Gatha: Relational Loss for enhancing text-based style transfer

Surgan Jandial, Shripad Deshmukh, Abhinav Java, Simra Shahid, Balaji Krishnamurthy
Adobe MDSR
{jandial,sdeshmu,ajava,sshahid,kbalaji}@adobe.com

Abstract

Text-based style transfer is a promising area of research that enables the generation of stylistic images from plain text descriptions. However, the existing text-based style transfer techniques do not account for the subjective nature of prompt descriptions or the nuances of style-specific vocabulary during the optimization process. This severely limits the stylistic expression of the predominant models. In this paper, we address this gap by proposing Gatha, which incorporates subjectivity by introducing an additional loss function that enforces the relationship between stylized images and a proxy style set to be similar to the relationship between the text description and the proxy style set. We substantiate the effectiveness of Gatha through both qualitative and quantitative analysis against the existing state-of-the-art models and show that our approach allows for consistently improved stylized images.

1. Introduction

The recent developments in Large Language Models (LLMs [2, 18, 27, 30]) and Visual Language Models (VLMs [16, 22, 23, 25, 28]) has paved the way for many existing and new applications. One such application – Style transfer – has traditionally been image-based [4, 7, 11, 15, 29], which offers limited expressiveness in terms of complex styles. For instance, consider the prompt: ‘Create an image that blends the classical style of a Renaissance painting with the futuristic aesthetic of a cityscape’. This prompt involves merging two distinct styles, which is challenging to express through the visual modality. On the contrary, with natural language, one can describe an endless array of styles - with the power of LLMs, text-based style transfer [5, 31] has begun addressing this issue.

The recent methods in the nascent field of text-conditioned style transfer, e.g. CLVA [5], StyleCLIP [21] and CLIPstyler [14], show tremendous promise in terms of the quality and diversity of the generated outputs. These approaches base their solution on aligning the directions of text embeddings and image embeddings [21], in that they direct the stylized image to be faithful to the direction of the text. Nevertheless, these models do not consider the subjective nature of style descriptions or the nuances of style-specific vocabulary in their optimization process. Therefore, moving the image in the direction of the text vector may not necessarily ensure that the relationship b/w the image and the style is the same as the relationship b/w text and style.

Thus, we introduce Gatha, a framework to incorporate a relational loss in style transfer systems. More specifically, we sample a proxy set of well-known style templates (see Fig. 1(a)), and hypothesize that the relation/similarity of these templates with the stylized image should match their relation/similarity with the target text description. We show the efficacy of Gatha, by employing it with the current state-of-the-art approach, CLIPstyler [14] as the baseline. We demonstrate how Gatha, with simply an additional loss, produces consistently better outputs than the baseline (CLIPstyler), both qualitatively and quantitatively. Also, this modification of loss requires changing only a few lines of code. Owing to its simplistic nature, Gatha can be transferred to other text-guided computer vision approaches via the introduction of an arbitrary style basis. Our contributions can be summarized below:

1. We propose a simple framework, Gatha, that incorporates the subjective nature of style descriptions.
2. Our framework achieves this by leveraging a proxy set of well-known styles (in text) and then enforcing the relationship of the target style description and style set to be similar to that of the stylized image and style set.
3. We empirically show the efficacy of Gatha over the existing state-of-the-art techniques in text-guided style transfer.

2. Related Works

Our work belongs to the Text-guided Style Transfer research area [5, 14, 21], which modifies the content image
using a natural language description of the desired style, leveraging the multimodal capabilities of VLMs such as CLIP [22], BLIP [16], and Kosmos [9].

While neural style transfer approaches [3, 4, 7, 10–13, 15, 17, 20, 26, 29], which rely on a target-style image as input were popular in the past, they have limited capability and expressiveness for transferring complex styles that require a textual description.

In the past, text-guided style transfer methods have utilized contrastive learning techniques to create representations of both text and image [5, 21]. The current state-of-the-art, CLIPstyler [14], has introduced a patch-level image-text matching loss that employs CLIP [22] embeddings along with several image augmentations to enhance the quality of text-guided transfer. While some extensions have been suggested [1, 8, 19], they are trained on general-purpose in-the-wild image captioning datasets and don’t include style-specific knowledge, which presents an opportunity for improvement.

3. Preliminaries of Text-guided Style Transfer

Recent state-of-the-art style transfer methods like StyleCLIP [21], StyleGAN-NADA [6] and CLIPstyler [14] leverage the representation capability of CLIP [22] and formulate style transfer as:

\[
\min_{\theta_f} [L_{\text{dir}}(\theta_f, \theta_C) + L_{\text{content}}(\theta_f, \theta_C)]
\]

where \( f \) is the stylized image generator (U-Net [24]), \( C \) is the frozen CLIP model, \( L_{\text{dir}} \) is the directional loss, and \( L_{\text{content}} \) is the content loss.

3.1. Directional Loss \( L_{\text{dir}} \)

This represents the cosine distance between the direction vectors of text and image modalities. Concretely, we first compute unit vector joining clip text embeddings of \( T_{in} \) – the placeholder textual description of content image \( I_{in} \) ("a photo"), and \( T_{tar} \) – the user-specified target style description (“A vintage photo of Brad Pitt.”), respectively. Likewise, we compute unit vector joining clip image embeddings of \( I_{in} \) and its desired stylised output \( I_{out} = f(I_{in}) \), respectively. Mathematically, if \( C_T \) and \( C_C \) denote the clip image and text encoders, the direction vectors are given as,

\[
T_{\text{dir}} = \frac{C_T(T_{tar}) - C_T(T_{in})}{\|C_T(T_{tar}) - C_T(T_{in})\|_2}
\]

\[
I_{\text{dir}} = \frac{C_C(I_{out}) - C_C(I_{in})}{\|C_C(I_{out}) - C_C(I_{in})\|_2}
\]

The directional loss \( L_{\text{dir}} \) is then given by

\[
L_{\text{dir}} = 1 - T_{\text{dir}} \cdot I_{\text{dir}}
\]

CLIPstyler applies \( L_{\text{dir}} \) at two levels: global image level \( L_{\text{dir}, \text{glob}} \), and image patch level \( L_{\text{dir}, \text{patch}} \).

3.2. Content Loss \( L_{\text{content}} \)

This represents the mean-squared error between the features of content and output stylized images both extracted from the pre-trained VGG-19 networks [7]. Additionally, [14] and other methods also deploy a total variation regularization loss \( L_{tv} \) to handle the artifacts from irregular pixels.

4. Proposed Methodology

We propose a novel approach to style transfer that goes beyond directional losses (Eqn 2) commonly used in the field. Our method introduces a relational loss that captures the nuanced relationships between image and text descriptions and improves the discriminative ability of the underlying style transfer model. We achieve this by comparing the relationship that a generated image forms with a set of style templates to the relationship that its target style text description forms with the same style templates.

In the following sections, we describe the building blocks of our model: Style Templates, Style Tensor, and Relational Loss. Then we show the step-by-step architecture of our model in figure 1.

4.1. Creating the Style Tensor

We demonstrate the creation of style tensor in Figure 1 (a). We first collect a list of 261 commonly used style templates from popular design platforms. Each style template (or basis) is preprocessed into a question format, and the resulting list of questions is encoded into a style tensor using an embedding size of 512. The question prompt is of the following format – “Is the style <Style>?”.

The style tensor, denoted as \( S \), has dimensions of \( N \times 512 \), where \( N \) is the number of style templates in the list. We encode the basis in a style tensor \( S \) using \( C_T \) as follows:

\[
S = C_T(T_{\text{style}}) \in \mathbb{R}^{N \times 512}
\]

4.2. Relational Loss \( L_{gatha} \)

Our proposed relational loss aims to ensure that the relationship between the generated stylized image and the style tensor is similar to the relationship between the target style text description and the style tensor. This captures the nuanced relationships between the image and text descriptions more effectively, resulting in better performance in the style transfer task. We discuss the comparison of the relational loss with directional loss in Figure 2.

To compute the relationship between the target text description \( T_{tar} \) and the style tensor \( S \), we use a similarity
score $T_{rel}$, obtained by projecting the target text embedding $C_T(T_{tar})$ over the style tensor $S$.

$$T_{rel} = S_{N \times 512} \times (C_T(T_{tar}))^T_{512 \times 1} \quad (4)$$

where $T_{rel} \in R^{N \times 1}$ is the relation vector independent of the embedding dimension of the VLM, where $N$ is the number of styles in the style tensor.

For the stylized image $I_{out}$, we consider its patches and encode their relationships independently. We first augment each patch, compute its clip image embedding $C_I(aug(I_{out}^i))$, and then compute its similarity score $I_{rel}^i$ as follows:

$$I_{rel}^i = S_{N \times 512} \times (C_I(aug(I_{out}^i)))^T_{512 \times 1} \quad (5)$$

To incorporate the relational loss, $L_{gatha}$, we measure the mean squared error between the image and text style relation vector averaged over all patches -

$$L_{gatha} = \frac{1}{M} \sum_{i=1}^{M} \| I_{rel}^i - T_{rel} \|^2_2 \quad (6)$$

5. Experiments and Results

In this section, we present qualitative and quantitative results for our approach. To ensure consistency, the dataset along with the resolution of content images and the other configuration settings exactly follow CLIPstyler (CS) [14].
image, and so the image almost looks similar to the original (unstyled) image. In contrast to this, in Table 1 (b), we observe our approach to outperform CS by considerable margins in both CLIP scores and SSIM. Furthermore, Fig. 3 qualitatively establishes the efficacy of Gatha over CS across multiple reference images, and text prompts.

5.2. Ablations

We experiment \(L_{gatha}\) with different weight (\(\beta\)) parameters in Table. 1 (b), and observe that increasing \(\beta\) beyond a point results in worse behaviors. Also, style labels used in our loss functions are currently sampled randomly across the web, and hence we probe the performance effects of their choice. In Table 2, we thus consider 3 different types of style basis and find that \(L_{gatha}\) by its design is able to show improvements in all three cases, irrespective of their variations.

<table>
<thead>
<tr>
<th>Setting</th>
<th>CLIP Score</th>
<th>SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>23.82</td>
<td>0.3895</td>
</tr>
<tr>
<td>CS w/o (L_{dir,patch})</td>
<td>19.73</td>
<td>\textbf{0.4005}</td>
</tr>
<tr>
<td>CS w/o (L_{dir,glob})</td>
<td>23.33</td>
<td>0.3886</td>
</tr>
<tr>
<td>CS w/o (L_{TV})</td>
<td>\textbf{23.91}</td>
<td>0.3880</td>
</tr>
</tbody>
</table>

Table 1. (a). Performance of CLIPstyler across a variation of losses. (b). Performance of \(L_{gatha}\) across multiple \(\beta\) parameters.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>CS</th>
<th>I</th>
<th>II</th>
<th>III</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP Score</td>
<td>23.82</td>
<td>24.07</td>
<td>\textbf{24.18}</td>
<td>24.02</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.3895</td>
<td>\textbf{0.3913}</td>
<td>0.3887</td>
<td>0.3909</td>
</tr>
</tbody>
</table>

Table 2. Experimenting \(L_{gatha}\) with different style basis \{I, II, III\}, where CS is CLIPstyler.

References


[17] 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. ii


[26] Artiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Bjorn Ommer. A style-aware content loss for real-time hd


