실제 장면 이미지에서 포맷 제어 가능한 텍스트 편집

Quang-Vinh Dang[1], Hyung-Jeong Yang[1], and Soo-Hyung Kim [1]
[1] 전남대학교 인공지능융합학과

quangvinh242003@yahoo.com, hjyang@jnu.ac.kr, shkim@jnu.ac.kr

Format-Controllable Text Editing in Real-Scene Images

Quang-Vinh Dang[1], Hyung-Jeong Yang[1], and Soo-Hyung Kim[1]
[1]Dept. of Artificial Intelligence Convergence, Chonnam National University

Abstract

Flexibility is crucial in applications where users or systems require precise control over the appearance of text in images, particularly in scene text editing tasks. However, previous methods have primarily focused on altering text content, often neglecting the important aspect of controlling text formatting. In this paper, we propose a text editing model that not only edits content but also provides control over the format, utilizing a diffusion model with denoising and text-aware losses. By integrating these mechanisms, the model is capable of generating high-quality scene text images based on user-specified inputs such as text, size, and font, ensuring that both the content and appearance align with user preferences. We evaluate the model's performance using OCR accuracy on the ICDAR FST dataset, and the results demonstrate that our approach is highly competitive and effective when compared to existing methods in the field.

1. Introduction

Scene text editing involves modifying text within real-world images while ensuring seamless integration with the background. This task is critical in applications such as updating advertisements, correcting documents, and enhancing augmented reality, where both the content and visual presentation of the text must be precise. Key formatting elements—including font, size, orientation, capitalization, and spacing—are essential for maintaining readability and ensuring the text naturally blends into the scene.

Despite recent advances in diffusion models, challenges remain in accurately rendering text while preserving its style. The DiffUTE model [1] addresses some of these challenges by enabling precise multilingual text editing through the use of character glyphs, positional information, and a self-supervised learning framework for large-scale unannotated data. Similarly, MOSTEL [2] employs semi-supervised learning to improve background consistency, achieving strong results on datasets such as Tamper-Syn2k and Tamper-Scene. RewriteNet [3] furthers the field by decomposing content and style features, enabling seamless text editing while preserving the original style, and bridging the gap between synthetic and real-world data through self-supervised learning. In related work, Dang et al. [4][5] proposed a GAN-based model to generate paired data for scene text segmentation, though it resulted in only moderate image quality.

However, these methods do not allow for control over the format of the edited text, limiting their flexibility, as shown in Figure 1a. In this paper, we propose a novel, flexible, and controllable text editing framework based on diffusion models, allowing users to modify not only the content but also the format of the text, as illustrated in Figure 1b.

During training, we fine-tuned the latent diffusion model with pixel-level text segmentation. To ensure the generated text maintains the desired format, we introduce a pixel-level text-aware loss, along with a denoising loss to improve image quality and background coherence. In testing, users provide text format specifications like font, size, and text location. The text generator creates a segmentation image based on these inputs, which is combined with the original scene text image and processed by the diffusion model to produce the final edited image.

Figure 1: Comparison of Samples from Previous Methods (Text-Only Editing) and the Proposed Method (Text and Format Editing).

2. Proposed Method

The diffusion model processes extracted features from image inputs based on the VAE model [6] to reduce complexity, while using image pixels for better visualization (Figure 2). In training, denoising and pixel-level text-aware losses are applied: the former ensures high-quality images, and the latter guarantees the text matches the specified format. Hugging Face Diffusers [5] is used for the diffusion process, with a pre-trained UNet. In inference, the text generator (via Matplotlib) uses user inputs for content, and format to create a text segmentation image, enabling the model to generate scene text images that match the desired format.

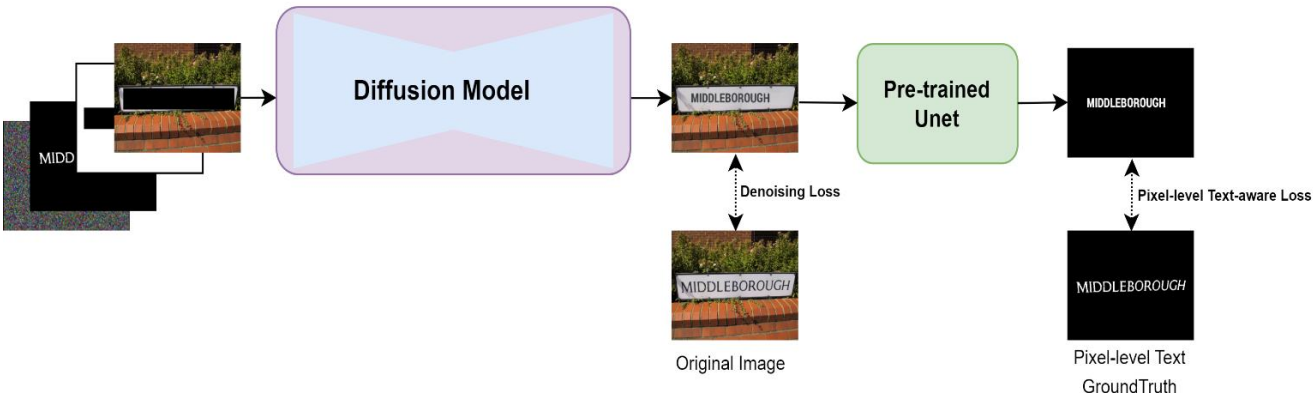

Figure 2: The diffusion model generates images conditioned on noisy features, pixel-level text segmentation, text region masks, and masked text region features.

3. Result and Discussion

Table 1 demonstrates the high quality of the generated scene text images. However, the primary contribution of this paper lies in the ability to control and preserve the format of the edited text, a feature that cannot be adequately measured by OCR accuracy alone.

Table 1: Quantitative comparison on the ICDAR13 dataset. Higher values indicate better performance, and OCR accuracy is reported by comparing the generated text to the target text.

Models	***OCR Accuracy***
Pix2Pix [7]	15.48
MOSTEL [2]	53.75
DiffSTE [9]	81.48
Ours	**83.82**

4. Conclusion

This paper introduced a diffusion-based framework for scene text editing that allows users to control both the content and format of text. Our model, which combines denoising loss and a novel text-aware loss, produces high-quality scene text images with user-defined formats. It achieved competitive OCR accuracy on the ICDAR13 dataset, slightly outperforming existing methods.

Acknowledgement:

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government (MSIT), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00219107).

References

[1] Chen H., Xu Z., Gu Z., Li Y., Meng C., Zhu H., Wang W., "DiffUTE: Universal text editing diffusion model," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[2] Qu Y., Tan Q., Xie H., Xu J., Wang Y., Zhang Y., "Exploring stroke-level modifications for scene text editing," Proceedings *of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, pp. 2119-2127, 2023.

[3] Lee J., Kim Y., Kim S., Yim M., Shin S., Lee G., Park S., "RewriteNet: Reliable scene text editing with implicit decomposition of text contents and styles," arXiv preprint, arXiv:2107.11041, 2021.

[4] Dang Q.V., Lee G.S., "Scene text segmentation via multi-task cascade transformer with paired data synthesis," *IEEE Access*, 2023.

[5] Dang Q.V., Lee G.S., "Scene text segmentation by paired data synthesis," *Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP)*, pp. 545-549, 2023.

[6] Kingma D.P., "Auto-encoding variational bayes," *arXiv preprint*, arXiv:1312.6114, 2013.

[7] Isola P., Zhu J.Y., Zhou T., Efros A.A., "Image-to-image translation with conditional adversarial networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125-1134, 2017.

[8] Wu L., Zhang C., Liu J., Han J., Liu J., Ding E., Bai X., "Editing text in the wild," *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1500-1508, 2019.

[9] 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

[10] Fang S., Xu C., Niu Y., Chen Z., Pu S., Huang F., "Read like humans: Autonomous, bidirectional and iteratively refining scene text recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7098-7107, 2021.

[11] Chen Z., Lin W., Huang J., Pu S., "TextDiffuser: Diffusion models for scene text editing," arXiv preprint, arXiv:2304.02328, 2024.

[12] Karatzas D., Shafait F., Uchida S., Iwamura M., Bigorda L., Mestre S.R., Mas J., Mota D.F., Almazan J., de las Heras L.P., "ICDAR 2013 Robust reading competition," Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 1484-1493, 2013.

[13] Ch'Ng S., Chan C.S., "Total-Text: A comprehensive dataset for scene text detection and recognition," Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 935-942, 2017.

[14] Xu Y., Wang X., Li X., Lv Z., Zhang Y., "Rethinking text segmentation: A novel dataset and method," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2563-2572, 2021.