SceneTextStyler: Editing Text with Style Transformation

Honghui Yuan[0009−0001−4334−9363] and Keiji Yanai[0000−0002−0431−183X]

The University of Electro-Communications, Chofu, Tokyo, JAPAN
{yuan-h, yanai}@mm.inf.uec.ac.jp

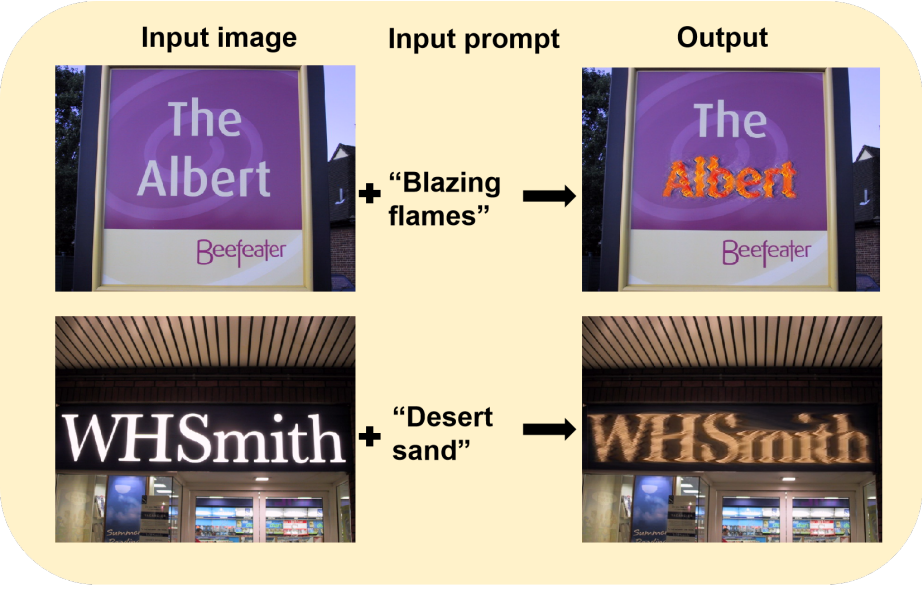

Fig. 1. The process of editing text with style transformation of our proposed system.

Abstract. Scene text editing has gained widespread use in fields like poster design. In this study, we propose a new task for scene text editing that keeps the image background and text content unchanged while only modifying the style of the text. To accomplish this, we developed a web-based application that enables style transformation of text in scene images using prompts. As shown in Fig.1, our application successfully transforms text in an image into any desired style using prompts. Experimental results demonstrate that the proposed application can naturally apply different styles to scene text without altering the image background and text content. Furthermore, our model does not require any additional training and is capable of real-time style transfer for scene text.

Keywords: scene text edit · style transfer · CLIPstyler

1 Introduction

Recently, with the advancement of deep learning, image editing has become more accessible and has achieved remarkable results. Scene text editing, a subset of

image editing, has garnered significant attention. Previous methods for scene text editing primarily focused on replacing text in scene images while preserving the background and maintaining the text's style (such as color, texture, etc.). However, these approaches only allow for the replacement of text content, offering no flexibility for users to freely modify the text style. To address this limitation, we propose a novel task: scene text style transformation, where only the text style is altered while keeping the text content and background intact. Traditional scene text editing methods typically involve three steps: background restoration, text conversion, and image re-composition. These methods rely on the original image as a style reference during text transformation to ensure that the newly generated text matches the style of the original text. However, they do not offer the ability to change the style of the text. Moreover, while state-of-the-art diffusion model-based methods can produce more natural scene text images, they generate text styles based on the existing text in the image, limiting users from freely altering the style during text editing.

Image style transformation typically involves generating new images by combining the content of one image with the style of another. Numerous methods like Stytr2 [4], and Drb-gan [23] have successfully generated attractive results. However, it can sometimes be challenging to find a suitable reference style image. In such cases, using prompts for style transformation provides a simpler and more efficient approach. Recently, the CLIP model [15], a multi-modal model that integrates language and images, has been employed to guide image generation using text prompts. CLIPstyler [11] leverages a lightweight CNN network in combination with the CLIP model to perform image style transformation based on simple text descriptions, eliminating the need for reference style images. This approach can maintain the main content of the original image while applying the desired style transformation. Unlike traditional image style transformation, the style transformation of text images is more complex, as it requires preserving the readability of the text while applying stylistic changes. Several studies like Dg-Font [22], and Cf-Font [20] have explored text image style transformation, typically using networks that learn the structure of the text to ensure content retention. However, these studies are often limited to images containing single characters and struggle to handle images containing multiple characters, such as words.

Therefore, we propose a new web-based application based on CLIPstyler [11] to enable style transformation specifically for the text in scene text images. Our application allows users to transform the style of the text into any desired style through prompts, without the need for a reference style image and without altering the image background or the text content.

2 Related Work

Scene text editing has made significant progress in replacing text within an image while preserving the appearance of the text. For example, STEFANN [17] designed two networks for font structure and color transformation, each focused on replacing text one character at a time, although it cannot handle cases where

the number of characters differs from the original. SRNet [21] replaced text using three sub-networks: background restoration, text transformation, and resynthesis. SwapText [24], based on SRNet, introduced a Thin-Plate-Spline (TPS) module to geometrically transform text using spatial points, while SimAN [12] introduced normalization for similarity recognition and uses a self-supervised learning method for network training. TextStyleBrush [10] incorporated text image style vectors into the generator, leveraging StyleGAN [9] to guide the generation of final images. Mostel [14] improved scene text editing performance by introducing stroke-level information and utilizing both synthetic and real-world data. However, these methods typically require a style reference image. Recent diffusion models [16] have achieved significant success in image editing. Methods such as DiffSTE [7], DiffUTE [2], GlyphDraw [13], GlyphControl [25], and TextDiffuser [3] use diffusion models for natural scene text generation and editing, though they lack control over the text's style. In contrast, our study does not rely on a style reference image and allows users to specify the style of scene text through prompts.

Image style transformation aims to transfer the style of a reference image to a target image. Numerous methods, such as StyleGAN [9], and StyTr2 [4] have achieved significant success in style transformation. However, these approaches generally require a style reference image. Recently, CLIPstyler has addressed this limitation by enabling arbitrary style transformations using text prompts. Methods like Sem-CS [8] and Gen-Art [26] have further refined this by using semantic segmentation to solve the over-stylization problem in the foreground. However, these techniques applied stylization to the entire image, lacking the ability to focus on specific targets within the image. Several methods, including Word as Image [5], CLIPFont [18], DS-Fusion [19], and Zero-shot Font Style Transfer [6], have demonstrated the ability to transfer font styles through prompts. However, these methods are limited to font-specific images. To solve these limitations, in this study, we propose an application that achieved style transformation of text areas within scene images based on prompts.

3 Proposed System

3.1 System Overview

In this study, we proposed a web-based application to achieve style transformation specifically for the text areas in scene text images. The overview of our application is shown in Fig 2. The input to our application consists of a scene text image and a style prompt. First, users can select the scene text image and input the desired style prompt. Then, they can choose the specific text area they wish to modify in the image. We provide sample scene text images within the app for users to select from. Additionally, users can input custom text to control the style, or they can choose from the provided samples. After selecting the scene text image and style prompt, users can click the "Run Generation" button, which will apply the chosen style to the text in the image. The styled result will then be displayed in the result bar.

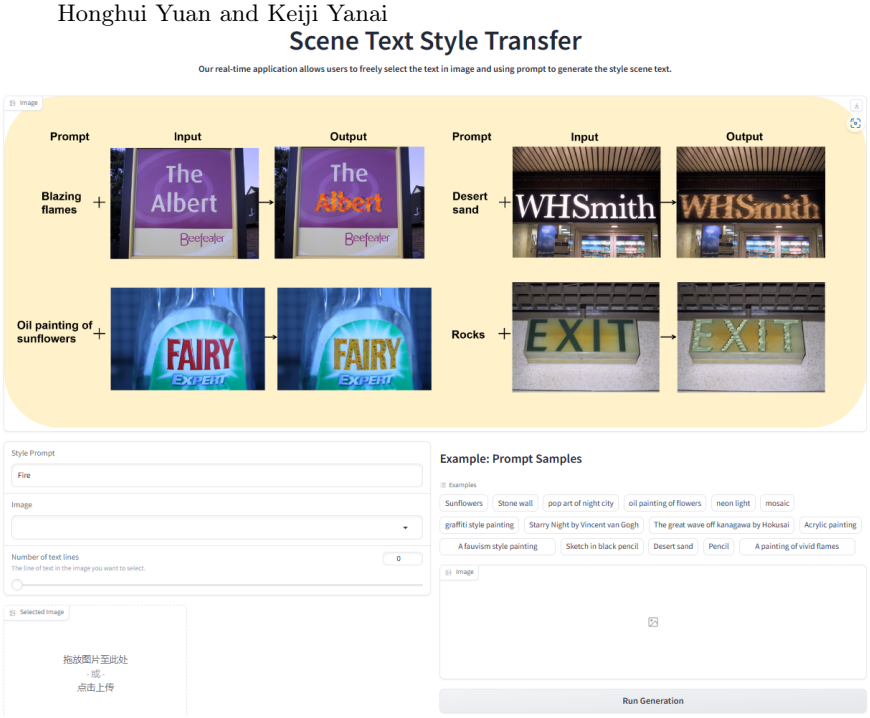

Fig. 2. The interface of our application.

3.2 Network of application

In our application, we designed a network to address the limitations of existing methods, which cannot perform style transfer on a specific part of the image. An overview of the proposed network is shown in Fig. 3. The proposed network mainly consists of a MaskNet network that extracts the mask image of the text part of the scene text image and a StyleNet network that performs style transformation. Using the pre-trained text image embedding model CLIP [15] and the loss function proposed in this study, the parameters of StyleNet are optimized to apply the style features based on the input prompts, generating the desired styled output.

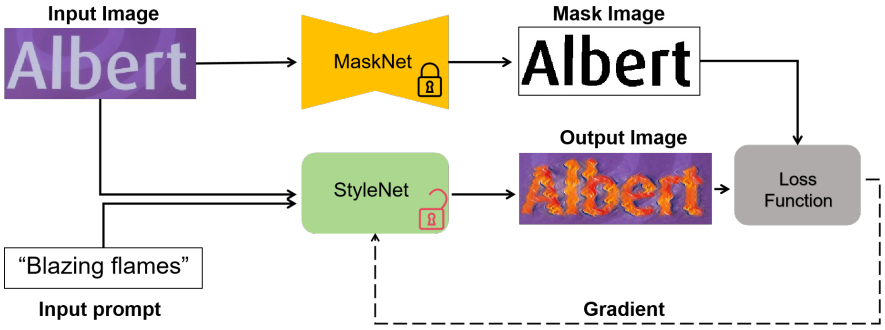

Fig. 3. The network of our application.

CLIPstyler has made image style transformations more accessible by using prompts, allowing for arbitrary style changes. However, CLIPstyler applies style transformations to the entire image and does not support transforming specific regions within an image. To address this limitation, we proposed a new network based on CLIPstyler that enables style transformation specifically for scene text. When altering the style of specific areas in an image without affecting other parts, a simple and effective approach is to use mask images to control the style transformation region. To achieve this, We utilize EasyOCR[1] to recognize each text part of the images and we propose a network called MaskNet, which extracts text masks from an image, allowing style changes to be applied solely to the text in a scene text image while preserving the background and text content. During the scene text style transformation process, MaskNet remains frozen and generates a mask image of the text in the scene image. Additionally, a CNN encoder-decoder network, StyleNet, is employed to generate stylized scene text images by optimizing its parameters using the loss function introduced in this study.

3.3 Loss Function

To transform the style in the text region of an image, We proposed a new loss function Text-aware Loss with the following components. Firstly, we introduced the Distance Transform Loss [1] in this study. Specifically, a distance transform map is generated from the mask image of the text region in the scene text image. Both the input and output images are multiplied by this distance transform map, and the mean squared error (MSE) is calculated.

We also modified the Patch CLIP Loss used in CLIPstyler. Specifically, the background region of the input image is extracted using the text mask, and the cosine similarity between the patches of the generated image and the original background is computed. Patches with significantly different similarities are identified as belonging to the text region, and the Patch CLIP Loss is then calculated only for those patches.

To minimize the influence of the original text style on the generated image and to ensure that the background remains unaltered, we introduced a Background Reconstruction Loss. Specifically, VGG Loss is calculated for the patches corresponding to the background region.

4 Experiments and results

MaskNet was trained using 2000 real-world scene text images collected from Mostel [14]. When using the proposed application, MaskNet is frozen and only StyleNet is optimized. The input scene text images are converted to 512×512 resolution, and after the optimization, the output result is returned to its original size. Set λ_d, λ_t, λ_b and λ_{tv} to 1×10^2, 9×10^3, 150 and 2×10^{-3}. The model used a learning rate of 5×10^{-4} and Adam optimizer. The iteration is set to 500 and the

[1] EasyOCR. https://github.com/JaidedAI/EasyOCR.

learning rate is halved every 100 iterations. A single NVIDIA TITAN RTX was used to test the model, and the generation time per image was approximately 90 to 120 seconds. The results of our application are shown in Fig. 4

Fig. 4. The results of our application. The first line is the input scene text, the second line is the result of our application, and the third line is the input prompt.

5 Conclusion

In this study, we proposed a web-based application to achieve style transformation specifically for scene text. Unlike previous methods, the proposed application does not require a style reference image and allows users to freely modify the style of text regions in a scene image using prompts. The proposed loss function and MaskNet in this study addressed the limitation of CLIPstyler, which could not apply style transformations to specific regions of an image. Experimental results confirmed that the proposed method generates visually appealing styled scene text images while preserving both the image background and the text content. Use our app can quickly edit posters, covers, and other artwork to generate stylized artwork. However, this research currently supports only English text and does not extend to languages like Kanji or Katakana, as MaskNet is limited to alphabetic characters. In future work, we aim to expand the functionality to include scene text style conversion for additional languages. The special structure of fonts results in some styles not being well expressed in the font. In the future, we will utilize a greater degree of font deformation to solve this problem.

References

1. Atarsaikhan, G., Iwana, B.K., Uchida, S.: Contained neural style transfer for decorated logo generation. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS). pp. 317–322 (2018)
2. Chen, H., Xu, Z., Gu, Z., Li, Y., Meng, C., Zhu, H., Wang, W., et al.: Diffute: Universal text editing diffusion model. Advances in Neural Information Processing Systems **36** (2024)
3. Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., Wei, F.: Textdiffuser: Diffusion models as text painters. Advances in Neural Information Processing Systems **36** (2024)
4. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: Stytr2: Image style transfer with transformers. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 11326–11336 (2022)
5. Iluz, S., Vinker, Y., Hertz, A., Berio, D., Cohen-Or, D., Shamir, A.: Word-as-image for semantic typography. ACM Transactions on Graphics (TOG) **42**(4), 1–11 (2023)
6. Izumi, K., Yanai, K.: Zero-shot font style transfer with a differentiable renderer. In: Proceedings of the 4th ACM International Conference on Multimedia in Asia. pp. 1–5 (2022)
7. Ji, J., Zhang, G., Wang, Z., Hou, B., Zhang, Z., Price, B., Chang, S.: Improving diffusion models for scene text editing with dual encoders. arXiv preprint arXiv:2304.05568 (2023)
8. Kamra, C.G., Mastan, I.D., Gupta, D.: Sem-cs: Semantic clipstyler for text-based image style transfer. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 395–399. IEEE (2023)
9. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
10. Krishnan, P., Kovvuri, R., Pang, G., Vassilev, B., Hassner, T.: Textstylebrush: Transfer of text aesthetics from a single example. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
11. Kwon, G., Ye, J.C.: Clipstyler: Image style transfer with a single text condition. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 18062–18071 (2022)
12. Luo, C., Jin, L., Chen, J.: Siman: exploring self-supervised representation learning of scene text via similarity-aware normalization. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 1039–1048 (2022)
13. Ma, J., Zhao, M., Chen, C., Wang, R., Niu, D., Lu, H., Lin, X.: Glyphdraw: Learning to draw chinese characters in image synthesis models coherently. arXiv preprint arXiv:2303.17870 (2023)
14. Qu, Y., Tan, Q., Xie, H., Xu, J., Wang, Y., Zhang, Y.: Exploring stroke-level modifications for scene text editing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2119–2127 (2023)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
16. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)

17. Roy, P., Bhattacharya, S., Ghosh, S., Pal, U.: Stefann: scene text editor using font adaptive neural network. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 13228–13237 (2020)
18. Song, Y., Zhang, Y.: Clipfont: Text guided vector wordart generation. In: 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022. BMVA Press (2022), https://bmvc2022.mpi-inf.mpg.de/0543.pdf
19. Tanveer, M., Wang, Y., Mahdavi-Amiri, A., Zhang, H.: Ds-fusion: Artistic typography via discriminated and stylized diffusion. In: Proc. of IEEE International Conference on Computer Vision. pp. 374–384 (2023)
20. Wang, C., Zhou, M., Ge, T., Jiang, Y., Bao, H., Xu, W.: Cf-font: Content fusion for few-shot font generation. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 1858–1867 (2023)
21. Wu, L., Zhang, C., Liu, J., Han, J., Liu, J., Ding, E., Bai, X.: Editing text in the wild. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 1500–1508 (2019)
22. Xie, Y., Chen, X., Sun, L., Lu, Y.: Dg-font: Deformable generative networks for unsupervised font generation. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 5130–5140 (2021)
23. Xu, W., Long, C., Wang, R., Wang, G.: Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer. In: Proc. of IEEE International Conference on Computer Vision. pp. 6383–6392 (2021)
24. Yang, Q., Huang, J., Lin, W.: Swaptext: Image based texts transfer in scenes. In: Proc. of IEEE Computer Vision and Pattern Recognition. pp. 14700–14709 (2020)
25. Yang, Y., Gui, D., Yuan, Y., Liang, W., Ding, H., Hu, H., Chen, K.: Glyphcontrol: Glyph conditional control for visual text generation. Advances in Neural Information Processing Systems **36** (2024)
26. Yang, Z., Song, H., Wu, Q.: Generative artisan: A semantic-aware and controllable clipstyler. arXiv preprint arXiv:2207.11598 (2022)