformers. In: CVPR (2022)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
13. Dufour, N., Picard, D., Kalogeiton, V.: Scam! transferring humans between images with semantic cross attention modulation. In: ECCV (2022)
14. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021)
15. Frans, K., Soros, L.B., Witkowski, O.: Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. arXiv preprint arXiv:2106.14843 (2021)
16. Fu, T.J., Wang, X.E., Wang, W.Y.: Language-driven artistic style transfer. In: ECCV (2022)
17. Gal, R., Patashnik, O., Maron, H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators. arXiv preprint arXiv:2108.00946 (2021)
18. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: CVPR (2016)
19. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
20. Huang, N., Tang, F., Dong, W., Xu, C.: Draw your art dream: Diverse digital art synthesis with multimodal guided diffusion. p. 1085–1094. MM '22 (2022)

21. Huang, N., Zhang, Y., Tang, F., Ma, C., Huang, H., Dong, W., Xu, C.: Diffstyler: Controllable dual diffusion for text-driven image stylization. IEEE Transactions on Neural Networks and Learning Systems pp. 1–14 (2024). https://doi.org/10.1109/TNNLS.2023.3342645
22. Huang, S., An, J., Wei, D., Luo, J., Pfister, H.: Quantart: Quantizing image style transfer towards high visual fidelity. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5947–5956 (jun 2023). https://doi.org/10.1109/CVPR52729.2023.00576
23. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017)
24. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016)
25. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2416–2425 (2022). https://doi.org/10.1109/CVPR52688.2022.00246
26. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: arXiv preprint arXiv:1312.6114 (2014)
27. Kolkin, N., Salavon, J., Shakhnarovich, G.: Style transfer by relaxed optimal transport and self-similarity. In: CVPR (2019)
28. Kuhnle, A., Copestake, A.: Shapeworld - a new test methodology for multimodal language understanding. arXiv preprint arXiv:1704.04517 (2017)
29. Kwon, G., Ye, J.C.: Clipstyler: Image style transfer with a single text condition. arXiv preprint arXiv:2112.00374 (2021)
30. Köhler, J., Krämer, A., Noé, F.: Smooth normalizing flows. In: NeurIPS (2021)
31. Li, X., Liu, S., Kautz, J., Yang, M.H.: Learning linear transformations for fast arbitrary style transfer. In: CVPR (2019)
32. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal style transfer via feature transforms. NeurIPS (2017)
33. 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: ECCV (2014)
34. 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: ICCV (2021)
35. Liu, X., Park, D.H., Azadi, S., Zhang, G., Chopikyan, A., Hu, Y., Shi, H., Rohrbach, A., Darrell, T.: More control for free! image synthesis with semantic diffusion guidance. arXiv preprint arXiv:2112.05744 (2021)
36. Liu, Z.S., Kalogeiton, V., Cani, M.P.: Multiple style transfer via variational autoencoder. In: ICIP (2021)
37. Liu, Z.S., Wang, L.W., Siu, W.C., Kalogeiton, V.: Name your style: text-guided artistic style transfer. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3530–3534 (2023). https://doi.org/10.1109/CVPRW59228.2023.00359
38. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020)
39. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: CLIP4Clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021)
40. van den Oord, A., Vinyals, O., kavukcuoglu, k.: Neural discrete representation learning. In: NeurIPS (2017)

41. Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: CVPR (2019)
42. Park, J., Kim, S., Kim, S., Yoo, J., Uh, Y., Kim, S.: Lanit: Language-driven image-to-image translation for unlabeled data. arXiv preprint arXiv:2208.14889 (2022)
43. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: ICCV (2021)
44. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4172–4182 (2023). https://doi.org/10.1109/ICCV51070.2023.00387
45. Pressman, J.D., Crowson, K., Contributors, S.C.: Simulacra aesthetic captions. Tech. Rep. Version 1.0, Stability AI (2022), https://github.com/crowsonkb/simulacra-aesthetic-models
46. 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 (2021)
47. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. vol. 139, pp. 8821–8831 (2021)
48. Rodent, N.: Vqgan-clip. https://github.com/nerdyrodent/VQGAN-CLIP (2022)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10674–10685 (2022). https://doi.org/10.1109/CVPR52688.2022.01042
50. Sanakoyeu, A., Kotovenko, D., Lang, S., Ommer, B.: A style-aware content loss for real-time hd style transfer. In: ECCV (2018)
51. Sanakoyeu, A., Kotovenko, D., Lang, S., Ommer, B.: A style-aware content loss for real-time hd style transfer. In: ECCV (2018)
52. Schaldenbrand, P., Liu, Z., Oh, J.: Styleclipdraw: Coupling content and style in text-to-drawing synthesis. NeurIPS Workshop on Machine Learning and Design (2021)
53. Schwettmann, S., Hernandez, E., Bau, D., Klein, S., Andreas, J., Torralba, A.: Toward a visual concept vocabulary for gan latent space. In: ICCV (2021)
54. Sheng, L., Lin, Z., Shao, J., Wang, X.: Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In: CVPR (2018)
55. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
56. Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.: Texture networks: Feed-forward synthesis of textures and stylized images (2016)
57. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
58. Wang, J., Yang, H., Fu, J., Yamasaki, T., Guo, B.: Fine-grained image style transfer with visual transformers (2022). https://doi.org/10.48550/ARXIV.2210.05176, https://arxiv.org/abs/2210.05176
59. Wang, W., Yang, S., Xu, J., Liu, J.: Consistent video style transfer via relaxation and regularization. IEEE TIP (2020). https://doi.org/10.1109/TIP.2020.3024018
60. Wang, Z., Zhao, L., Xing, W.: Stylediffusion: Controllable disentangled style transfer via diffusion models. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7643–7655 (2023). https://doi.org/10.1109/ICCV51070.2023.00706

61. Xu, W., Long, C., Nie, Y.: Learning dynamic style kernels for artistic style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10083–10092 (June 2023)
62. Xu, Z., Sangineto, E., Sebe, N.: Stylerdalle: Language-guided style transfer using a vector-quantized tokenizer of a large-scale generative model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7601–7611 (October 2023)
63. Yang, S., Hwang, H., Ye, J.C.: Zero-shot contrastive loss for text-guided diffusion image style transfer. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22816–22825 (2023). https://doi.org/10.1109/ICCV51070.2023.02091
64. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. CVPR (2018)
65. Zhang, X., Sha, Y., Kampffmeyer, M.C., Xie, Z., Jie, Z., Huang, C., Peng, J., Liang, X.: Armani: Part-level garment-text alignment for unified cross-modal fashion design. p. 4525–4535. MM '22 (2022)
66. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10146–10156 (jun 2023). https://doi.org/10.1109/CVPR52729.2023.00978
67. Zhang, Y., Tang, F., Dong, W., Huang, H., Ma, C., Lee, T.Y., Xu, C.: Domain enhanced arbitrary image style transfer via contrastive learning. In: ACM SIGGRAPH (2022)
68. Zhu, M., He, X., Wang, N., Wang, X., Gao, X.: All-to-key attention for arbitrary style transfer. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 23052–23062. Los Alamitos, CA, USA (oct 2023). https://doi.org/10.1109/ICCV51070.2023.02112