Multimodality-guided Image Style Transfer using Cross-modal GAN Inversion

Hanyu Wang[1]†, Pengxiang Wu[2], Kevin Dela Rosa[2], Chen Wang[2], Abhinav Shrivastava[1]
[1]University of Maryland, College Park [2]Snap Inc.
hywang66@umd.edu {pwu,kevin.delarosa,chen.wang}@snapchat.com abhinav@cs.umd.edu

Figure 1. **Our multimodality-guided image style transfer results.** We show style transfer results guided by two text styles (left) and two imag styles (right). 15 stylized images are synthesized by evenly interpolating between these four styles.

Abstract

Image Style Transfer (IST) is an interdisciplinary topic of computer vision and art that continuously attracts researchers' interests. Different from traditional Image-guided Image Style Transfer (IIST) methods that require a style reference image as input to define the desired style, recent works start to tackle the problem in a text-guided manner, i.e.*, Text-guided Image Style Transfer (TIST). Compared to IIST, such approaches provide more flexibility with text-specified styles, which are useful in scenarios where the style is hard to define with reference images. Unfortunately, many TIST approaches produce undesirable artifacts in the transferred images. To address this issue, we present a novel method to achieve much improved style transfer based on text guidance. Meanwhile, to offer more flexibility than IIST and TIST, our method allows style inputs from multiple sources and modalities, enabling MultiModality-guided Image Style Transfer (MMIST). Specifically, we realize MMIST with a novel cross-modal GAN inversion method, which generates style representations consistent with specified styles. Such style representations facilitate style transfer and in principle generalize any IIST methods to MMIST. Large-scale experiments and user studies demonstrate that our method achieves state-of-the-art performance on TIST task. Furthermore, comprehensive qualitative results confirm the effectiveness of our method on MMIST task and cross-modal style interpolation.*

1. Introduction

As a research topic at the intersection of computer vision and art, Image Style Transfer (IST) aims to apply certain style patterns to a given content image. The seminal work of Gatys *et al*. [13] proposed to transfer the style of

[2]Work done during internship at Snap Inc.

one image to another content image by optimizing the pixel values using both style and content losses, inspiring many subsequent works in this field. To speed up the style transfer process, [17] trained a feed-forward neural network for each style to transfer it to different contents. Going beyond single style transfer, [15] introduced the idea of arbitrary style transfer and aimed to transfer arbitrary styles to any content in a single forward pass. Based on this formulation, [6, 16, 27, 29, 33, 37, 42, 48] improved [15] on multiple aspects.

The above-mentioned methods can be classified as *Image-guided Image Style Transfer (IIST)*. They rely on reference style images, which are not always accessible in real-world scenarios. For example, artists may conceive novel styles that can be easily described via texts but never exist in previous artworks. Such a dependence on reference style images limits the application of IIST methods [25].

Recently, based on the large-scale image-text pretrained model CLIP [35], several methods proposed to edit images purely conditioned on text descriptions, achieving *Text-guided Image Style Transfer (TIST)* [12, 21, 25, 34]. Notably, by training a lightweight U-Net on a single content image using CLIP loss, CLIPStyler [25] can synthesize stylized images from arbitrary content images and style text descriptions, setting a state-of-the-art (SOTA) performance on this task. However, although the styles of transferred images by CLIPStyler are generally consistent with the corresponding text descriptions, CLIPStyler often adds undesirable local patterns to the stylized images, distorting the original content severely, as shown in Figure 2. This indicates CLIPStyler fails to disentangle the style and content information from both text and image.

To address these issues, we propose a novel framework to better manipulate images based on reference style texts. Meanwhile, to offer more flexibility than IIST and TIST methods, our framework is designed to accept style guidance from multiple sources and modalities, enabling *MultiModality-guided Image Style Transfer (MMIST)*. The ability to exploit multimodal style references can be useful in many scenarios. For example, an artist may design new artistic styles by modifying styles of existing artworks; such modified styles can be easily defined by combining text descriptions and existing art images, yet are difficult to describe with text or image reference only.

To realize MMIST, we propose a novel cross-modal GAN inversion method which generates diverse style representations according to multi-modal style inputs (*e.g.*, text and image). Such generated style representations allow us to generalize any IIST methods to tackle the problem of MMIST. Specifically, we leverage a pretrained GAN model and invert style text descriptions and/or style images into GAN's latent space to get the corresponding style reference images. In this process, style-specific CLIP-based guidance is used to connect the domains of text and image. After obtaining the style reference images, we then feed them into an existing IIST approach which is adapted to take multiple style references as input. To further enhance the quality of stylized images, we propose a novel multi-style boosting strategy which enriches the style patterns. Similar to learned model parameters, the style representations can be reused at test time, allowing our method to stylize arbitrary contents in a single forward pass.

Figure 2. **Failure case of CLIPStyler.** We show one content and one style text description, together with the results from CLIPStyler and our method. CLIPStyler adds many small face-like patterns to the stylized images.

We evaluate our framework on 44 style text descriptions and 61 content images, which result in 2,684 style-content combinations. Both qualitative and user study results clearly show the improvement of our framework over previous methods on TIST task. Furthermore, extensive experiments also confirm the effectiveness of our framework on MMIST task and cross-modal style interpolation.

Our main contributions can be summarized as follows:

- We introduce a more general problem than IIST and TIST, *i.e.*, MultiModality-guided Image Style Transfer (MMIST), and solve it with a novel framework. The proposed framework can transfer styles from arbitrary number of reference images/texts to arbitrary content, a task which is not feasible for all existing methods to the best of our knowledge.
- We propose a novel cross-modal GAN inversion method to distill styles from different modalities. This inversion procedure also enables our method to interpolate between different styles arbitrarily.
- Extensive experiments and large-scale user studies (5,041 users) confirm the effectiveness of our model in terms of both qualitative results and user preference.

2. Related Work

2.1. Image-guided Image Style Transfer

Image-guided Image Style Transfer (IIST) has become a popular research topic since the seminal work of Gatys *et al*. [13]. Using a deep feature-based style loss and content loss, Gatys *et al*. [13] directly optimized pixel values to obtain decent style transfer results. To address the slow op-

timization problem in this method, [17] and several subsequent works [39–41] proposed to train feed-forward neural networks that can apply the pretrained style to arbitrary images in a single forward pass. Given its success, this idea was further developed by allowing one single trained model to store multiple styles. For example, [5] used multiple convolutional filter banks to explicitly represent multiple styles; [10] proposed conditional instance normalization to achieve the same goal; and [28] introduced a selector structure to support incremental learning for new styles.

Recently, arbitrary IIST started to attract widespread attention due to its effectiveness, efficiency, and flexibility. [15] proposed AdaIN to perform IST by adaptively aligning the mean and variance of content features with those of style features. With similar motivation, [6, 16, 27, 33] introduce new loss functions or novel mechanisms to improve the style transfer quality. [29] adaptively performs attentive normalization on a per-point basis. Besides quality, some works focus on other properties of IST methods, *e.g.*, domain-awareness [14] or brushstroke-level optimization [24].

2.2. Text-guided Image Manipulation

Text-guided image manipulation aims to manipulate the input image based on a text description while preserving text-irrelevant parts in the original image. [9] employed a GAN-based encoder-decoder model to achieve this goal. [32] further introduced a text-adaptive discriminator to ensure that only text-related regions are modified. By extending GAN-based text-to-image generators [44–46], Li *et al.* [26] proposed ManiGAN to manipulate images in a multi-stage manner. [43] utilized GAN inversion, visual-linguistic similarity learning, and instance-level optimization to build a unified framework for multimodal image generation and manipulation with text.

Recently, the success of CLIP [35] in connecting the domains of image and text has inspired a new direction to achieve text-guided image manipulation. By modifying the latent space of StyleGAN [18–20] using CLIP guidance, StyleCLIP [34] can perform text-guided manipulation in three different ways. VQGAN-CLIP [7] achieved the same goal by using CLIP loss to optimize the latent space of VQGAN [11]. StyleGAN-NADA [12] proposed a directional CLIP loss to optimize a GAN model instead of the latent space, resulting in more accurate manipulation effects. More recently, diffusion models [8, 22, 36, 38] have been combined with CLIP to obtain better performance [3, 21].

Most of these works are intended for content or attribute editing. Although some of them [12, 21] can be applied to style editing or transfer, the quality they can achieve is far from desirable. By contrast, CLIPStyler [25] is specifically designed to solve the task of TIST, outperforming all previous image manipulation methods on this task. However, CLIPStyler still suffers from certain drawbacks, as illustrated in Figure 2. Similar to CLIPStyler, [30] proposed to transfer an image's style conditioned on an artist's name. In this paper, we generalize the tasks of TIST and IIST to MMIST, and solve it under a unified framework.

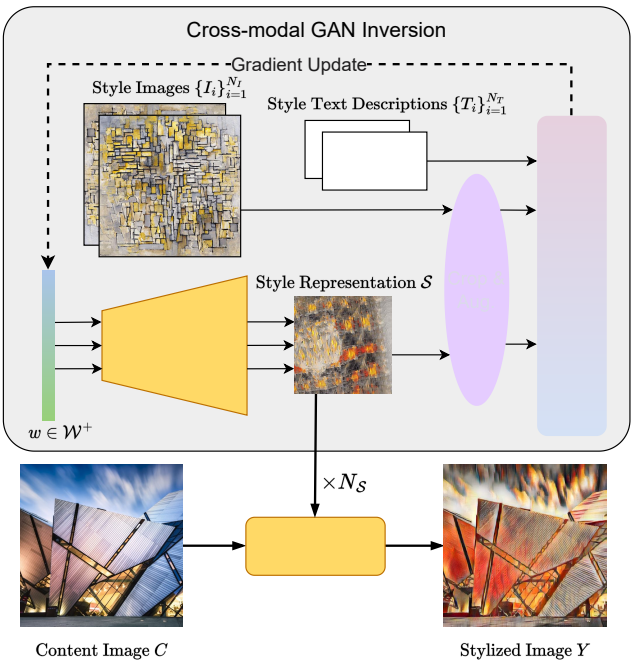

Figure 3. **The overview of our method.** Taking styles from multiple modalities as input, our method generates style representations using cross-modal GAN inversion. With the adapted IIST model, our method can apply the cached styles to any unseen content image in a single forward pass.

2.3. GAN Inversion

Traditional GAN inversion aims to invert a given image back into the latent space of a pretrained GAN generator. It emphasizes the accuracy and fidelity of the reconstructed image. GAN inversion can be applied to a wide range of downstream tasks, including image manipulation [12, 34, 43], image interpolation [1], image generation [43], etc. Learning-based [49], optimization-based [1], or hybrid [2] methods have been developed to invert a GAN. All of them use a reconstruction loss such as L_2 loss or LPIPS loss [47].

In this paper, we propose cross-modal GAN inversion. Different from traditional methods that pursue a perfect reconstruction of the whole input image, our cross-modal GAN inversion only reconstructs partial information of the input, *i.e.*, style, which is defined by inputs of multiple modalities such as text and image.

3. Method Overview

Given a set of style images $\{I_i\}_{i=1}^{N_I}$ and a set of style text descriptions $\{T_i\}_{i=1}^{N_T}$, our framework applies these specified styles to a set of content images $\{C_i\}_{i=1}^{N_C}$ and synthesizes a corresponding set of stylized images $\{Y_i\}_{i=1}^{N_C}$. Different from previous image editing methods that directly optimize the stylized image [25], the latent of the content image [7, 34], or parameters of a generative model [12, 21], we train a model for certain styles in a content-agnostic manner.

Our key insight is that MMIST can be achieved with the aid of style representations that comply with the input style text descriptions and image patterns. More specifically, we can generalize IIST methods to leverage such style representations and thereby create stylized images guided by multiple modalities. To this end, we propose a novel cross-modal GAN inversion method to map all input multimodal style references into the $\mathcal{W}^+$ space [1] of a pretrained StyleGAN3 [18] generator $\mathbf{G}$, and thereby generate intermediate style representations $\{\mathcal{S}_i\}_{i=1}^{N_\mathcal{S}}$ in the image space. As is shown in Figure 3, the style representations are images consisting of style *patterns* without meaningful contents. The cross-modal GAN inversion ensures that $\{\mathcal{S}_i\}_{i=1}^{N_\mathcal{S}}$ summarizes and combines the style information from all input styles $\{I_i\}_{i=1}^{N_I}$ and $\{T_i\}_{i=1}^{N_T}$. To leverage $\{\mathcal{S}_i\}_{i=1}^{N_\mathcal{S}}$, we adapt an existing IIST method to make it compatible with multiple style inputs. Denote the adapted IIST method by $\mathbf{M}$. We use $\mathbf{M}$ to stylize the content images $\{C_i\}_{i=1}^{N_C}$ with intermediate style representations $\{\mathcal{S}_i\}_{i=1}^{N_\mathcal{S}}$, producing stylized outputs $\{Y_i\}_{i=1}^{N_C}$.

Separating style representation generation from stylized image synthesis is the key to the success of our framework. When performing TIST task, previous methods [7, 21, 25] always deal with style alignment and content preservation simultaneously, resulting in distorted content or irrelevant artifacts that appear in the results. In contrast, by leveraging the strong style-content disentangling ability of IIST approaches, our method can put the entire focus on style representation generation for creating high-quality stylized images. Besides, with generated intermediate style representations, only a single forward pass is needed for our method to apply a learned style to any unseen content image.

4. Cross-modal GAN Inversion

To generate style representations from modalities other than image or text only, we propose cross-modal GAN inversion. In Table 1, we compare it with traditional GAN inversion. The goal of traditional GAN inversion is to faithfully reconstruct the original input image. Naturally, this only works for the image modality, and only one image at a time. Since it is targeting pixel-wise reconstruction, all information from the input image is supposed to be stored in the latent space of the GAN generator. However, the goal of cross-modal GAN inversion is completely different. It aims to combine different styles together to generate intermediate style representations. Therefore, the inversion should be able to accept multiple inputs from different references and modalities. Besides, only the style components of inputs are required to be inverted as their content parts are irrelevant to the downstream task.

Table 1. **Comparison between traditional GAN inversion and cross-modal GAN inversion.** "Ref." means reference.

Method	Ref. Modality	Number of Ref.	Inversion Target
Traditional	Image Only	Single	Original Image
Cross-modal	Multiple	Multiple	Style

4.1. Style-specific CLIP Loss

We employ CLIP [35] to connect image with other modalities, as well as to extract style components. However, naively applying CLIP cosine similarity loss does not result in accurate style representation, since the content components are also entangled in the CLIP embedding space.

Following [25] and [12], we employ patch-wise CLIP loss to address this problem. Formally, denote the pretrained CLIP image encoder by E_I and text encoder by E_T. For each style text description T_i, we use the text-image patch-wise directional CLIP loss proposed by [25], *i.e.*,

$$\begin{aligned} &\mathcal{S} = \mathbf{G}(w), \\ &\{\mathcal{S}^j\}_{j=1}^{N_{\text{crop}}} = \text{aug}(\text{crop}(\mathcal{S})), \\ &\Delta\mathcal{S}^j = E_I(\mathcal{S}^j) - E_I(I_{\text{src}}), \\ &\Delta T = E_T(T_i) - E_T(T_{\text{src}}), \\ &L_{T_i} = \frac{1}{N_{\text{crop}}} \sum_{j=1}^{N_{\text{crop}}} \left(1 - \frac{\Delta\mathcal{S}^j \cdot \Delta T}{\|\Delta\mathcal{S}^j\| \|\Delta T\|}\right), \end{aligned} \tag{1}$$

where $w \in \mathcal{W}^+$ is a vector in StyleGAN3 latent space that we optimize, $\mathcal{S}$ is the style representation generated by $\mathbf{G}$ from w. $\text{aug}(\cdot)$ is the augmentation function, $\text{crop}(\cdot)$ is the patch crop function, and N_{crop} is the number of cropped patches. T_{src} and I_{src} are the source text and source image used to compute CLIP embedding directions, respectively. For simplicity, T_{src} is set to be "a photo" following [25], whereas I_{src} is an arbitrary photo-realistic image.

Eq. 1 effectively measures the style similarity between the input text T_i and the generated style $\mathcal{S}$. However, we want our model to handle style inputs from the image modality as well. To this end, we propose an image-image patch-wise directional CLIP loss as below:

$$\begin{aligned} &\{I_i^k\}_{k=1}^{N_{\text{crop}}} = \text{aug}(\text{crop}(I_i)), \\ &\Delta I_i^k = E_I(I_i^k) - E_I(I_{\text{src}}), \\ &L_{I_i} = \frac{1}{N_{\text{crop}}^2} \sum_{j=1}^{N_{\text{crop}}} \sum_{k=1}^{N_{\text{crop}}} \left(1 - \frac{\Delta\mathcal{S}^j \cdot \Delta I_i^k}{\|\Delta\mathcal{S}^j\| \|\Delta I_i^k\|}\right), \end{aligned} \tag{2}$$

Algorithm 1: Cross-modal GAN Inversion

Data: A set of style images $\{I_i\}_{i=1}^{N_I}$, a set of style text descriptions $\{T_i\}_{i=1}^{N_T}$, and corresponding style weights $\{\alpha_i^I\}_{i=1}^{N_I}$, $\{\alpha_i^T\}_{i=1}^{N_T}$.
Result: The generated style representation $\mathcal{S}$ and its corresponding latent $w^* \in \mathcal{W}^+$.

Randomly initialize w;
repeat
  $\mathcal{S} \leftarrow \mathbf{G}(w)$;
  Run aug and crop on $\mathcal{S}$ and $\{I_i\}_{i=1}^{N_I}$ to obtain $\{\mathcal{S}^j\}_{j=1}^{N_{\text{crop}}}$, $\{I_i^k\}_{i,k=1,1}^{N_i,N_{\text{crop}}}$;
  Calculate L_{sty} using Eq. 1, Eq. 2, and Eq. 3;
  Adam update for w with $\nabla_w L_{\text{sty}}$;
until *L_{sty} is converged*;
return $(\mathcal{S}, w^*)$;

where I_i is the input style image, and $\Delta\mathcal{S}^j$ is calculated using Eq. 1. Specifically, for every image I_i or $\mathcal{S}$ involved in the CLIP embedding computation, we first randomly crop a large number of patches and then augment them. After computing the loss for each patch, we average them together to obtain the final loss value. By computing the averaged cosine similarity between each direction pair, Eq. 2 accurately estimates the style similarity between the input image I_i and the generated style $\mathcal{S}$.

In the general case where multiple style images $\{I_i\}_{i=1}^{N_I}$ and style text descriptions $\{T_i\}_{i=1}^{N_T}$ are given, we calculate the style-specific CLIP loss L_{sty}, and solve the following optimization problem:

$$w^* = \arg\min_{w \in \mathcal{W}^+} L_{\text{sty}} = \arg\min_{w \in \mathcal{W}^+} \sum_{i=1}^{N_I} \alpha_i^I L_{I_i} + \sum_{i=1}^{N_T} \alpha_i^T L_{T_i}, \tag{3}$$

where $\{\alpha_i^I\}_{i=1}^{N_I}$ and $\{\alpha_i^T\}_{i=1}^{N_T}$ are the style weights.

4.2. Inversion Algorithm

With the style-specific CLIP guidance, it is straightforward to run our cross-modal GAN inversion algorithm. As shown in Algorithm 1, after initializing w, we repetitively calculate L_{sty} and use Adam optimizer [23] to update w, until L_{sty} has converged.

Note that in this algorithm, both $\{I_i\}_{i=1}^{N_I}$ and $\{T_i\}_{i=1}^{N_T}$ are optional. If only $\{T_i\}_{i=1}^{N_T}$ is given, it degenerates to text-guided style generation, converting our framework to a method for TIST. Similarly, if only $\{I_i\}_{i=1}^{N_I}$ is given, it degenerates into mixing multiple style images, *i.e.*, generating a mixture of styles represented by multiple input images. When both $\{I_i\}_{i=1}^{N_I}$ and $\{T_i\}_{i=1}^{N_T}$ are provided, cross-modal style interpolation, *e.g.*, interpolating a style between a given text and a given image, can be naturally achieved by adjusting the style weights $\{\alpha_i^I\}_{i=1}^{N_I}$ and $\{\alpha_i^T\}_{i=1}^{N_T}$.

Algorithm 2: Multimodality-guided Style Transfer

Data: Input styles $\{I_i\}_{i=1}^{N_I}$, $\{T_i\}_{i=1}^{N_T}$, style weights $\{\alpha_i^I\}_{i=1}^{N_I}$, $\{\alpha_i^T\}_{i=1}^{N_T}$, and a set of content images $\{C_i\}_{i=1}^{N_C}$.
Result: A set of stylized images $\{Y_i\}_{i=1}^{N_C}$.

if *The aggregated feature F is not cached* **then**
  Run Algorithm 1 $N_{\mathcal{S}}$ times to obtain $\{\mathcal{S}_i\}_{i=1}^{N_{\mathcal{S}}}$;
  $\{F_i\}_{i=1}^{N_C} \leftarrow \mathbf{M}_f(\{\mathcal{S}_i\}_{i=1}^{N_{\mathcal{S}}})$;
  $F \leftarrow \text{aggregate}(\{F_i\}_{i=1}^{N_C})$;
end
$\{Y_i\}_{i=1}^{N_C} \leftarrow \mathbf{M}_t(\{C_i\}_{i=1}^{N_C}, F)$;
return $\{Y_i\}_{i=1}^{N_C}$;

5. Multimodality-guided Image Style Transfer

5.1. Multi-style Boosting

Due to the internal randomness of GAN inversion, one single intermediate style representation may not cover all style patterns specified by the input references, impairing the style transfer quality of final results. To address this problem, we propose a multi-style boosting algorithm. We aim to enrich the intermediate style representations, while keeping them compatible with the adapted IIST model $\mathbf{M}$. Specifically, for each set of style inputs, we run cross-modal GAN inversion multiple times, resulting in a set of style representations $\{\mathcal{S}_i\}_{i=1}^{N_{\mathcal{S}}}$. Then we feed them into $\mathbf{M}$ separately, and aggregate the outputs together to exploit $\{\mathcal{S}_i\}_{i=1}^{N_{\mathcal{S}}}$. The aggregation strategy depends on the specific implementation of $\mathbf{M}$. We describe one instance of $\mathbf{M}$ and the aggregation strategy in the supplementary material.

5.2. Style Transfer Algorithm

We detail our style transfer method in Algorithm 2. For brevity, we assume the adapted IIST model $\mathbf{M}$ can be decomposed into a feature extraction network $\mathbf{M}_f$ and a style transfer module $\mathbf{M}_t$, *i.e.*, $\mathbf{M}(C, \mathcal{S}) = \mathbf{M}_t(C, \mathbf{M}_f(\mathcal{S}))$. Similar to Algorithm 1, Algorithm 2 can be used for TIST by only providing $\{T_i\}_{i=1}^{N_T}$, or MMIST by providing both $\{I_i\}_{i=1}^{N_I}$ and $\{T_i\}_{i=1}^{N_T}$. In particular, if only one T_0 is given, our method degenerates to the IIST approach it generalizes since cross-modal GAN inversion is not necessary in this case. The aggregated style feature F from each unique set of input styles can be cached for later use once it is produced. With cached F, when new content images arrive with the same input styles, only the style transfer part $\mathbf{M}_t$ needs to be executed. Since $\mathbf{M}_t$ is a feed-forward network, our method runs significantly faster than previous meth-

Figure 4. **Comparison with other TIST methods.** Our method delivers more accurate styles than others. Meanwhile, as opposed to other methods which often distort the content or add unreasonable patterns, content information is greatly preserved in results of our method.

ods [7,21,25] that require ad-hoc optimization for each pair of style and content.

6. Experiments

We consider three tasks to evaluate our method: (1) TIST; (2) MMIST with one style image and one style text description; and (3) MMIST with style interpolated from multiple references in different modalities. Note that IIST is trivial and unnecessary to consider because our method degenerates to the IIST approach it includes, as mentioned in Section 5.2. We employ AdaAttN [29] as our adapted IIST method in all our experiments. Please see the supplementary material for more implementation details.

6.1. Comparison with TIST Methods

Qualitative comparison. We conducted a qualitative comparison between our method and several TIST methods. In this task, style transfer is conditioned on a single text description. Therefore, we simply set $N_T = 1$ and $N_I = 0$ in our method. We consider CLIPStyler [25], CLIPStyler-F [25], DiffusionCLIP [21], Null-text Inversion [31], and InstructPix2Pix [4] as our baselines. These methods either exclusively focus on this task [25] and achieve state-of-the-art performance, or treat TIST as one of their practical applications. It is worth noting that CLIPStyler and CLIPStyler-F are proposed in the same paper, and the latter is an extension of the former and achieves TIST in a single forward pass for learned style text descriptions. By caching style representations, the speed of our method is similar to CLIPStyler-F, which is significantly faster than CLIPStyler

Table 2. **Quantitative user study results on four TIST methods.** For each method, we report the number of positive responses received from raters, as well as its percentage (over 161,040 responses in total). Our method outperforms baselines.

Method	Style (%) ↑	Content (%) ↑	Overall (%) ↑
CLIPStyler	10631 (39.6)	9686 (36.1)	8966 (33.4)
CLIPStyler-F	14626 (54.5)	13453 (50.1)	11776 (43.9)
DiffusionCLIP	9768 (36.4)	7929 (29.5)	7395 (27.6)
Null-text Inversion	7375 (27.5)	16876 (62.8)	7166 (26.7)
InstructPix2Pix	10580 (39.4)	17453 (65.0)	8268 (30.8)
Ours	**17151 (63.9)**	**18061 (67.3)**	**15125 (56.4)**

and DiffusionCLIP.

Figure 4 shows comparison results on 9 content-style pairs. We list the style text descriptions and the content image in the first and the second rows, respectively. Style transfer results of different methods are listed in the remaining rows. As is shown, CLIPStyler often breaks or distorts the original contents (1st and 7th columns). For results generated by DiffusionCLIP, their styles do not match the corresponding text descriptions well (1st, 2nd, and 7th columns). In addition, these methods tend to add undesirable local patterns to stylized images (4th and 9th columns). The absence of an original image caption likely explains why Null-text Inversion's outcomes closely mirror the content images, displaying minimal style variations. InstructPix2Pix sometimes compromises the content (3rd column) or introduces inaccurate styles (2nd and 7th columns). In contrast, our method delivers more accurate styles while greatly preserving the content information.

Quantitative user study. We conduct a large-scale quantitative user study to better understand the performance of our method. We still use CLIPStyler [25], CLIPStyler-F [25], DiffusionCLIP [21], Null-text Inversion [31], and InstructPix2Pix [4] as our baselines. We apply 44 distinctive text-described styles to 61 different content images, giving 2,684 stylized images. For each of them, we ask 10 different raters to evaluate it from three aspects: style consistency, content preservation, and overall quality. For each aspect, raters are asked if the stylized image *respects the aspect well* (positive) or *not* (negative). In total we obtain 161,040 responses where each method receives 26,840 responses and each stylized image receives 10 responses. The total number of raters involved is 5,041. We report the number and the percentage of positive responses. Table 2 shows that our method outperforms all baselines in every aspect. Interestingly, CLIPStyler-F received more positive responses than CLIPStyler, although the former is designed to be a fast extension in [25]. This user study result is consistent with the qualitative results shown in Figure 4.

Running speed. Our pre-computed style representations enable *fast stylization* at *test time*. We compare speed of our method with other TIST methods in Table 3. All methods are under their fastest setting and run on an RTX A6000.

Table 3. **Speed comparison with other TIST methods.** Our method is the fastest. ($\star$) Null-text Inversion does not need per-style training but has a per content inverison time of 80.7s.

Method	Per-style Training Time (s) ↓	Stylization FPS ↑
CLIPStyler	0	0.03
CLIPStyler-F	44.1	4.40
DiffusionCLIP	1885	0.05
Null-text Inversion	$\star$	0.07
InstructPix2Pix	0	0.08
Ours	12.6	**6.00**

Table 4. **Ablation study on different design choices.** The performance is evaluated through the user study.

Setting	Preference % ↑	Setting	Preference % ↑
CropSize128	39.5	PatchLoss500	49.9
NoCrop	41.9	PatchLoss2500	48.2
NoAug	47.9	NoBoosting	34.5

6.2. MMIST and Cross-modal Style Interpolation

Qualitative results of MMIST with one style image and one style text description. To the best of our knowledge, MMIST is infeasible for all the existing methods. However, under our unified framework, it can be easily performed by providing style references from more than one modality. We first consider the case where input style is specified through one style image and one style text description. Figure 5 shows MMIST results of our method on 8 different text-image styles and 4 content images. Our method can successfully summarize styles from a pair of style image and text description, and apply it to various contents. Moreover, contents are consistently preserved in all stylized images.

Qualitative results of MMIST with style interpolated from multiple references in different modalities. As mentioned in Section 5.2, our method is able to handle arbitrary number of style inputs from different modalities. To finer control the mixture degree of two or more given styles, we interpolate the styles with various ratios. This is done by setting different $\{\alpha_i^I\}_{i=1}^{N_I}$ and $\{\alpha_i^T\}_{i=1}^{N_T}$, while keeping their summation fixed. Note that under our unified framework, style interpolation between any number or kinds of modalities can be achieved in the same way. Figures 1 and 6 demonstrate the style interpolation results defined by different mixtures of style images and text descriptions. Our method produces decent and reasonable stylized images. More results can be found in the supplementary material.

6.3. Ablation Study

We investigate a few design choices of our method and quantitatively measure their effects in practice. To this end, we consider the TIST task. In our experiment, we randomly pair the style text descriptions and content images, and obtain 1,012 pairs for stylized image generation. For each

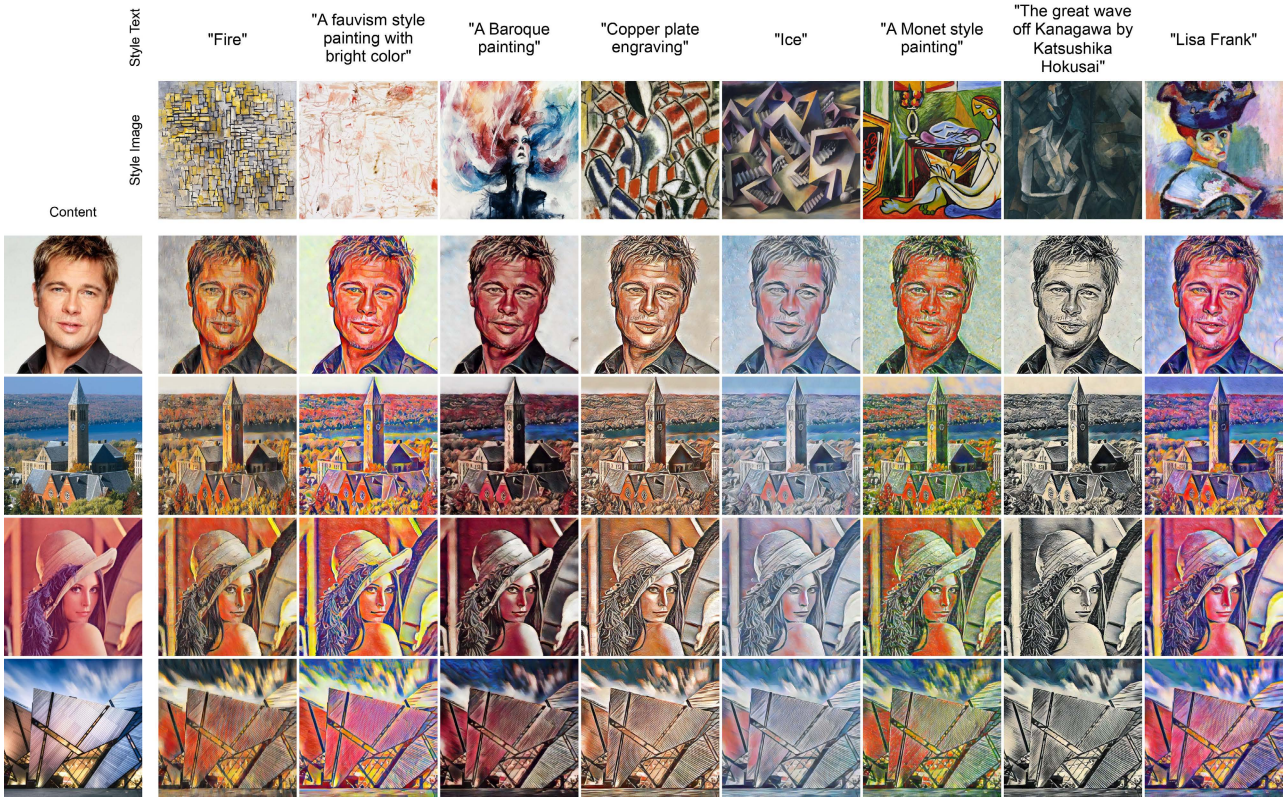

Figure 5. **MMIST results.** The first and second rows show style text descriptions and style images, respectively. The first column shows content images. Our method successfully mixes multimodal styles and applies them to various content images.

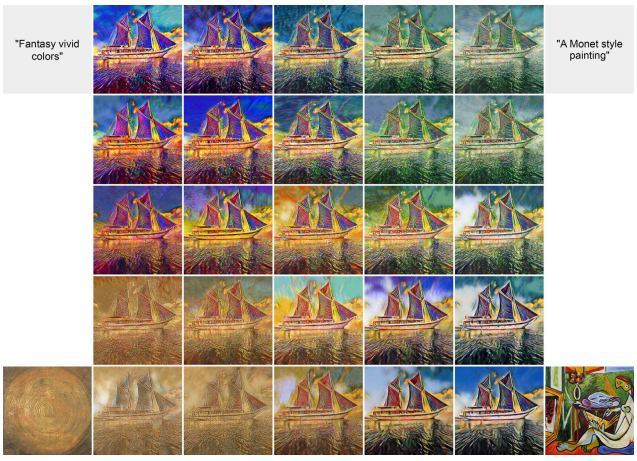

Figure 6. **MMIST and Style interpolation with four image and text styles.** Please zoom in to see the details.

specific pair, 10 raters are asked to pick their preferred stylized image between the one generated by a certain design choice and the one generated by our full method. We report the user preference percentage for all design choices in Table 4. In the table, CropSize128 means we crop 128×128 patches when calculating style-specific CLIP loss. NoCrop means we do not crop or augment the image, and directly apply the loss. NoAug means we do not apply augmentation to the cropped patches. PatchLoss500 and PatchLoss2500 mean we set $\alpha_0^T = 500$ and $\alpha_0^T = 2500$, respectively. NoBoosting means we do not use multi-style boosting when applying the style to contents. We observe that proper crop size and multi-style boosting are critical to the good performance of our method, whereas the effects of augmentation and style weight are relatively minor to our method. Changing the style weight from 500 to 2500 almost does not affect the performance, indicating that our method is quite robust to this hyperparameter. Qualitative ablation study results are available in the supplementary material.

7. Conclusions and Future Work

In this paper, we present a unified style transfer framework to transfer styles defined by multiple modalities. The proposed cross-modal GAN inversion enables our framework to combine different styles and faithfully transfer them to arbitrary images. Extensive experiments demonstrate that our method achieves SOTA performance on TIST. In addition, the proposed method handles the new MMIST problem and cross-modal style interpolation task effectively.

While our work only considers style information from image and text, there is no theoretical restriction on our method to obtain styles from other modalities, *e.g.*, audio. We leave the exploration of this direction as future work.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 3, 4

[2] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18511–18521, 2022. 3

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 3

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 6, 7

[5] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1897–1906, 2017. 3

[6] Haibo Chen, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, Dongming Lu, et al. Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems*, 34:26561–26573, 2021. 2, 3

[7] 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. *arXiv preprint arXiv:2204.08583*, 2, 2022. 3, 4, 6

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[9] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE international conference on computer vision*, pages 5706–5714, 2017. 3

[10] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 3

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[12] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021. 2, 3, 4

[13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 1, 2

[14] Kibeom Hong, Seogkyu Jeon, Huan Yang, Jianlong Fu, and Hyeran Byun. Domain-aware universal style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14609–14617, 2021. 3

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2, 3

[16] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. Dynamic instance normalization for arbitrary style transfer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4369–4376, Apr. 2020. 2, 3

[17] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 2, 3

[18] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 3, 4

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[21] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 2, 3, 4, 6, 7

[22] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 3

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[24] Dmytro Kotovenko, Matthias Wright, Arthur Heimbrecht, and Bjorn Ommer. Rethinking style transfer: From pixels to parameterized brushstrokes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12196–12205, 2021. 3

[25] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 2, 3, 4, 6, 7

[26] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020. 3

[27] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3809–3817, 2019. 2, 3

[28] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Diversified texture synthesis with feed-forward networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3920–3928, 2017. 3

[29] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6649–6658, 2021. 2, 3, 6

[30] Zhi-Song Liu, Li-Wen Wang, Wan-Chi Siu, and Vicky Kalogeiton. Name your style: An arbitrary artist-aware image style transfer. *arXiv preprint arXiv:2202.13562*, 2022. 3

[31] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 6, 7

[32] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. *Advances in neural information processing systems*, 31, 2018. 3

[33] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5880–5888, 2019. 2, 3

[34] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2, 3, 4

[35] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[37] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Bjorn Ommer. A style-aware content loss for real-time hd style transfer. In *proceedings of the European conference on computer vision (ECCV)*, pages 698–714, 2018. 2

[38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[39] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. *arXiv preprint arXiv:1603.03417*, 2016. 3

[40] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 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*, pages 6924–6932, 2017. 3

[41] Xin Wang, Geoffrey Oxholm, Da Zhang, and Yuan-Fang Wang. Multimodal transfer: A hierarchical deep convolutional neural network for fast artistic style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5239–5247, 2017. 3

[42] Zijie Wu, Zhen Zhu, Junping Du, and Xiang Bai. Ccpl: contrastive coherence preserving loss for versatile style transfer. In *European Conference on Computer Vision*, pages 189–206. Springer, 2022. 2

[43] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021. 3

[44] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 3

[45] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 3

[46] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018. 3

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3

[48] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH*, pages 1–8, 2022. 2

[49] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016. 3