Our research paper titled “Scene Text Image Super-resolution based on Text-conditional Diffusion Models” has accepted for presentation at WACV2024, an international conference in the field of computer vision.

Toyota Motor Corporation
Social System PF Development Div.
InfoTech-AS Data Analysis Team
Chihiro Noguchi
Shun Fukuda
Masao Yamanaka

Toyota Motor Corporation has developed image generation AI technology related to text recognition. Our research paper titled "Scene Text Image Super-resolution based on Text-conditional Diffusion Models" was accepted at the WACV2024 (Winter Conference on Applications of Computer Vision 2024), an international conference in the field of computer vision, and was presented at the main conference on January 4, 2024.

Backgrounds

Our daily lives are filled with various forms of text information, and the ability to accurately read this information is beneficial in many applications. For example, images captured by onboard cameras contain text information such as traffic signs and license plates. For services like autonomous driving and automatic map maintenance to be realized, it is extremely important to accurately read this information. However, text areas in images are often small, and are strongly affected by blur and noise, presenting various challenges for accurate reading.

Scene text image super-resolution is a super-resolution technology specialized for text images. Its purpose is to restore images from poor-condition text images (low resolution, containing strong blur/noise, etc.) to their pre-degradation state (Figure 1). The main use case is as a preprocessing step for text recognition models, where converting to more readable text images can improve the recognition accuracy of subsequent recognition models. This technology is particularly useful in situations where it is difficult to directly tune text recognition models (for example, when using APIs from other companies, or when it is already integrated into a

product and cannot be frequently updated).

Figure 1 Examples of text images. The recognition results of the text recognition model are shown below the images. Red text indicates misrecognized characters. (Top) The image resized using bilinear interpolation of the input text image. The degradation is severe, making it difficult to read the text. (Middle) The image after applying super-resolution. Readability is greatly improved, and the high-resolution image is well restored. (Bottom) The correct high-resolution image.

Technical Overview

One of the major challenges in the field of scene text image super-resolution is the high cost of collecting training datasets. For training, low-resolution and high-resolution pair images, which are pixel-wise positionally aligned, are required, but preparing these in large quantities necessitates significant human effort. Although methods have been proposed that combine synthetic blur and noise to replicate realistic degradation patterns, simulating image degradation that occurs in real environments is challenging, especially for restoring text images with severe degradation. Therefore, the proposed method learns degradation patterns from a limited number of existing low-resolution and high-resolution pair images and applies them to text images to generate a large quantity of synthetic paired images that are closer to real images. By using this augmented dataset for training existing scene text image super-resolution models, we have achieved a significant improvement in performance.

In the proposed method, we use a diffusion model as the image generation model. Diffusion models are used as conditional generative models in various domains due to their stability in learning. While diffusion models have a very high expressive ability, their long inference time is known as a major drawback. In the proposed method, the generated pair images are used as the training dataset for existing scene text image super-resolution methods, so the slowness

of inference time is not a major problem. On the other hand, the quality of the generated text images is very important in constructing the training dataset. In the paper, we first consider the diffusion model-based method as a natural extension of existing scene text image super-resolution models. In our experiments, despite this simple extension, we have obtained results that significantly surpass existing methods. However, we are not satisfied with the performance improvement here because we consider the approach of using images generated by the learned diffusion model as the training dataset for existing models.

In scene text image super-resolution, there are two types of ground truth data. One is the high-resolution image that is the target for generation. This data is difficult to obtain in large quantities due to cost, as mentioned above. The other type of ground truth data is the text in the text image. This is relatively easy to obtain and can be acquired in large quantities from datasets for text recognition. The proposed method considers a diffusion model conditioned not only on low-resolution images but also on the ground truth text, and learns the distribution for generating high-resolution images. The effect of using ground truth text as a condition is extremely significant; in our experiments, we have obtained results where super-resolution images of quality comparable to that of the ground truth high-resolution images can be restored (Figure 3). The proposed method uses a diffusion model conditioned on the ground truth text to generate paired images, and by using the generated images as a training dataset for existing models, we have achieved a dramatic improvement in performance.

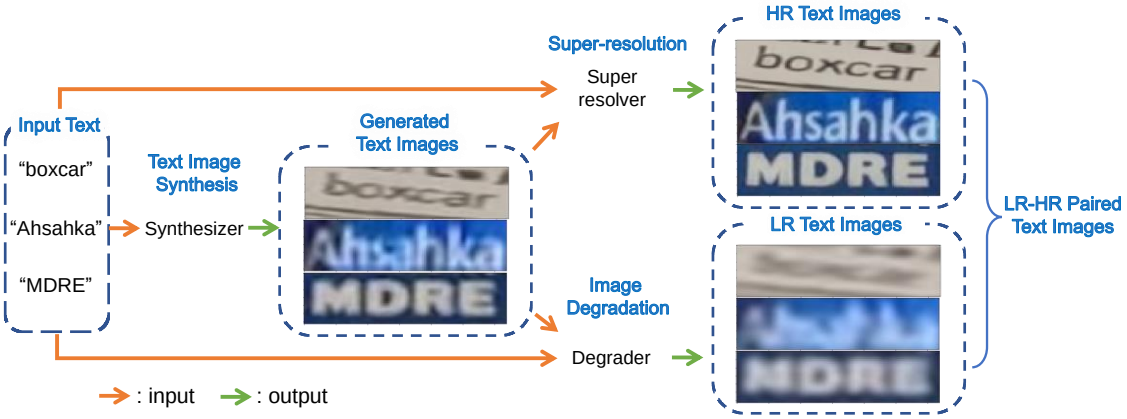

Figure 1　Overview of proposed framework

Figure 1 shows the overview of the low-resolution and high-resolution pair image generation framework we propose. The proposed method consists of three diffusion models. The first is a text image generation model (Synthesizer). This model generates text images corresponding to the inputted text. Note that the objective here is to prepare text images and corresponding

texts, and instead, it is possible to use existing text recognition models' training datasets. In the experiments in the paper, we examined two patterns: using the existing dataset for text recognition and using images generated from the text image generation model trained with it. The second is the scene text image super-resolution model (Super-resolver). As mentioned above, it transforms the input text image into a high-quality super-resolution image using the ground truth text as a condition. The third is the image degradation model (Degrader). It transforms the input text image into a degraded image. The text super-resolution model and the image degradation model were trained using an existing pair image dataset called TextZoom. Also, the datasets for text recognition often contain relatively less degraded images, and applying the image degradation model enables the generation of images more suitable for the super-resolution model's training.

Method	GT Text	SSIM ($\times 10^{-2}$)	PSNR	Acc. (%)
Bicubic		69.61	20.35	26.8
HR		-	-	72.4
SRCNN [10]		72.27	20.78	29.2
SRResNest [25]		74.03	21.03	35.1
TSRN [51]		76.90	21.42	41.4
STT [5]		76.14	21.05	48.1
PCAN [61]		77.52	21.49	47.4
TG [6]		74.56	21.40	48.9
TATT [29]		79.30	21.52	52.6
C3-STISR [62]		77.21	21.51	53.7
DDPM*		79.50	22.70	55.0
TCDM		**79.58**	**22.83**	**55.7**
TATT	✓	79.34	22.35	61.0
TCDM	✓	**80.25**	**22.86**	**68.1**

Figure 2　Results of quantitative evaluation

Figure 3 shows the results of the quantitative evaluation of scene text image super-resolution using a diffusion model. SSIM and PSNR are metrics widely used in the general super-resolution, indicating the average difference between the generated super-resolution images and the ground truth high-resolution images. Acc. indicates the recognition accuracy by a pretrained text recognition model. The text conditional diffusion model (TCDM) employed in the proposed method achieves higher performance (55.7%) than existing methods, even without using the ground truth text (GT Text). When the ground truth text is used as input, performance dramatically improves (68.1%), achieving recognition accuracy comparable to that in the ground truth high-resolution images (72.4%).

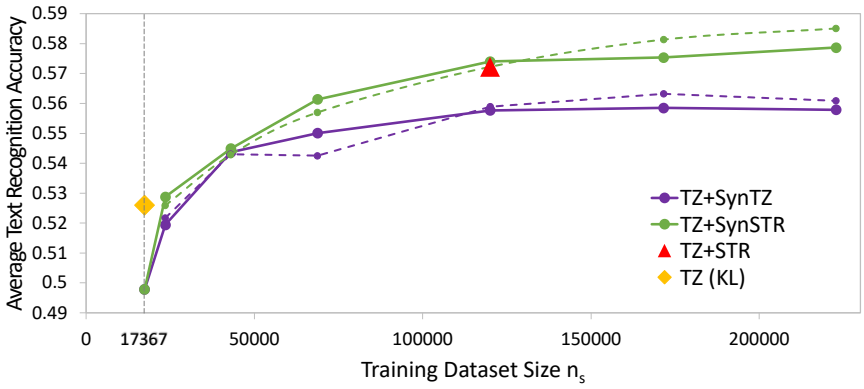

Figure 3　Results of training existing scene text image super-resolution method (TATT) with generated pair images.

Figure 4 shows the evaluation results when using pair images generated by the proposed method as training data for an existing scene text image super-resolution model (TATT). The horizontal axis represents the number of training datasets, and the vertical axis represents the recognition accuracy evaluated using a pretrained text recognition model. The results confirm the usefulness of the proposed method, as increasing the number of training datasets improves recognition accuracy.

Summary and Future Works

In this paper, we have provided a technical explanation of our research paper on scene text image super-resolution. Our proposed method uses a diffusion model to learn degradation patterns from existing low-resolution and high-resolution pair datasets, and constructs an augmented dataset from the existing dataset. Our experimental findings reveal that by using ground truth text as an additional condition for the diffusion model, it is possible to more accurately restore the original high-resolution image. Using this property, our proposed method constructs an augmented dataset and uses it to train existing scene text image super-resolution models, achieving a significant performance improvement in recognition accuracy.

Our proposed method required a low-resolution and high-resolution pair dataset for learning degradation patterns. However, the pair datasets available as public datasets are very limited and are currently restricted to English text images. Therefore, the applicability of our proposed method is limited to English text images, presenting challenges for practical use outside the English-speaking world. As a future development, we are advancing the study of

methods that can adapt to non-Latin characters by learning degradation patterns in a way that does not depend on pair images and applying them in a manner suitable for text images.