Supplementary Material of

DCDM: Diffusion-Conditioned-Diffusion Model for Scene Text Image Super-Resolution

Shrey Singh[1], Prateek Keserwani[1], Masakazu Iwamura[2], and Partha Pratim Roy[1]

[1] Indian Institute of Technology, Roorkee, Uttarakhand-247667, India
{ssingh19, pkeserwani, partha}@cs.iitr.ac.in
[2] Osaka Metropolitan University, Sakai, Osaka 599-8531, Japan
masa.i@omu.ac.jp

A The advantages of TCDM [23], which should be taken into account for a fair comparison

We discuss two advantages of TCDM and how they affect the experimental results and our findings.

Two advantages of TCDM TCDM has two advantages compared to other conventional methods and our proposed method. For a fair comparison, we should take them into account.

The first advantage of TCDM is to use ground truth text in training, as described in the part of "training strategy" in Section 4.2. In other words, TCDM always uses correct labels in the training. In contrast, the proposed method was trained using the recognition results of the HR images like other conventional methods. As shown in Tab. 1, the accuracy of HR images is no better than 82% on average. Since it is known that wrong labels significantly spoil recognition accuracy [50], this is a significant advantage of TCDM.

The second advantage of TCDM is to use large synthetic data in training. Figure 3 of the TCDM paper shows the effect of using synthetic data in training, as shown in Fig. A.1. The figure shows that TCDM gains its accuracy thanks to their synthetic data. Since the primary contribution of the TCDM paper is the introduction of large synthetic data in training, such a result is reasonable. However, comparing methods under different conditions (i.e., some methods are trained with large synthetic data and others are without) is not fair.

How do the advantages of TCDM affect? In the quantitative evaluation in Section 4.3, we compared the proposed method with conventional methods. As shown in Tab. 1, our proposed method clearly outperformed all conventional methods except TCDM and achieved comparable accuracy with TCDM. Therefore, considering the purpose of this paper—to seek an answer to the question,

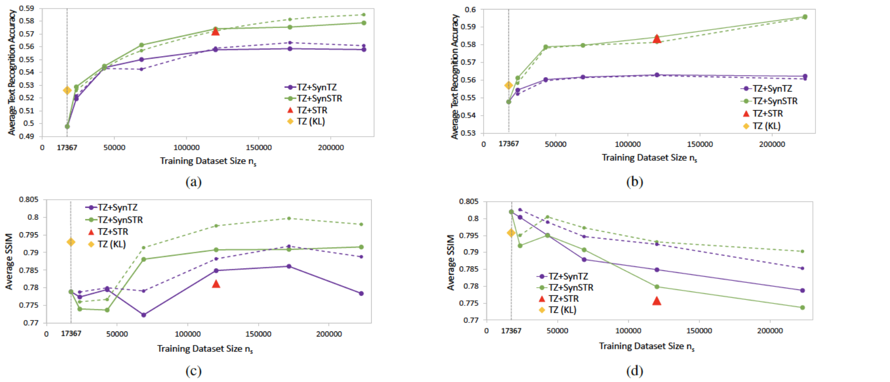

Figure 3. Evaluation results of TATT and the text-conditional DM trained on the augmented datasets. (a) and (b) show average recognition accuracy of TATT and the text-conditional DM, respectively. (c) and (d) show average SSIM of TATT and the text-conditional DM, respectively. Dotted lines show the results of fine-tuning using TextZoom only. The solid and dotted lines in the same color correspond to the same augmented dataset. CTC loss was used to train the text prior generator except in the case of "TZ (KL)." Here, "TZ (KL)" indicates the case where KL loss was used. The size of TextZoom is 17,367.

Fig. A.1: Figure 3 of the TCDM paper [23]

"Is a text recognizer necessary for STISR in inference?"—we could find an answer that utilizing a latent text diffusion model can effectively replace a text recognizer.

How do the advantages of TCDM affect it? In Tabs. 1 and 2, TCDM with the first advantage was evaluated. Suppose we can compare the proposed method with TCDM in a fair condition–without the first advantage. In that case, the proposed method may outperform TCDM, and we may conclude that *utilizing a latent text diffusion model can be a better option than a text recognizer.* However, our paper could not show this for two reasons. First, we were unable to complete an experiment using ground truth text in training soon after the TCDM paper was published in WACV 2024. Second, we were unable to obtain neither the synthetic data nor pretrained models used in the TCDM paper because they have not yet been released while they state they are going to do so. Although the source code of TCDM was publicly available on their GitHub page (https://github.com/toyotainfotech/stisr-tcdm), we first needed to reproduce their results, which was difficult to do in a short time. Therefore, this is a limitation of our paper and part of future work.

B Failure examples of the proposed method, corresponding to the successful cases in the qualitative evaluation shown in Fig. 3

In Fig. A.2, we show failure examples of the proposed method, although it is not common to show negative examples in previous research papers on this reserach topic. In general, the images generated by the proposed method are less blurred and of better quality than those by other STISR methods. In column 1,

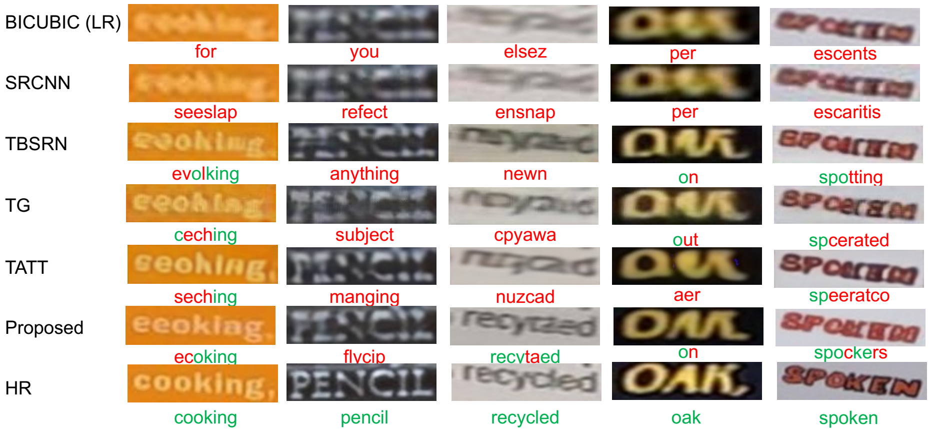

Fig. A.2: High resolution images generation by various state-of-the-art methods and the proposed method, failure cases for the proposed method. BICUBIC (LR) and HR indicate the input to the methods and ground truth, respectively. The text recognition results by ASTER [33] are shown under the images. The characters in green and red indicate correct and incorrect recognition results, respectively.

while the LR images are blurred and contain ambiguous stroke information, the STISR methods generated clearer images. However, the first letter, 'c,' tends to be misrecognized as 'e' and 's' in all methods except TG, which may be caused by a noisy text prior. Column 2 exhibits LR images with discontinuous strokes blending into the background, complicating accurate recognition. Columns 3 and 4 show severe blurriness in the LR images, leading to distorted information and mispredictions despite continuous strokes. Column 5 presents difficulties when heavy blur text fonts protrude from the background, making accurate prediction challenging even with correct information. These examples highlight the model's struggles with ambiguous strokes, discontinuities, blurriness, and font complexities, affecting its super-resolution and recognition accuracy.

References

50. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. In: Proc. International Conference on Learning Representations (ICLR) (2017), https://arxiv.org/abs/1611.03530 1