WAS: Dataset and Methods for Artistic Text Segmentation
Supplementary File

Xudong Xie[1], Yuzhe Li[1], Yang Liu[1], Zhifei Zhang[2], Zhaowen Wang[2], Wei Xiong[2], and Xiang Bai[1](✉)

[1] Huazhong University of Science and Technology, China
{xdxie,yzli12,yangliu1213,xbai}@hust.edu.cn
[2] Adobe, USA
{zzhang,zhawang}@adobe.com, wxiongur@gmail.com

1 More Details on Synthetic Dataset Construction

As stated in the paper, we first construct the training pipeline for a text image generation model, learning to generate text images spatially aligned with text masks. Then we construct an inference pipeline to input new masks and prompts into the trained generation model, generating new text images. Here we add the details of prompt generation in the training and inference pipelines.

1.1 Training Pipeline

To train the text image generation model such as ControlNet [5], it is necessary to obtain training data of <caption, text mask, text image> triplets. Text masks and text images are from our proposed real dataset. Captions should be detailed descriptions of the text images. To this end, we utilize a large multi-modal model, Monkey [2], to caption the images. Monkey is an open-source model and can handle vision-language tasks with high-resolution input and detailed scene understanding. It performs well on Image Captioning and various Visual Question Answering (VQA) tasks. Therefore, we feed a text image and a prompt *"generate the detailed caption in English"* to Monkey and let it output a detailed description. The examples of the generated captions are shown in Fig. 1. We found that, in many cases, Monkey is able to recognize and describe the text in images. To ensure the accuracy of the descriptions and to highlight the importance of the text, we add a sentence after each caption: *This image contains the text "text in the image".*

1.2 Inference Pipeline

During the inference phase, we first need to produce new binary masks of text through the Mask Render introduced in the paper. Moreover, it is crucial to

✉ Corresponding author

This image features a white t-shirt with a cartoon character on it, as well as a green and white flower with a rainbow on it. The character has green hair and is wearing a green hat. There is also a necklace with a cross on it in the image. The text "BILLIE ELLISH" can be seen in the top left corner of the image. Overall, this image seems to be related to music or fashion, possibly featuring Billie Eilish or her merchandise. This image contains the text “BILLIE, EILISH”.

This image features a tote bag with the phrase "your feelings are valid" printed on it. The bag is hanging from a pink curtain and there is a pink and white flag nearby. The bag also has a brown and white logo with the word "your" on it. In the background, there is a pink and white striped wall. The overall message of the image is to encourage people to embrace their emotions and feelings. This image contains the text “your, feelings, are, valid”.

This image shows a sign for Bay Forest Estates, located in a residential area. The sign is white with green lettering and features a tree logo. There is also a white pole nearby. In the background, there is a large tree with green leaves and a red and white sign. The sign for Bay Forest Estates is positioned in the center of the image, with the other objects surrounding it. This image contains the text “Bay, Estates, Forest”.

Fig. 1: The captions generated by Monkey [2] for training.

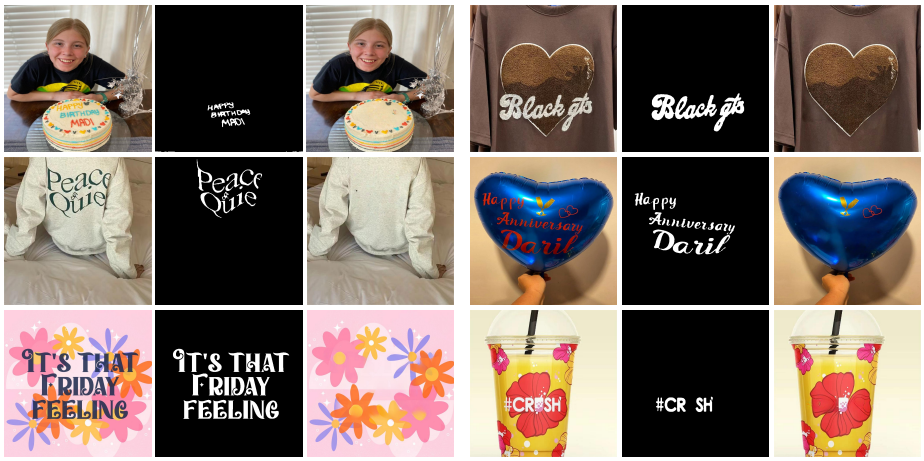

Fig. 2: Text removal visualization using predicted text masks from our WASNet and inpainting model LaMa [4]. Each sample includes the original image, the predicted mask, and the text removal result from left to right.

generate new prompts that describe more complex scenes. Combining the masks with rich descriptions of scenes, the trained model can generate new and realistic text images. We use GPT-4 [1] to generate the prompts. To ensure that the new prompts and the training prompts are in the same domain, and avoid domain gaps in the images generated by the model, we first provide GPT-4 with 50 caption examples produced by Monkey. Then we ask GPT-4 to mimic the style of these captions and synthesize new prompts. The instruction is *Please follow the above caption examples and generate a similar caption, which must contain some double-quoted spaces " ".* Next, we insert the text corresponding to each new mask into the quotation marks in the new prompt, forming the final prompt. The generated <prompt, mask, image> triplet is shown in Fig. 3 of the paper.

2 Applications

2.1 Text Removal

Text removal refers to the process of erasing or deleting text regions from an image. The finer the text mask, the better the erasing performance, as it preserves more background pixels. Therefore, stroke-level text segmentation can greatly benefit this task. Text removal is essentially an image inpainting task, so LaMa [4] is employed and the results are shown in Fig. 2.

Fig. 3: Visualization for text background replacement. The first column displays the original images. The second column shows the predicted masks from our WASNet. The remaining three columns show the images whose backgrounds have been replaced.

2.2 Text Background Replacement

Once we have obtained the fine mask of the text, we can freely replace the background of the image, embedding the text into various scenes. We use ControlNet [5] to replace the background and Fig. 3 presents the results.

Fig. 4: Visualization for text style transfer. The first column displays two style reference images. The second and third columns show the original images and the predicted text masks. The last column displays the images with stylized text.

2.3 Text Style Transfer

Text style transfer is a task that renders text in natural images into artistic text according to a style reference image while keeping the text content unchanged.

It usually relies on accurate text masks. We use Intelligent Typography [3] as the style transfer model and input the predicted text masks to it. The stylized text is shown in the last column of Fig. 4.

References

1. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y.: Sparks of artificial general intelligence: Early experiments with gpt-4 (2023)
2. Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., Bai, X.: Monkey: Image resolution and text label are important things for large multi-modal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26763–26773 (2024)
3. Mao, W., Yang, S., Shi, H., Liu, J., Wang, Z.: Intelligent typography: Artistic text style transfer for complex texture and structure. IEEE Transactions on Multimedia (2022)
4. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2149–2159 (January 2022)
5. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 3813–3824. IEEE (2023)