This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Style-Editor: Text-driven object-centric style editing

Jihun Park*, Jongmin Gim*, Kyoungmin Lee*, Seunghun Lee, Sunghoon Im†
DGIST, Daegu, Republic of Korea
{pjh2857, jongmin4422, kyoungmin, lsh5688, sunghoonim}@dgist.ac.kr

Figure 1. Results of our Style-Editor under diverse textual conditions. To the left and right of the arrow(→) indicate the source text T^{src} and style text T^{sty}, respectively.

Abstract

We present Text-driven object-centric style editing model named Style-Editor, a novel method that guides style editing at an object-centric level using textual inputs. The core of Style-Editor is our Patch-wise Co-Directional (PCD) loss, meticulously designed for precise object-centric editing that are closely aligned with the input text. This loss combines a patch directional loss for text-guided style direction and a patch distribution consistency loss for even CLIP embedding distribution across object regions. It ensures a seamless and harmonious style editing across object regions. Key to our method are the Text-Matched Patch Selection (TMPS) and Pre-fixed Region Selection (PRS) modules for identifying object locations via text, eliminating the need for segmentation masks. Lastly, we introduce an Adaptive Background Preservation (ABP) loss to maintain the original style and structural essence of the image's background. This loss is applied to dynamically identified background areas. Extensive experiments underline the effectiveness of our approach in creating visually coherent and textually aligned style editing.

1. Introduction

In the realm of creative digital industries such as advertising, film, and video game development, the demand for advanced image manipulation is surging. The introduction of an object-focused style editing model, driven by textual commands, is transforming these sectors. It allows for detailed and user-friendly adjustments to the visual aspects of objects in images. This innovation empowers designers to bypass traditional manual editing, enabling them to define stylistic alterations using just text, thus facilitating

* Equal Contribution
†Corresponding author

Figure 2. Our editing results in industrial applications.

rapid concept development and customization. Consider the ease with which digital fashion elements can be modified, car hues altered for online showrooms, or furniture designs changed in virtual environments, all through simple text instructions, as illustrated in Fig. 2.

Style editing began with the foundational concept of style transfer, which traditionally relies on reference images to guide the transformation process, as demonstrated by seminal works [6, 12, 17, 28, 35, 57, 58]. While these methods have been foundational, their dependency on visual templates can constrain creative possibilities. In contrast, a burgeoning wave of research in text-guided style transfer [2, 15, 27, 29, 34, 66] is redefining the landscape. By dispensing with the need for reference images, these novel approaches broaden the horizons for stylistic editing, harnessing the power of textual descriptions to steer the creative process. The pioneering approach [2, 4, 40, 43, 56] empowers object-centric editing directly steered by textual descriptions. Yet, this method also presents certain drawbacks; it risks altering the original content's fidelity, may fail to accurately capture the intended textual descriptions, or may unintentionally edit undesired parts of the image.

To address this issue, we introduce Text-driven object-centric style editing model named Style-Editor, a novel approach aimed at transforming the appearance of objects based on textual descriptions. This approach is designed to retain the structural integrity of objects and preserve both the appearance and structure of the background. At the core of Style-Editor lie three pivotal components: the Patch-wise Co-Directional (PCD) loss, the Adaptive Background Preservation (ABP) loss, and the Text-Matched Patch Selection (TMPS) with Pre-fixed Region Selection (PRS) module.

Leveraging the robust zero-shot image classification capabilities of the CLIP model, the TMPS and PRS modules are designed to identify and style the locations of objects related to the text. Unlike traditional directional loss [16, 34, 63] that may cause distortion due to a focus on vector directions, our PCD loss, augmented by the TMPS module, directs style editing in foreground regions aligned with the source text. This ensures a consistent CLIP-embedding distribution across patches from both source and stylized images, promoting a cohesive style editing within object areas. Our ABP loss is designed to preserve the style and structure of the background in the source image, specifically in non-object areas. It adaptively targets and applies the loss to dynamically detected background regions, ensuring a natural and seamless transition in object-centric style editing and blending stylized elements with their surroundings as shown in Fig. 2.

In summary, our primary contributions include:

- We present the Text-Matched Patch Selection (TMPS) and Pre-fixed Region Selection (PRS), which identify regions of objects related to text for object-centric style editing.
- We propose the Patch-wise Co-Directional (PCD) loss to enable precise style editing on targeted objects, maintaining the integrity and coherence of the source's visual aesthetic.
- We introduce the Adaptive Background Preservation (ABP) loss, effectively maintaining the original style and structure of designated background areas.

2. Related Works

2.1. Style Transfer

Neural Style Transfer (NST) represents a significant advancement in the domain of image stylization, introduced by [17]. This pioneering approach utilized a pre-trained Convolutional Neural Network (CNN), particularly VGGNet, to extract distinct content and style features. However, a notable constraint of NST lies in its substantial computational demand, stemming from the per-image optimization approach. To overcome this, [22] introduced Adaptive Instance Normalization (AdaIN), a technique that aligns the mean and variance of source image features with those of the style image. Building on this, [36, 37] proposed the Whitening and Coloring Transform (WCT), which aligns the entire covariance matrix of the features, resulting in more refined and superior stylization outcomes. With the advent of attention mechanisms in neural networks [11, 59], new style transfer models have emerged that utilize these mechanisms to achieve impressive results [8, 20, 39, 42, 64], reflecting the ongoing evolution of style transfer technology.

2.2. Text-Guided Image Synthesis

The Contrastive Language-Image Pretraining (CLIP) model, trained on a large-scale image-text dataset [49], has

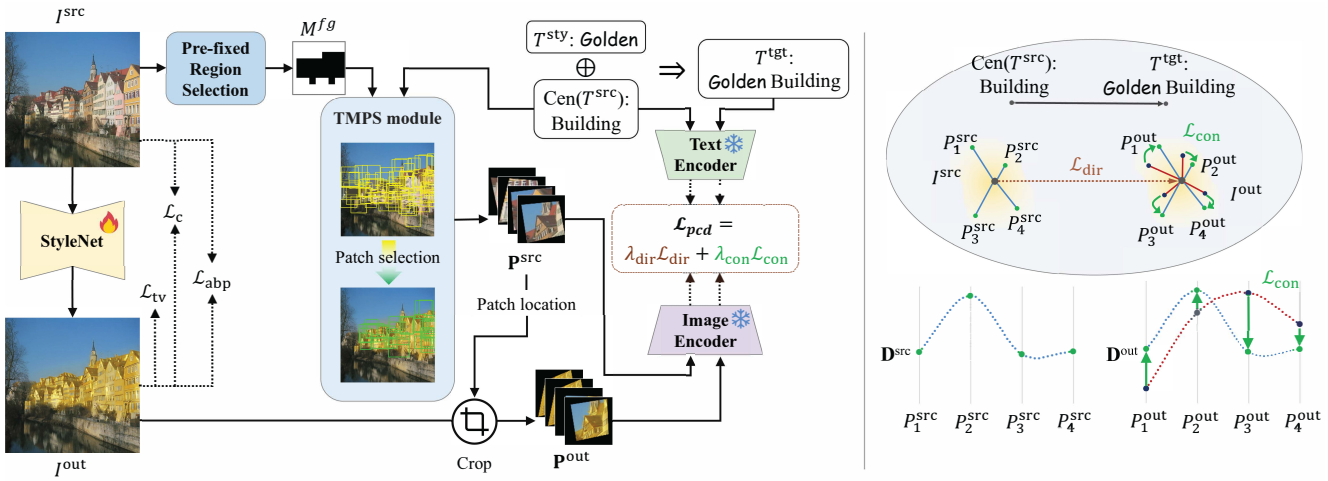

Figure 3. **(Left)** Overall pipeline of our Style-Editor consisting of a style editing network (StyleNet), Pre-fixed Region Selection (PRS), Text-Matched Patch Selection (TMPS) module and pretrained CLIP encoders. The StyleNet takes a source image I^{src} and generates an object-wise stylized image I^{out}. The TMPS module is responsible for pinpointing patches that most closely correspond to T^{src} from the foreground regions identified by the PRS. The selected and augmented patches, $\mathbf{P}^{\text{src}}, \mathbf{P}^{\text{out}}$, are then aligned with $T^{\text{src}}, T^{\text{tgt}}$ in the CLIP embedding space using Patch-wise Co-Directional (PCD) loss $\mathcal{L}_{\text{pcd}}$. The target text T^{tgt} is derived by central word selection. Additionally, we apply a content loss $\mathcal{L}_{\text{c}}$, an Adaptive Background Preservation (ABP) loss $\mathcal{L}_{\text{abp}}$ to enhance object-centric style editing, along with a total variance loss $\mathcal{L}_{\text{tv}}$ for regularization. **(Right)** Illustration of the functionality of the PCD loss $\mathcal{L}_{\text{pcd}}$ in feature space. It is composed of a patch-wise directional loss $\mathcal{L}_{\text{dir}}$ and a patch distribution consistency loss $\mathcal{L}_{\text{con}}$.

significantly impacted image synthesis applications [14, 18, 47, 55]. Research [16, 45] building upon the StyleGAN architecture [30, 31] has shown how text descriptors can adapt the style of source images. For instance, the directional loss introduced in [16] showcases remarkable stability and diversity in this domain, overcoming the previous requirement of a style image for transfer. Parallel to these developments, the field has seen a surge in research focusing on diffusion model [19].

Recent works [7, 41, 53, 54, 63] are exploring new frontiers in image editing and generation, using text-driven diffusion models. Despite these advancements, a persisting challenge for both Generative Adversarial Networks (GANs) and diffusion models is maintaining the subject's identity consistently in the generated images.

2.3. Text-driven Object-Centric Style Editing

Style editing modifies the color or texture of the original image to match a desired style while preserving the image's content [24]. Early methods [25, 26, 33] utilized reference images for style information, applying style editing to targeted areas through segmentation networks or clustering methods. StyleGAN-NADA [16] introduced a paradigm shift by eliminating the need for reference images and instead leveraging textual directions in the CLIP space for domain transfer. Building on this, CLIPstyler [34] successfully applied this concept in text-driven style editing, advancing this field. Following this trajectory, [2] proposed a text-driven object-centric style editing approach by extracting a relevancy map [5].

Another approach to text-driven object-centric style editing involves diffusion models trained on large vision-text datasets, such as [50, 54]. Within these approaches, some involve training the model on newly generated text-image pairs [4] or optimizing null-text [40]. Additionally, methods [3, 43, 46, 56] have been developed to edit feature maps of the diffusion model, enabling more precise control over style editing based on specific textual guidance. Despite these diverse advancements, a notable limitation remains: these techniques often lead to unintended alterations of both style and content, and they often fail to achieve high levels of style editing fidelity.

3. Method

3.1. Overview Framework

The overall pipeline of our model is illustrated in Fig. 3-(Left). We train a style editing network (StyleNet) to generate a stylized image I^{out}, given a source image I^{src}, along with accompanying source text T^{src}, and a style text T^{sty}. We use the text encoder and the image encoder derived from the pre-trained CLIP model, freezing their parameters during the training process. The training of StyleNet incorporates a composite loss function consisting of four distinct components: Patch-wise Co-Directional loss ($\mathcal{L}_{\text{pcd}}$), Adaptive Background Preservation loss ($\mathcal{L}_{\text{abp}}$), a content loss

ALGORITHM 1: Text-Matched Patch Selection

Input: Image patch set $\mathbf{P}$ and source text T^{src}
Output: Patches corresponding to the source text $\mathbf{P}^{\text{sel}}$
Parameter: K: # of patches in $\mathbf{P}$, M: # of patches similar to source text

$\mathbf{S}, \hat{\mathbf{S}} \leftarrow \emptyset, \emptyset$
for $i = 1$ to K **do**
  $f_i \leftarrow E_I(P_i)$
  $s_i \leftarrow \frac{f_i \cdot E_T(T^{\text{src}})}{\|f_i\| \cdot \|E_T(T^{\text{src}})\|}$
  $\mathbf{S} \leftarrow \mathbf{S} \cup \{s_i\}$
end for
$\mathbf{I} \leftarrow \{i \mid s_i \geq$ the M^{th} largest value in $\mathbf{S}\}$
$f_{\text{avg}} \leftarrow \frac{1}{|\mathbf{I}|} \sum_{i \in \mathbf{I}} f_i$
for $j = 1$ to K **do**
  $\hat{s}_j \leftarrow \frac{f_j \cdot f_{\text{avg}}}{\|f_j\| \cdot \|f_{\text{avg}}\|}$
  $\hat{\mathbf{S}} \leftarrow \hat{\mathbf{S}} \cup \{\hat{s}_j\}$
end for
$\hat{s}_k \leftarrow$ the $(\text{round}(\frac{K}{2}))^{\text{th}}$ largest value from $\hat{\mathbf{S}}$
$\mathbf{J} \leftarrow \{j \mid \hat{s}_j \geq \hat{s}_k$ and $\hat{s}_j > 0.8\}$
return $\mathbf{P}^{\text{sel}} \leftarrow \{P_j \mid j \in \mathbf{J}\}$

ALGORITHM 2: Pre-fixed Region Selection

Input: Source image I^{src} and source text T^{src}
Output: A binary mask containing objects corresponding to the source text M^{fg}
Parameter: τ: Threshold for # of selections.

Divide I^{src} into L uniform square grids.
Generate a set of three distinct-sized patches per grid: $\mathbf{P}^{\text{grid}} = \{P_1^{\text{grid}}, \ldots, P_{3L}^{\text{grid}}\}$.
Obtain selected patches $\mathbf{P}^{\text{grid\_sel}}$ using TMPS module: $\mathbf{P}^{\text{grid\_sel}} = \textit{TMPS}(\mathbf{P}^{\text{grid}}, T^{\text{src}})$.
Initialize a voting matrix $V \in \mathrm{R}^{H \times W}$ with all elements set to zero.
for each pixel in the selected patches $\mathbf{P}^{\text{grid\_sel}}$ **do**
  Increment the corresponding element in V.
end for
Determine the pre-fixed foreground region M^{fg}:

$$M^{\text{fg}}(i,j) = \begin{cases} 1, & \text{if } V(i,j) \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

return M^{fg}

($\mathcal{L}_{\text{c}}$), and a total variation regularization loss ($\mathcal{L}_{\text{tv}}$). These are weighted by the balance terms λ_{abp}, λ_{c} and λ_{tv} as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pcd}} + \lambda_{\text{abp}}\mathcal{L}_{\text{abp}} + \lambda_{\text{c}}\mathcal{L}_{\text{c}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}}. \quad (1)$$

This paper primarily delves into the Patch-wise Co-Directional loss $\mathcal{L}_{\text{pcd}}$ in Sec. 3.3, and the Adaptive Background Preservation loss $\mathcal{L}_{\text{abp}}$ in Sec. 3.4. The content loss $\mathcal{L}_{\text{c}}$ imposes the mean-square error between the features of an input and a stylized image extracted from the pre-trained VGG-19 networks [17]. The total variation loss $\mathcal{L}_{\text{tv}}$ is a loss function designed based on the noise removal method [52].

3.2. Text-Matched Patch Selection (TMPS)

A critical aspect of object-centric style editing is the accurate identification of the object's location within an image. To address this challenge, we introduce the TMPS module. This module pinpoints and selects image patches that correlate to a specified object as described by the source text, leveraging the zero-shot image classification capability of the CLIP model. Utilizing the image encoder E_I and the text encoder E_T from the CLIP architecture, TMPS establishes a strong link between the object and its textual descriptor. We define a representative feature vector f_{avg} in Alg. 1, capturing the essential characteristics of the object in the source image. Then, we identify patches $\mathbf{P}^{\text{sel}}$ with features similar to this representative vector. The detailed mechanism of TMPS is explained in Alg. 1.

To streamline the search process for TMPS during the object-centric style editing, we introduce the Pre-fixed Region Selection (PRS) module within the source image. Applied during the initial iterations, this module delineates a foreground region M^{fg}, configured to coarsely isolate object areas as outlined in Alg. 2. This early demarcation of the object's location allows for generating patches specifically within the foreground region (M^{fg}) in later iterations. This strategy enhances the precision and efficiency of the style editing, focusing the modification on the most relevant sections of the image.

3.3. Patch-Wise Co-Directional Loss (PCD)

Our PCD loss $\mathcal{L}_{\text{pcd}}$ incorporates a patch-wise directional loss $\mathcal{L}_{\text{dir}}$ and a patch distribution consistency loss $\mathcal{L}_{\text{con}}$ with balance terms λ_{dir} and λ_{con} as follows:

$$\mathcal{L}_{\text{pcd}} = \lambda_{\text{dir}}\mathcal{L}_{\text{dir}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}. \quad (2)$$

Patch-wise directional loss The foundational concept of directional loss, as initially introduced in [16] and [34], is further refined in our approach. We adapt and advance this concept specifically for object-centric style editing, with a targeted application on image patches that correspond to the source text T^{src}. We commence by randomly selecting a subset of patches within the foreground region M^{fg} of the source image. The selected patches, represented as $\mathbf{P}^{\text{src}} \in \{P_1^{\text{src}}, ..., P_N^{\text{src}}\}$, are then extracted using our TMPS, as detailed in Alg. 1. Given the patches and the texts, the

patch-wise directional loss $\mathcal{L}_{\text{dir}}$ is defined as follows:

$$\mathcal{L}_{\text{dir}} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{\Delta P_i \cdot \Delta T}{|\Delta P_i||\Delta T|}\right),$$
$$\Delta P_i = E_I\left(\text{aug}\left(P_i^{\text{out}}\right)\right) - E_I\left(\text{aug}\left(P_i^{\text{src}}\right)\right),$$
$$\Delta T = E_T\left(T^{\text{tgt}}\right) - E_T\left(T^{\text{src}}\right),\ T^{\text{tgt}} = T^{\text{sty}} \oplus \text{Cen}(T^{\text{src}}), \tag{3}$$

where the operator $\oplus$ denotes text combination. We utilize the function $\text{Cen}(\cdot)$ for central word selection and $\text{aug}(\cdot)$ for patch augmentation. The patches P_i^{out} are derived from output stylized images I^{out} using the TMPS algorithm in Alg. 1.

Target text generation To formulate the target text T^{tgt}, we apply central word selection technique $\text{Cen}(\cdot)$ in (3) leveraging Spacy [21], and merge the source text T^{src} and the style text T^{sty}. This design is rooted in the manifold augmentation technique proposed in [65] leveraging text embeddings from models like BERT [9], GPT [48], and CLIP. This amalgamation of texts aligns more closely with the user's intended styling objectives. For instance, to adapt the style of 'red apple' to appear green, we engage TMPS using 'red apple' but modify the target text to 'green apple', with an emphasis on 'apple' as the central word. This refined approach to text input in TMPS enables more accurate patch selection, facilitating precise object-centric style editing.

Patch distribution consistency loss Traditional directional loss functions typically emphasize the direction of vectors while often neglecting their semantic information. This emphasis can inadvertently lead to a misalignment between patches in the source image and those in the stylized image as depicted in Fig. 3-(Right). Such a discrepancy can lead to the collapse of semantic information, resulting in a distorted transformation that causes the loss of information from the source image, contradicting the fundamental aim of style editing. To address this issue, we design the patch distribution consistency loss $\mathcal{L}_{\text{con}}$ as follows:

$$\mathcal{L}_{\text{con}} = \text{JSD}(\mathbf{D}^{\text{src}}, \mathbf{D}^{\text{out}}),$$
$$\mathbf{D}^{\text{src}} = \frac{\hat{\mathbf{D}}^{\text{src}}}{\sum_{i=1}^{N}\hat{\mathbf{D}}_i^{\text{src}}},\ \hat{\mathbf{D}}^{\text{src}} = \left(\frac{E_I\left(P_i^{\text{src}}\right)\cdot E_I(I^{\text{src}})}{|E_I\left(P_i^{\text{src}}\right)|\cdot|E_I(I^{\text{src}})|}\right)_{i=1}^{N},$$
$$\mathbf{D}^{\text{out}} = \frac{\hat{\mathbf{D}}^{\text{out}}}{\sum_{i=1}^{N}\hat{\mathbf{D}}_i^{\text{out}}},\ \hat{\mathbf{D}}^{\text{out}} = \left(\frac{E_I\left(P_i^{\text{out}}\right)\cdot E_I(I^{\text{out}})}{|E_I\left(P_i^{\text{out}}\right)|\cdot|E_I(I^{\text{out}})|}\right)_{i=1}^{N}, \tag{4}$$

where $\text{JSD}(\cdot)$ is Jensen–Shannon divergence, which is employed to align the feature distributions of patches $\mathbf{D}^{\text{src}}$ from the source image with those $\mathbf{D}^{\text{out}}$ from the stylized image.

3.4. Adaptive Background Preservation (ABP) Loss

We introduce the ABP loss, designed to ensure the non-style editing of the background region. Our approach begins by identifying foreground regions as outlined in Alg. 1, Alg. 2. In each iteration, the adaptive foreground mask M^{fg^*} and background mask M^{bg^*} are dynamically updated as follows:

$$M^{\text{bg}^*} = 1 - M^{\text{fg}^*},\ M^{\text{fg}^*} = \left(\bigvee_{i=1}^{N_{\text{iter}}} M_i^{\text{src}}\right), \tag{5}$$

where N_{iter} represents the number of patches in each iteration, and the binary mask M_i^{src} is assigned a value of one over the area of patches $\{P_i^{\text{src}}\}_{i=1,\ldots N_{\text{iter}}}$, which are selected from the source image through the TMPS.

Next, we apply the MS-SSIM and L1 loss functions to ensure that the original styles of background regions are retained as follows:

$$\mathcal{L}_{\text{abp}} = \mathcal{L}_{\text{MS\_SSIM}}(I^{\text{out}} \odot M^{\text{bg}^*}, I^{\text{src}} \odot M^{\text{bg}^*}) + \mathcal{L}_{\text{L1}}(I^{\text{out}} \odot M^{\text{bg}^*}, I^{\text{src}} \odot M^{\text{bg}^*}). \tag{6}$$

This loss enables a well-balanced integration of the stylized and original regions within the image. Specifically, while the foreground areas undergo style editing, the background's integrity is preserved.

4. Experiments

4.1. Implementation Details

We employ the pretrained CLIP model on VIT-B/32 [11]. Our stylization network is based on the U-net architecture [51] consisting of three downsample and three upsample layers featuring channel sizes of 16, 32, and 64. The source images have a resolution of 512×512 pixels. We set λ_{dir}, λ_{con}, λ_{abp}, λ_{c}, and λ_{tv} to 1.5×10^4, 3×10^4, 3×10^4, 4×10^2, and 2×10^{-3}, respectively. For content loss, we follow [17] by utilizing features from the "conv4_2" and "conv5_2" layers.

We commence our training with an initial set of 20 early iterations and completed a total of 200 iterations using the Adam optimizer [32]. The training began with an initial learning rate of 5×10^{-4}, halved after the first 100 iterations. We incorporate the perspective augmentation function from the PyTorch library [44]. Notably, our method processes each source image independently, using only the source image, source text, and style text, without requiring additional training data. The average training time for each text prompt is approximately 45 seconds on an NVIDIA A6000 GPU. For further details on our training procedure, please refer to Appendix A.

4.2. Evaluation metrics

We conduct a comprehensive quantitative evaluation using a variety of metrics that are well-established in the field of text-driven style editing as outlined in Tab. 1. The similarity metric (Sim), a basic metric in the style editing field, measures the similarity between image and text embeddings.

Methods	Foreground quality metrics		Background quality metrics					
	$\text{Sim}_F \uparrow$	$\text{Con}_F \downarrow$	$\text{L1}_B \downarrow$	$\text{Con}_B \downarrow$	$\text{Sty}_B \downarrow$	$\text{SSIM}_B \uparrow$	$\text{DISTS}_B \downarrow$	$\text{PSNR}_B \uparrow$
FlexIT [7]	0.24	8.08	0.25	4.78	0.73	0.60	0.17	19.69
ZeCon [63]	0.23	7.49	0.28	5.33	0.63	0.64	0.24	19.35
Null-text inversion [40]	0.20	4.22	0.16	2.58	0.22	0.74	0.10	23.48
Instruct Pix2Pix [4]	0.22	7.42	0.44	4.66	0.97	0.62	0.20	17.25
Plug and Play [56]	0.23	6.51	0.33	4.26	0.63	0.63	0.22	18.26
pix2pix-zero [43]	0.22	8.12	0.40	4.91	1.07	0.61	0.26	17.11
LEDITS++ [3]	0.22	6.81	0.18	2.92	0.44	0.74	0.14	21.66
local-prompt-mixing [46]	0.20	9.65	0.24	4.65	0.61	0.67	0.18	20.01
StylerDALLE [62]	0.26	8.35	0.51	5.71	1.22	0.52	0.28	15.42
CLIPstyler [34]	0.28	5.16	0.66	3.27	0.35	0.51	0.25	13.20
Text2LIVE [2]	<u>0.32</u>	<u>4.13</u>	<u>0.14</u>	<u>1.22</u>	<u>0.18</u>	<u>0.87</u>	<u>0.09</u>	<u>24.69</u>
Ours	**0.33**	**3.75**	**0.10**	**1.15**	**0.10**	**0.90**	**0.07**	**27.65**

Table 1. Quantitative comparison of our method with the recent text-guided style editing methods using foreground quality metrics and background quality metrics. The symbols ↑ and ↓ indicate higher values are better and lower values are better, respectively.

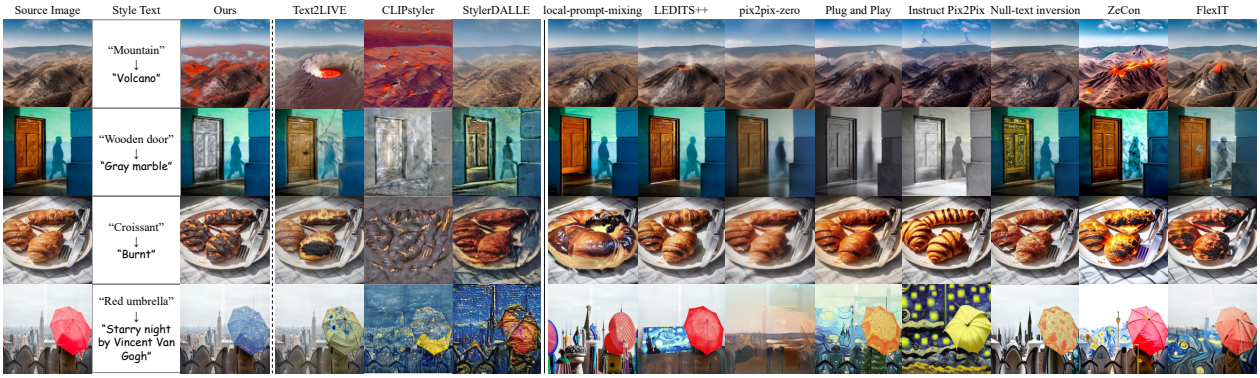

Figure 4. Comparison of our method with various text-guided style editing models. To the left of the solid line are the qualitative results of our model and non-diffusion based models, to the right are the results from diffusion-based methods.

We also apply content metrics (Con) using the VGG content loss [28] and style metrics (Sty) using the mean-variance style loss [23]. These metrics are adapted from traditional calculations but specifically mask either the foreground or background of input images to focus analysis on selected areas. To thoroughly evaluate the quality of the stylized image backgrounds from multiple aspects, we conduct additional assessments specifically targeting the background masks. For this purpose, we utilize the SSIM metric [60], which evaluates structural, luminance, and color attributes. Additionally, we employ the DISTS measure [10] to ensure deep structural and textural consistency in these areas. Furthermore, we use the PSNR [13] metric and absolute difference (L1) for an overall assessment of signal fidelity and basic visual quality in the backgrounds. For a comprehensive quantitative evaluation of object-centric style editing, we split the image region into foreground and background components. We utilize Ground Truth (GT) segmentation (binary) masks for the object $M^{\text{fg-gt}}$ and background $M^{\text{bg-gt}}(= 1 - M^{\text{fg-gt}})$. For example, the Sim_F and L1_B are formulated as follows:

$$\begin{aligned} \text{Sim}_F &= \frac{E_T\left(T_i^{\text{tgt}}\right) \cdot E_I(I^{\text{out}} \odot M^{\text{fg-gt}})}{|E_T\left(T_i^{\text{tgt}}\right)| \cdot |E_I(I^{\text{out}} \odot M^{\text{fg-gt}})|}, \\ \text{L1}_B &= \frac{\sum |I^{\text{src}} \odot M^{\text{bg-gt}} - I^{\text{out}} \odot M^{\text{bg-gt}}|}{255 \cdot \text{num}(I^{\text{src}} \odot M^{\text{bg-gt}})}, \end{aligned} \tag{7}$$

where E_I and E_T denote the pre-trained CLIP image and text encoders, respectively. $\text{num}(\cdot)$ indicates the number of pixels in the image. We randomly selected 16 images from the MSCOCO 2017 dataset [38], each accompanied by GT segmentation masks, to conduct evaluation. Each image was paired with a total of 10 style text descriptions based on three different categories: color (red, blue, green), texture (golden, frosted, stained glass, neon light), and artistic style (Starry Night by Vincent van Gogh, a watercolor

#	Components $\mathcal{L}_{\text{dir}}$	$\mathcal{L}_{\text{con}}$	$\mathcal{L}_{\text{abp}}$	Foreground quality metrics $\text{Sim}_F \uparrow$	$\text{Con}_F \downarrow$	Background quality metrics $\text{L1}_B \downarrow$	$\text{Con}_B \downarrow$	$\text{Sty}_B \downarrow$	$\text{SSIM}_B \uparrow$	$\text{DISTS}_B \downarrow$	$\text{PSNR}_B \uparrow$
(a)				0.29	4.31	0.60	2.72	0.25	0.56	0.22	14.16
(b)	✓			0.32	4.72	0.49	2.07	0.21	0.64	0.15	16.02
(c)	✓	✓		**0.33**	4.62	0.48	2.00	0.21	0.65	0.15	16.16
(d)	✓		✓	0.32	4.16	**0.10**	1.28	0.12	0.89	0.08	27.28
(e)	✓	✓	✓	**0.33**	**3.75**	**0.10**	**1.15**	**0.10**	**0.90**	**0.07**	**27.65**

Table 2. Ablation study examining the effects of the proposed losses $\mathcal{L}_{\text{dir}}$, $\mathcal{L}_{\text{con}}$, and $\mathcal{L}_{\text{abp}}$. The quantitative results of (a)-(e) correspond to the qualitative results shown in Fig. 5. The symbols ↑ and ↓ indicate higher values are better and lower values are better, respectively.

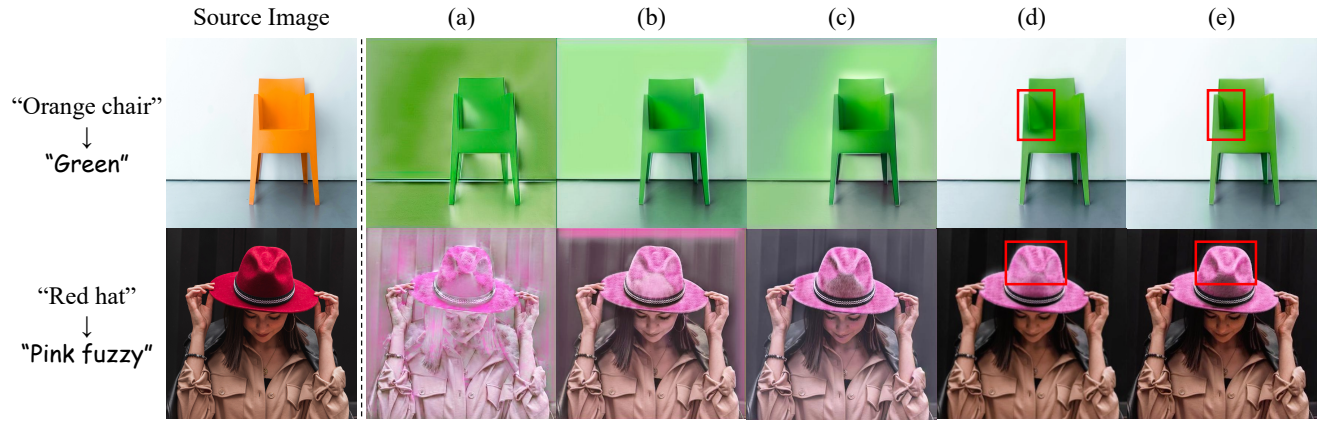

Figure 5. Qualitative results showcasing the impact of applying/omitting the proposed losses. Configurations (a)-(e) correspond to the settings detailed in Tab. 2.

painting of flowers, cubism style). This resulted in a total of 160 stylized images for evaluation.

4.3. Comparison to state-of-the-arts

In Tab. 1 and Fig. 4, we present both quantitative and qualitative comparisons of our Style-Editor model against various text-driven editing models including various diffusion models, such as Text2LIVE [2], CLIPstyler [34],Null-text inversion [40], ZeCon [63], FlexIT [7], StylerDALLE [62], Instruct Pix2Pix [4], Plug and Play [56], pix2pix-zero [43], LEDITS++ [3] and local-prompt-mixing [46]. Our quantitative evaluation indicates that our method surpasses all compared models in performance. Specifically, the object regions in stylized images from our method exhibit higher Sim_F and lower Con_F scores compared to competitive methods. This indicates that our proposed method effectively styles the object regions according to the corresponding text while preserving the structure of the source image. In terms of background quality, the metrics show that the proposed method subtly changes the style of background regions in the source image, as indicated by Sty_B. Meanwhile, it maintains the integrity of the source image's structure, as evidenced by scores from L1_B, Con_B, SSIM_B, DISTS_B, and PSNR_B. This dual capability highlights our method's adeptness at enhancing foreground elements distinctly from the background, ensuring a balanced and coherent visual output.

In the qualitative evaluation, the first row of Fig. 4 showcases the style editing of a mountain into a volcano using the style text "Volcano". Our method adeptly preserves the inherent structure of the mountain while seamlessly integrating the volcanic style. In contrast, Text2LIVE often reconstructs an entirely new mountain structure, typically characterized by a singular hole indicative of magma. This comparison underscores the nuanced capability of our approach in maintaining the original form while applying distinct style editing. Utilizing source texts like "wooden door", "croissant" and "red umbrella", our method demonstrates a precise capture of object locations, contrasting with other models that often indiscriminately apply style editing to the entire image, leading to unfavorable background metric scores. Additionally, the six diffusion-based style editing models frequently alter the target object, resulting in poor Con_F scores. While Text2LIVE shows results similar to ours, it notably applies style editing to unintended areas (*e.g.*, wooden door with "gray marble" as style text) and

fails to properly reflect the style indicated by the style text (*e.g.*, red umbrella in "Starry night by Vincent Van Gogh").

4.4. Ablation study

To demonstrate the efficacy of our method, we conduct an ablation study, with quantitative results presented in Tab. 2 and qualitative insights in Fig. 5. Notably, the quantitative results in Tab. 2-(e) which incorporates all proposed losses $\mathcal{L}_{\text{dir}}$, $\mathcal{L}_{\text{con}}$, and $\mathcal{L}_{\text{abp}}$, indicate that our method achieves the best performance. Notably, the absence of the adaptive background preservation loss $\mathcal{L}_{\text{abp}}$ in Tab. 2-(c) results in a significant drop in the background absolute difference metrics L1_B. Omitting the patch distribution consistency loss $\mathcal{L}_{\text{con}}$ in Tab. 2-(d) leads to a decrease in performance across all foreground quality metrics. These findings clearly demonstrate the critical role of each proposed loss, affirming their essential contributions to our design.

The qualitative results in Fig. 5 further illustrate this point. In Fig. 5-(a), a scenario where patches are randomly selected and directional loss based on the text direction is applied without incorporating all proposed losses and modules, demonstrates a significant limitation. Here, the entire source image undergoes style editing, which leads to a distortion of the semantic content originally present in the image. Fig. 5-(b) demonstrates that applying only $\mathcal{L}_{\text{dir}}$ leads to focused style editing on the targeted object, enabled by TMPS, but also results in notable background changes and some loss or alteration of object details. In contrast, Fig. 5-(c), where $\mathcal{L}_{\text{dir}}$ is combined with $\mathcal{L}_{\text{con}}$, effectively preserves crucial object details like the chair's shadow and shape of the hat, though background alterations remain. The comparison of Fig. 5-(d) and (e) reveals the impact of the $\mathcal{L}_{\text{con}}$; in (d), despite achieving object-centric style editing, there is a noticeable loss in object details, particularly in cropped areas. Meanwhile, the comparison of Fig. 5-(c) and (e) demonstrates that with the full complement of losses, both the background remains unchanged and the object details are preserved without distortion, highlighting the effectiveness of the ABP loss.

4.5. Comparison to mask-guided generative models

To highlight the differences from mask-based models and showcase the natural style-editing capabilities of our Style-Editor model, we conduct a comparative analysis with existing mask-guided generative models [1, 41]. These models typically rely on masks to specific areas for image editing. Our comparison, as depicted in Fig. 6, highlights a key distinction. Unlike mask-based models, which may inadvertently distort vital details of the object, Style-Editor consistently maintains the structural integrity of the object, focusing solely on altering its style. This process is crucial in preserving the overall coherence and authenticity of the image. Moreover, Style-Editor effectively identify relevant objects

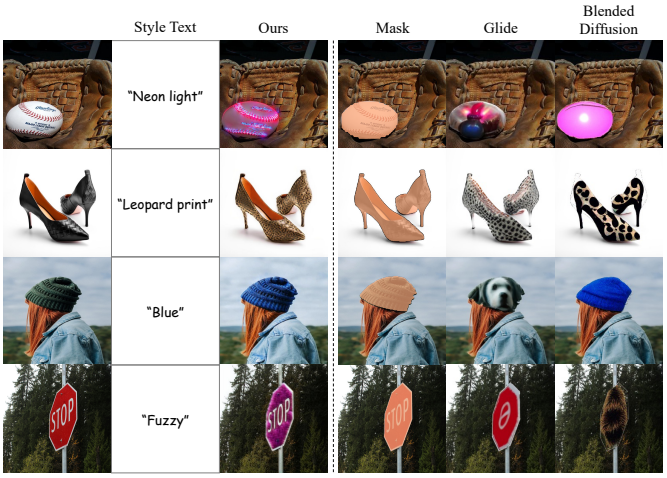

Figure 6. Comparative analysis of Style-Editor with other generative models using mask input. This figure demonstrates the superiority of our model in producing enhanced results compared to methods that use object masks as guidance for style editing tasks.

based on the provided text and apply style edits directly to these areas, eliminating the need for labor-intensive and sometimes imprecise manual mask creation. Style-Editor, thus, not only preserves the integrity of the objects but also enhances efficiency, offering a more user-friendly alternative to traditional mask-guided generative models.

5. Conclusion

In this paper, we have presented Text-driven Object-Centric Style editing (Style-Editor), a new approach for editing the appearance of objects in images based purely on textual descriptions. This method eliminates the need for reference style images or segmentation masks, facilitating complex and tailored style editing directly from text. At the heart of Style-Editor are three pivotal components: the Text-Matched Patch Selection with Pre-fixed Region Selection module, the Patch-wise Co-Directional (PCD) loss and the Adaptive Background Preservation (ABP) loss. The TMPS and PRS modules are pivotal in accurately identifying the location of objects linked to the source text. The PCD loss, complemented by our Text-Matched Patch Selection (TMPS), precisely identifies and selects patches for style editing, ensuring an artifact-free style editing through consistent CLIP-embedding distribution. Meanwhile, the ABP loss plays a critical role in preserving the original style and structure of the background. It adeptly adjusts to dynamically identified background areas, effectively preventing unintended alterations during the style editing process. Extensive experiments validate the superior performance of our Style-Editor, demonstrating its capability to achieve high-quality, text-driven object-centric style editing while preserving the overall visual coherence and integrity of images.

Acknowledgments

This work was supported by the 2025 innovation base artificial intelligence data convergence project project with the funding of the 2025 government (Ministry of Science and ICT) (S2201-24-1002). The supercomputing resources for this work was supported by Grand Challenging Project of Supercomputing AI Education and Research Center, DGIST.

References

[1] 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. 8, 18

[2] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022. 2, 3, 6, 7, 18

[3] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8861–8870, 2024. 3, 6, 7, 18

[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. 2, 3, 6, 7

[5] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 397–406, 2021. 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

[7] Guillaume Couairon, Asya Grechka, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Flexit: Towards flexible semantic image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18270–18279, 2022. 3, 6, 7, 18

[8] Yingying Deng, Fan Tang, Weiming Dong, Wen Sun, Feiyue Huang, and Changsheng Xu. Arbitrary style transfer via multi-adaptation network. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2719–2727, 2020. 2

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 5

[10] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 6

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5

[12] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 2

[13] Fernando A Fardo, Victor H Conforto, Francisco C de Oliveira, and Paulo S Rodrigues. A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms. *arXiv preprint arXiv:1605.07116*, 2016. 6

[14] Kevin Frans, Lisa Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *Advances in Neural Information Processing Systems*, 35:5207–5218, 2022. 3

[15] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-driven image style transfer. 2021. 2

[16] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 2, 3, 4

[17] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 4, 5

[18] Jiayi Guo, Chaofei Wang, You Wu, Eric Zhang, Kai Wang, Xingqian Xu, Shiji Song, Humphrey Shi, and Gao Huang. Zero-shot generative model adaptation via image-specific prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11494–11503, 2023. 3

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[20] Kibeom Hong, Seogkyu Jeon, Junsoo Lee, Namhyuk Ahn, Kunhee Kim, Pilhyeon Lee, Daesik Kim, Youngjung Uh, and Hyeran Byun. Aespa-net: Aesthetic pattern-aware style transfer networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22758–22767, 2023. 2

[21] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, 2015. Association for Computational Linguistics. 5

[22] 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 (ICCV)*, 2017. 2

[23] 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. 6

[24] Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Shifeng Chen, and Liangliang Cao. Diffusion model-based image editing: A survey. *arXiv preprint arXiv:2402.17525*, 2024. 3

[25] Zixuan Huang, Jinghuai Zhang, and Jing Liao. Style mixer: Semantic-aware multi-style transfer network. In *Computer Graphics Forum*, pages 469–480. Wiley Online Library, 2019. 3

[26] Jing Huo, Shiyin Jin, Wenbin Li, Jing Wu, Yu-Kun Lai, Yinghuan Shi, and Yang Gao. Manifold alignment for semantically aligned style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14861–14869, 2021. 3

[27] Surgan Jandial, Shripad Deshmukh, Abhinav Java, Simra Shahid, and Balaji Krishnamurthy. Gatha: Relational loss for enhancing text-based style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3545–3550, 2023. 2

[28] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 2, 6

[29] Chanda Grover Kamra, Indra Deep Mastan, and Debayan Gupta. Sem-cs: Semantic clipstyler for text-based image style transfer. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 395–399. IEEE, 2023. 2

[30] 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

[31] 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

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

[33] Lironne Kurzman, David Vazquez, and Issam Laradji. Class-based styling: Real-time localized style transfer with semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 3

[34] 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

[35] 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. 2

[36] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017. 2

[37] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 453–468, 2018. 2

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 6, 15

[39] 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

[40] 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. 2, 3, 6, 7, 18

[41] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3, 8, 18

[42] 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

[43] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 2, 3, 6, 7

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[45] 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. 3

[46] Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 23051–23061, 2023. 3, 6, 7, 18

[47] Justin NM Pinkney and Chuan Li. clip2latent: Text driven sampling of a pre-trained stylegan using denoising diffusion and clip. *arXiv preprint arXiv:2210.02347*, 2022. 3

[48] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya

Sutskever, et al. Improving language understanding by generative pre-training. 2018. 5

[49] 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

[50] 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

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 5

[52] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992. 4

[53] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3

[54] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[55] Peter Schaldenbrand, Zhixuan Liu, and Jean Oh. Styleclipdraw: Coupling content and style in text-to-drawing translation. *arXiv preprint arXiv:2202.12362*, 2022. 3

[56] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2, 3, 6, 7

[57] 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. 2

[58] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 2

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[60] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 6

[61] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 15

[62] Zipeng Xu, Enver Sangineto, and Nicu Sebe. 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*, pages 7601–7611, 2023. 6, 7

[63] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22873–22882, 2023. 2, 3, 6, 7

[64] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. Attention-aware multi-stroke style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1467–1475, 2019. 2

[65] Moon Ye-Bin, Jisoo Kim, Hongyeob Kim, Kilho Son, and Tae-Hyun Oh. Textmania: Enriching visual feature by text-driven manifold augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2526–2537, 2023. 5

[66] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jiahui Zhang, Shijian Lu, Miaomiao Cui, Xuansong Xie, Xian-Sheng Hua, and Chunyan Miao. Towards counterfactual image manipulation via clip. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3637–3645, 2022. 2